Main Session Talks

WHORE in the 17th century: Exploring the development of meaning across four registers

Helen Baker, Vaclav Brezina and Tony McEnery
Lancaster University

New resources such as the EEBO-TCP corpus are allowing us to look at the changing nature of word meaning over time with a level of detail not previously possible for the great variety of words. The EEBO-TCP corpus has been developed at Lancaster University using texts available from the EBBO-TCP. In constructing the corpus, we retained all metadata provided by EEBO-TCP and supplemented that with part of speech annotations, spelling regularizations and genre information. The corpus is currently in its third version, hence is referred to as the EEBO-TCP v3 corpus.

While previous historical corpora, such as ARCHER or the Helsinki corpora, were of use for exploring frequent, principally grammatical, words, corpora of the scale of EEBO-TCP v3 corpus, which provides nearly one billion words with genre classifications for the seventeenth century alone, allow a systematic study of words which barely occur, or do not occur at all, in existing historical corpora. In addition to being able to look at the dynamic nature of word meaning using this corpus, we are also able to see the interaction between register and word meaning over time.

We had previously used this corpus to analyse the shifting semantics of 'whore' manually, using collocation and concordance analysis to develop a sense of how the word shifted in meaning across the century (McEnery and Baker, 2016). However, such analyses are highly time consuming, though they provide well contextualized analyses. In this paper we will compare the findings of that manual analysis with an automated approach to identifying where meaning shifts occur in the century. Our approach does not offer an automation of the analysis of the meaning of the word – that is still best done, in our view, using machine aided corpus lexicographic methods. What we wish to achieve is a technique which focuses the analysis at points in time when meaning is in transition. This would speed up the process of exploring meaning variation significantly, allowing yet more time for the analyst to explore the context and causes of meaning change. Given that the analysis of McEnery and Baker (ibid) had already shown a strong genre effect in meaning change, our overall analysis in this paper will compare the results of McEnery and Baker (ibid), both in overall terms and with regard to genre to assess the extent to which the method develops would have proved a useful guide to that original study.

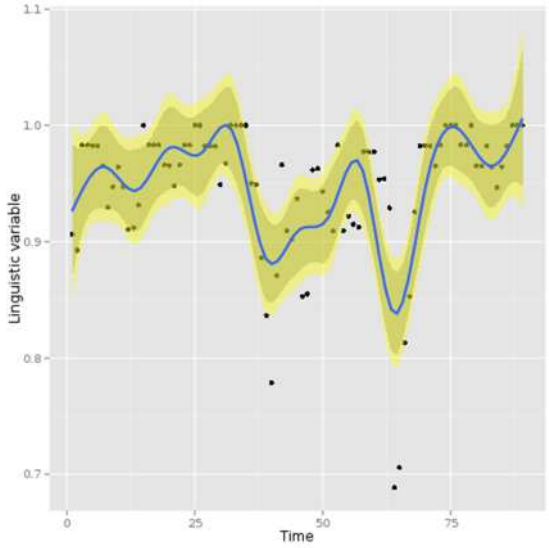
To demonstrate this, in this paper, we use a subset of the EEBO v.3 corpus to investigate the change of meaning of the term 'whore' in the 17th century across four genres. We selected four subcorpora representing four prominent genres – plays, poetry, religious writing and treatises – for this purpose. The details about the size of the subcorpora are displayed in Table 1.

Table 1: EEBO genre-based subcorpora

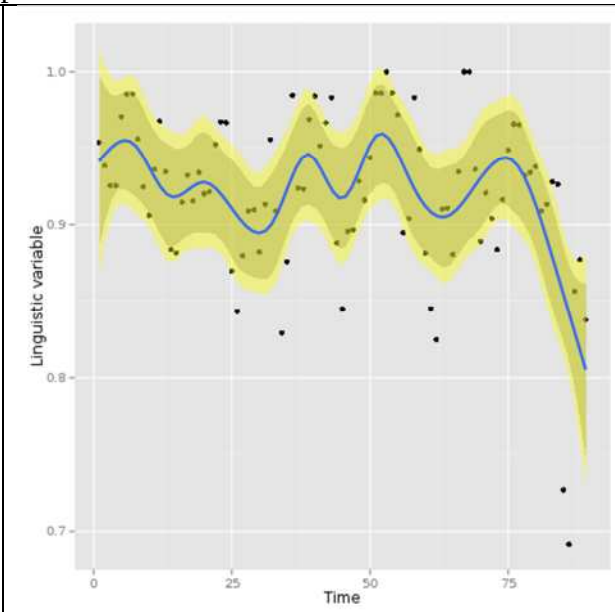
Genre	Words (tokens)	Texts
Plays (comedies, tragedies, histories..)	22,116,923	806
Poetry	33,975,788	1,669

Religious writing	162,581,090	3,462
Treatises	66,483,534	879
TOTAL	285,157,335	6,816

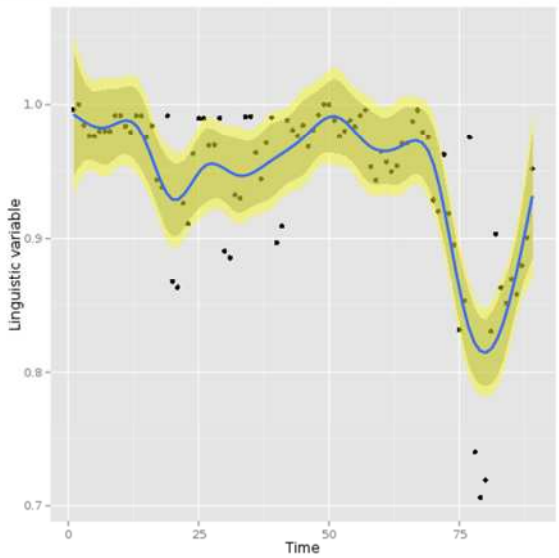
In the paper we will outline the technique we have developed for measuring and visualizing meaning change. This has four main components: i.) the use of collocation statistics; ii.) overlapping sliding windows which move through the century within which collocates are determined; iii.) an estimation of difference between sliding collocate windows in sequence and iv.) a non-parametric regression model which is applied to the data (Gabrielatos et al. 2012) to trace statistically significant points of departure where semantic change took place. Four examples of the visualizations from the quantitative analysis are displayed in Figure 1. These will be explained in the presentation. In brief, the greater a 'peak' or 'trough' on the curve, the greater a perturbation of meaning is assumed to happen at that point, i.e. meaning is in transition at that point.



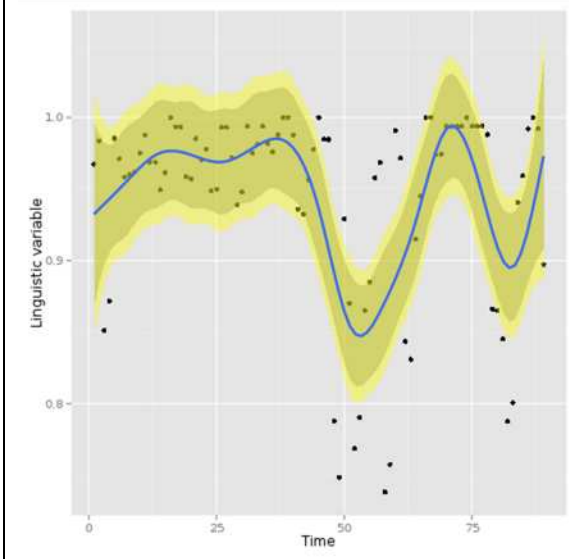
Whore_plays



Whore_religious writing



Whore_poetry



Whore_treatise

Figure 1: The development of meaning of the term 'whore' in the 17th century

The paper will present the findings of our investigation, showing how well the technique outlined above mirrors the previous manual analysis of the word 'whore'. In doing so, we will explore and explain the interaction between genre and the meaning of the word 'whore' in the seventeenth century. To conclude we will reflect on the need for a historically informed analysis of the graphs produced by the technique.

The missing components in the culinary recipes of the early Modern English period

Magdalena Bator
University of Social Sciences, Warsaw

The recipe as a text type has already been discussed by such scholars as Görlach, Carroll, or Taavitsainen. Görlach (1992, 2004) discussed the culinary recipe, Taavitsainen (2001) dealt with the medical texts, whilst Carroll (1999) offered examples from both groups. They presented a variety of linguistic features which make the recipe a distinctive text type. Additionally, a number of publications offer an analysis of the structure of the medical recipe, see for instance Stannard and Hunt. The majority of these studies concentrated on medieval recipes.

In the present paper, we will concentrate on the structure of the culinary recipe in the Early Modern English period, i.e. at the time when its form has already been fixed. And by using the patterns offered by Stannard (1982) and Hunt (1990) in their analysis of the medical texts, we will demonstrate that the culinary recipe differed from the medical one, and thus it should be treated as a separate subcategory of the text type. The elements which deserve special attention are: the heading and the procedure.

The corpus for the present study consists of the culinary recipes taken from four early Modern English collections: *A book of cookrye, The compleat cook, A new book of cookery,* and *A proper new book of cookery*. Additionally, a number of examples from the Middle English period will be shown in order to provide a certain background for the development of the analysed text type.

REFERENCES

Carroll, R. 1999. "The Middle English recipe as a text-type", *Neuphilologische Mitteilungen* 100: 27-42.

Görlach, Manfred. 1992. "Text types and language history: the cookery recipe", in: Rissanen, M. (et al.) (eds), *History of Englishes. New methods and interpretations in historical linguistics* Berlin: Mouton de Gruyter, 736-761.

Görlach, M. 2004. *Text types and the history of English*. Berlin; New York: Mouton de Gruyter.

Hunt, T. 1990. *Popular Medicine in 13th-century England. Introduction and Texts*. Cambridge: Brewer.

Stannard, J. 1982. "Rezeptliteratur as Fachliteratur", in: Eamon, W. (ed.) *Studies on Medieval Fachliteratur*. Scripta 6. Brussels, 59-63.

Taavitsainen, I. 2001a. "Middle English recipes: Genre characteristics, text type features and underlying traditions of writing", *Journal of Historical Pragmatics* 2: 85-113.

Taavitsainen, I. 2001b. "Changing conventions of writing: the dynamics of genres, text types, and text traditions", *European Journal of English Studies* 5/2: 139-150.

An investigation of diachronic change in hypotaxis and parataxis in German through language contact with English in translation

Mario Bisiada
Universitat Pompeu Fabra

Translation is a language contact situation that can influence language change (cf. Kranich, Becher & Höder 2011, Kranich 2014). However, whether such language contact with English has led to change in German is still the subject of debate (Becher, House & Kranich 2009, Hansen-Schirra 2011, Neumann 2011, Bisiada 2013). This paper investigates a frequency shift from hypotactic to paratactic constructions in concessive clauses in German management and business articles. The research hypothesises that the influence of the English verb-second word order may cause language users of German to prefer verb-second, paratactic constructions to verb-final, hypotactic ones. A previous study has provided some evidence that parataxis may be in the process of replacing hypotaxis in concessive clauses in the popular science genre, based on a diachronic corpus study of texts between 1978–1982 and 1999–2002 (Becher 2011).

This paper challenges that claim based on a 1 million word diachronic corpus, with texts dating from 1982–3 and 2008. The corpus combines a parallel corpus architecture of German translations and their source texts with a comparable corpus of German non-translations. The time span under analysis is adopted from Becher (2011) and will thus allow us to compare differences in the way the English concessive conjunctions *although*, *though*, *even though* and *while* have been translated in management articles compared to popular science articles.

The study finds that the concessive conjunctions under analysis were translated mainly hypotactically (63%) in 1982–3, and show a decrease in hypotactic translation in 2008 to 43%. At the same time, the frequency with which they were translated paratactically has increased from 43% in 1982–3 to 52% in 2008. In the corpus of non-translated texts, however, the number of hypotactic and paratactic structures has remained stable over the period of analysis.

These findings indicate that parataxis is indeed coming to be used more often in German translations. As the non-translations show no diachronic change, however, the development seems to be limited to translated language. Thus, no evidence has been gathered in support of the thesis that language contact with English in translation may lead to language change even in originally produced German business texts. Instead, the increase in parataxis seems to be driven by a development towards syntactically simpler constructions in the business genre, which is most evident, first, in the strong tendency towards sentence splitting in translation (reported in Bisiada 2014) and, second, in an increasing use of sentence-initial conjunctions (also observed by Becher, House & Kranich 2009), especially in the case of *doch* and *denn*.

The latter development may indeed be argued to be influenced by contact with English, which frequently uses such constructions. This type of research is limited, of course, in that it cannot show whether language contact has taken place in translation or simply by reading English-language articles or using English as a *lingua franca*. But taking into account the observation of an increasing frequency of sentence splitting in translation, there might be a tendency for German language users in this genre to express logical relationships using anaphoric cohesive reference rather than subordinative conjunction.

REFERENCES

Becher, Viktor. 2011. Von der Hypotaxe zur Parataxe: Ein Wandel im Ausdruck von Konzessivität in neueren populärwissenschaftlichen Texten. In Eva Breindl,

Gisella Ferraresi & Anna Volodina (eds.), Satzverknüpfungen. Zur Interaktion von Form, Bedeutung und Diskursfunktion, 181–209. Berlin: de Gruyter.
Becher, Viktor, Juliane House & Svenja Kranich. 2009. Convergence and divergence of communicative norms through language contact in translation. In Kurt Braunmüller & Juliane House (eds.), Convergence and divergence in language contact, 125–152. Amsterdam: John Benjamins.
Bisiada, Mario. 2013. Changing conventions in German causal clause complexes: A diachronic corpus study of translated and non-translated business articles. Languages in Contrast 13(1). 1–27.
Bisiada, Mario. 2014. "Lösen Sie Schachtelsätze möglichst auf": The impact of editorial guidelines on sentence splitting in German business article translations. Applied Linguistics Advance online access.
Hansen-Schirra, Silvia. 2011. Between normalization and shining-through: Specific properties of English–German translations and their influence on the target language. In Svenja Kranich, Viktor Becher, Steffen Höder & Juliane House (eds.), Multilingual discourse production: diachronic and synchronic perspectives, 135–162. Amsterdam: John Benjamins.
Kranich, Svenja. 2014. Translations as a locus of language contact. In Juliane House (ed.), Translation: A multidisciplinary approach, 96–115. Basingstoke: Palgrave Macmillan.
Kranich, Svenja, Viktor Becher & Steffen Höder. 2011. A tentative typology of translation-induced language change. In Svenja Kranich, Viktor Becher, Steffen Höder & Juliane House (eds.), Multilingual discourse production, 11–44. Amsterdam: John Benjamins.
Neumann, Stella. 2011. Assessing the impact of translations on English–German language contact. In Svenja Kranich, Viktor Becher, Steffen Höder & Juliane House (eds.), Multilingual discourse production: diachronic and synchronic perspectives, 233–256. Amsterdam: John Benjamins.

Connectives and the spoken/written dimension: Connective profiling Old Norse texts

Hannah Booth
University of Manchester

It is traditionally assumed that spoken and written language differ greatly in their respective use of connectives (e.g. Chafe 1985). The syntax of speech has been considered to be 'simpler', characterised by simple structures, parataxis and fragmentation, accompanied by a limited set of semantically 'basic' connectives (e.g. *and, or, but*). By contrast, written language has been thought to show higher levels of syntactic complexity, including more dependent clauses, thus requiring a greater variety of complex connectives (e.g. *due to the fact that, whereas*). A significant amount of recent research on the spoken/written distinction domains has, however, prompted many scholars to claim that the distinction is in fact not so clear (e.g. Biber 1988; Miller & Fernandez-Vest 2006). This paper contributes to this debate by demonstrating that an understanding of the usage of connectives must also take into account concerns of genre, alongside those of the spoken/written dimension.

I present a study of connective usage across various Old Norse textual genres. The Old Norse corpus (attestation c.1150-1350) is astoundingly rich for a medieval language, spanning a range of genres (e.g. poetry, literary and learned prose, translation and legal texts). It therefore offers great potential for genre-based studies, a potential which has scarcely been exploited to date. Crucially, certain genres of the period are strongly associated with a native oral tradition (e.g. saga narrative), whilst others are the fruit of an explicitly literate process (e.g. learned writings — see Quinn 2009; Raschellà 2007; Sigurðsson 2004). Moreover, text types generally associated elsewhere with the written medium are in Old Norse strongly rooted in orality (e.g. legal texts — see Fix 1993; McGlynn 2009). These special characteristics have been thoroughly explored by historians and literary scholars of medieval Scandinavia, yet rarely feed into linguistic scholarship on the period.

Using the IcePaHC corpus (Wallenberg et al. 2011), I examine the 'connective profiles' of four Old Norse (Icelandic) texts, which together represent different points on the orality/literacy scale: (1) *The First Grammatical Treatise* (expository prose); (2) *Grágás* (law code); (3) *Icelandic Book of Homilies* (sermon); (4) *The Saga of Icelanders* (narrative). I

apply a methodology developed by Kohnen (2007), whereby the frequency of the major coordinators and subordinators in each text is investigated to produce four distinct 'connective profiles'. Specifically, the proportion of coordinators to subordinators is considered for each text. Comparison of the four profiles reveals a number of patterns, some rather unexpected, such as the surprisingly high frequency of subordinators in *Grágás,* the 'orally-derived' legal text. Most importantly, the outcome of the profiling investigation prompts a reassessment of the context-specificity of connectives in spoken and written language. The traditional assumption is shown to be too simplistic, given that other genre-related concerns which transcend the spoken/written dimension also appear to play a crucial role, as seen with the frequent use of a conditional construction 'if X does Y, then he will receive punishment Z' in legal texts. The analysis of individual connectives and semantic sets of connectives (e.g. conditionals) will be used to reveal these patterns.

REFERENCES

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

Chafe, Wallace L. 1985. Linguistic differences produced by differences between speaking and writing. In Olson, David R., Nancy Torrance & Angela Hildyard (eds.) *Literacy, Language and Learning*. 105-23. Cambridge: Cambridge University Press.

Fix, Hans. 1993. Grágás. In Pulsiano, Phillip & Kirsten Wolf (eds.) *Medieval Scandinavia: An encyclopedia.* 234-5. New York: Garland.

Kohnen, Thomas. 2007. 'Connective Profiles' in the history of English texts. In Lenker, Ursula & Anneli Meurman-Solin (eds.) *Connectives in the History of English: Selected Papers from 13th ICEHL, Vienna, 23-28 August 2004*. 289-308. Amsterdam: John Benjamins.

McGlynn, Michael P. 2009. Orality in the Old Icelandic Grágás: Legal formulae in the assembly procedures section. *Neophilologus* 93. 521-36.

Miller, Jim & M. M. Jocelyne Fernandez-Vest. 2006. Spoken and Written Language. In Bernini, Giuliano & Maria Schwartz (eds.) *Eurotype: Typology of Languages in Europe. Pragmatic Organization of Discourse in the Languages of Europe.* 9-64. Berlin: de Gruyter.

Quinn, Judy. 2000. From orality to literacy in medieval Iceland. In Clunies Ross, Margeret (ed.) *Old Icelandic Literature and Society*. 30-60. Cambridge: Cambridge University Press.

Raschellà, Fabrizio D. 2007. Old Icelandic grammatical literature: The last two decades of research (1983-2005). In Quinn, Judy, Kate Heslop & Tarrin Wills (eds.) *Learning and Understanding in the Old Norse World: Essays in honour of Margaret Clunies Ross*. 341-72. Turnhout: Brepols.

Sigurðsson, Gísli. 2004. *The Medieval Icelandic Saga and Oral Tradition: A discourse on method.* Translated by Nicholas Jones. Cambridge, MA: The Milman Parry Collection of Oral Literature.

Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC).* Version 0.9. http://www.linguist.is/icelandic_treebank.

Gender differences in *soc.men* and *soc.women*: A diachronic study

Elli E. Bourlai
Indiana University Bloomington

Gender has been one of the main research interests in the area of Computer-Mediated Communication in the past two decades. Because of the anonymity and distance provided by the medium, many scholars argued that the linguistic boundaries between the two genders would start blurring. Several studies support this Technological Determinism framework (Whitmer and Katzman, 1997; De Oliveira, 2003; Huffaker and Calvert, 2005), but many others have found that traditional offline linguistic patterns have simply transferred into online communication supporting Social Constructionism (Selfe and Meyer, 1991; Herring, Johnson and DiBenedetto, 2003; Waseleski, 2006; Guiller and Durndell, 2007; Herring, 2010). However, most of the studies exploring this debate about gender language online have

either used synchronic corpora or attempted a diachronic comparison without a systematic sampling method.

This study presents the results of analyzing a set of linguistic features expressing gender in a diachronic corpus of approximately 2,000 e-mail messages. The messages were collected from the Usenet Newsgroups *soc.men* and *soc.women* in 4-year intervals, covering an 18-year period from 1989 to 2005. While most of the features analyzed seem to follow traditional gender patterns (1st person singular pronoun, words per sentence, dictionary words), some show convergence (exclamation marks, swear words), and others show reverse patterns (1st person plural pronoun, sexual words). Further analysis of the genders comparing the two lists indicated that the context of the online environment plays a very important role in the communication of the two genders, with women adapting their language to a more "male" linguistic style in *soc.men*, while men seemed to adapt to a more "female" linguistic style in *soc.women*. Thus, it may be argued that certain gender markers have been affected by the medium, whereas others are more susceptible to external societal changes; the frequency of both, nevertheless, depends on the context of the online environment.

This is a pilot study for a larger research project that will comprise a diachronic corpus of approximately 180 Usenet Newsgroups on a variety of different topics; as such, it has certain limitations. However, it brings up important issues regarding the collection, annotation, archiving and analysis of CMC diachronic corpora, since there is currently no methodological framework for designing diachronic corpora for online communication data. Consequently, this work not only aims at addressing the lack of diachronic CMC studies and the ongoing debate about the linguistic expression of gender online, but also methodological issues of designing diachronic CMC corpora.

REFERENCES

De Oliveira, S. M. (2003). Breaking Conversational Norms on a Portuguese Users Network: Men as Adjudicators of Politeness? *Journal of Computer-Mediated Communication, 9*(1).

Guiller, J., & Durndell, A. (2007). Students' linguistic behavior in online discussion groups: Does gender matter? *Computers in Human Behavior, 23*(5), 2240–2255.

Herring, S. C. (2010). Who's got the floor in computer-mediated conversation? Edelsky's gender patterns revisited. *Language@Internet 7*, article 8.

Herring, S. C., Johnson, D. A., & DiBenedetto, T. (1992). Participation in electronic discourse in a 'feminist' field. In *Locating Power: Proceedings of the 1992 Berkeley Women and Language Conference* (pp. 250-262). Berkeley: Berkeley Women and Language Group.

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2), article 1.

Selfe, C. L., & Meyer, P. R. (1991). Testing claims for on-line conferences. *Written Communication, 8*(2), 163–192.

Waseleski, C. (2006). Gender and the use of exclamation points in computer-mediated communication: An analysis of exclamations posted to two electronic discussion lists. *Journal of Computer-Mediated Communication*, 11(4), article 6.

Witmer, D. F. and Katzman, S. L. (1997). "On-Line Smiles: Does Gender Make a Difference in the Use of Graphic Accents?". *Journal of Computer-Mediated Communication, 2*(4).

A diachronic study of HAVE contractions in written American English

Joanne Close & Naomi Gradwell
University of Chester

This study investigates patterns of auxiliary and negation contraction with present tense auxiliary *HAVE* in written American English. The aims are to uncover patterns of variation and change between full and contracted forms and to consider factors which

condition the use of the contracted forms, such as genre and type of subject. There have been a number of studies of contracted auxiliary verbs, many of which have found that their usage within the English language is increasing (Krug 1994; Kjellmer 1997; Leech, Hundt, Mair & Smith 2009) and that informal and fiction texts contain a higher density of contracted auxiliaries than other texts (Kjellmer 1997; Leech et al. 2009). This study contributes to the field by conducting a study using the recently-compiled Corpus of Historical American English (COHA; Davies 2012).

COHA contains 400 million words of written American English from 1810 to 2009 across a range of genres. The corpus is POS tagged, making it possible for frequency counts to be generated according to specific, grammatical criteria e.g. a form of auxiliary *HAVE* followed by a past participle. Data from three decades (1810-1819, 1900-1909 and 2000-2009) were selected for analysis. Following Aarts, Close & Wallis (2013), the frequency of contracted forms of *HAVE* was calculated as a percentage of the total number of instances of *HAVE*. Therefore, we can be certain that any changes in frequency of contracted forms are not simply a reflex of changes in the frequency of full forms, i.e. an increase in contracted *'ve* is not simply the result of an increase in the use of auxiliary *HAVE*.

Results indicate a number of changes in the use of contracted forms *'ve* and *haven't*. A comparison of the full and contracted forms of HAVE in positive declarative contexts shows that the use of contracted auxiliary *HAVE* is increasing at the expense of the full form. The decade between 1810 and 1819 saw only 6.76% of instances of *HAVE* contracted, which increased to 13.93% by 1900-1910, and reached 27.95% by 2000-2009. A comparison of full and contracted forms in negative clauses (e.g. *have not* vs. *haven't*) shows a more dramatic increase: 5.71% of opportunities for negation contraction were realised between 1810 and 1819, by 1900-1909 the full and contracted forms were almost equal in their frequency of occurrence, and by 2000-2009 contracted *haven't* was used in the 75.62% of instances.

Preliminary results from the analysis of 1,000 random examples of *'ve* across the three periods shows that the preference for co-occurrence with the first person pronoun *I* persists, although in the two later decades the proportion of *'ve* occurring with other pronouns has increased; by 2000-2009, *'ve* occurs more frequently with other pronouns than it did between 1810 and 1819. In terms of genre, the use of contractions seems to be spreading from fiction to other genres. In the two later time periods, *'ve* and *haven't* have become more popular across in magazine and news texts and, to a lesser extent, the non-fiction genre.

REFERENCES

Aarts, B., Close, J., & Wallis, S. (2013). Choices over time: methodological issues in investigating current change. In: B. Aarts, J. Close, G. Leech & S. Wallis (Eds.), *The verb phrase in English: Investigating recent language change with corpora* (pp.14-45). Cambridge: Cambridge University Press,

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora 7*(2), 121–157. DOI: 10.3366/corp.2012.0024

Kjellmer, G. (1997). On contraction in Modern English. *Studia Neophilologica, 69*(2), 155-186.

Krug, M. (1994) *Contractions in spoken and written English: A corpus-based study of short-term developments since 1960.* Unpublished MA thesis, University of Exeter.

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

Towards a diachronic analysis of classical Ethiopic (Ge'ez)

Wolfgang Dickhut, Andreas Ellwardt, Cristina Vertan
Universität Hamburg

Ge'ez, a Semitic language of the South Semitic language branch, which has remained the by far most important vehicle of written knowledge throughout Ethiopian history until the nineteenth century, offers a particular case study for a Semitic language that developed on African soil, with remarkably little influence from non-Semitic Afro-Asiatic languages (Cushitic), also preserving, due to its peripheral position, a number of ancient Semitic features. The Ge'ez language was used for centuries and grew to reflect all the changes in the tradition. The existence of witnesses, with manuscripts being available from over eight centuries, and the span growing to two millennia if one considers the epigraphic evidence, offers the unique possibility of a diachronic approach to an African language. The Ethiopian textual heritage is profoundly marked by the pervasive intertextual phenomenon of the translation that takes place in distinct times (Aksumite and Post-Aksumite periods) and from (we do not know if exclusively) two different languages: from Greek in the first, and from Arabic in the latter period. There is no material gap between the two, since both of them (with unique exceptions) are materially documented and transmitted in manuscripts dating from the fourteenth century at the earliest.

With the help of a new tagged and parsed corpus and the incorporated search and visualization tools it shall be possible for the first time to present a new diachronic picture of the Ge'ez language which, even if not spoken for most of the past millennium, was actively used for the entire period and remained until well into the nineteenth century the only medium of written expression. For the first time, changes in linguistic patterns, in word use and stylistic choices shall be analysed and accounted for on an objective basis. It will be possible to contrast different style and language standards representative for various genres and periods, possibly and ultimately also for dialectal areas of text production and/or transmission ('scriptoria'). Each relevant sub-standard, defined by the corresponding corpus representative for the period – once contrasted with each other, and of course with the standard grammatical descriptions of Ge'ez – will define diachronic stylistic and linguistic features. Marking features will be singled out for each genre within one substandard. It is of course not predictable which kind of correlation might prevail, and this is of course one of the most important expected result of the research. From all the above sub-tasks, the implementation of the final project scope becomes possible, that is, a new approach to the history of Ethiopic language and literature in their interconnections. In our contribution, we will limit ourselves to presenting the annotation schema and tool which will allow us to perform this diachronic analysis of Ge'ez.

'From above', 'from below' and regionally balanced: Towards a new corpus of Nineteenth Century German

Stephan Elspaß & Konstantin Niehaus
Universität Innsbruck

In this talk, we report on an ongoing project on creating a corpus of German in the 19th century. The recent and current projects on building historical corpora from OHG to NHG until 1800 (e.g. early New High German 1650–1800 cf. Scheible et al. 2011), as well as the 'Variantengrammatik' project on present-day German (cf. Dürscheid & Elspaß 2015), all account for regional diversity in German. The 19th and 20th centuries remain the only period for which no big corpora with a focus on regional variation exists. The *Nineteenth Century German Corpus* (*NiCeGerman*) seeks to fill this gap. In addition to building a regionally balanced corpus, it will also account for an increasing demand for register variation, e.g. by considering not only conceptual literacy, but also orality in the history of New High German

(cf. Ágel & Hennig 2006). Thus, text genres 'from above' and 'from below' will be incorporated (cf. Elspaß & Niehaus 2014): A corpus of private letters (with a total of ca. 500,000 words) and a newspaper corpus (ca. 100,000 words), which is being expanded at present, form the basis of our corpus design (cf. Elspaß & Niehaus 2014). With respect to newspaper texts, the corpus design will consider the text-linguistic heterogeneity of this genre. Therefore, a distinction between regional news, 'entertainment' news and news agencies'/foreign correspondents' texts will be drawn. This will allow for a finer grained stylistic differentiation of newspaper texts.

A case study from 19th century grammar, the so-called *Ausklammerung* ('exbraceration') of phrases and sentences and its allegedly increasing usage, will be presented in the talk which can demonstrate the benefits of such a corpus design for the study of language continuity and language change.

REFERENCES

Ágel, Vilmos & Mathilde Hennig. 2006. Theorie des Nähe- und Distanzsprechens. In Vilmos Ágel & Mathilde Hennig (eds.). *Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650–2000*. Tübingen: Niemeyer, pp. 3–31.

Dürscheid, Christa & Stephan Elspaß. 2015. Variantengrammatik des Standarddeutschen. In Roland Kehrein, Alfred Lameli & Stefan Rabanus (eds.), *Regionale Variation des Deutschen. Projekte und Perspektiven.* Berlin, Boston: de Gruyter, pp. 563–584.

Elspaß, Stephan & Konstantin Niehaus. 2014. The standardization of a modern pluriareal language. Concepts and corpus designs for German and beyond. In: *Orð og tunga* 16/2014, pp. 47–67.

Scheible, Silke, Richard Jason Whitt, Martin Durrell, Paul Bennett. 2011. Investigating diachronic grammatical variation in Early Modern German. Evidence from the GerManC corpus. In Marek Konopka, Jacqueline Kubczak, Christian Mair, František Štícha & Ulrich H. Waßner (eds.), *Grammatik und Korpora 2009/Grammar & Corpora 2009. Third International Conference, Mannheim 22.-24.09.2009* (Corpus Linguistics and Interdisciplinary Perspectives on Language, 1). Tübingen: Narr, pp. 539–548.

Perception Verbs in Old Turkic: Text type, context and meaning extensions

Zeynep Erk Emeksiz & Buğra Oğuzhan Uluyüz
Anadolu University

This study aims at describing the challenges of studying on historical texts with a specific focus on the meaning extensions of the visual and auditory perception verbs *körmek* (to see), *bakmak* (to look) and *eşitmek* (to hear) and *tınglamak* (to listen) in the Old Turkic Period (OTP). Our observations are mainly based on the data collected from the inscriptions and texts of OTP consisting Köktürk and Yenisei Inscriptions, old Uighur texts, Kutadgu Bilig (KB), and Divanı Lügatit Türk (DLT), dating back approximately from 8th to 13th century (see Clauson,1973; Basin, 1991; Erdal 2004; Berta,2000; Ölmez, 2013 for a detailed analysis and discussion on the dates of the inscriptions and texts). The corpus of Old Uighur VATEC (http://vatec2.fkidg1.uni-frankfurt.de/) and Eski Türkçe Derlemi (http://derlem.cu.edu.tr/index.php?a=tarihsel/search) are also online sources to serach for the morphology and grammar of specific word strings. However, studying polysemy requires an analysis both on a discourse and text levels. These corpora may not be sufficient for such analysis since they provide only the co-text information. The translations provided by Tekin (1994), Rybatzki (1997), Berta (2000), Ölmez (2013) for Köktürk İnscription; Hamilton (1971,), Röhrborn (1991, 1996), Yakub (2010), Ayazlı (2012) for Old Uyghur texts; Dankoff (1985), Atalay (2006) for Divanı Lügatit Türk; and Dankoff (1983), Rahmeti (1979, 1997, 1998), Tezcan (1981), Ata (1993) for Kutadgu Bilig are also used to observe how the

polysemic meaning of the perception verbs were translated to Turkish, German, Russian and English.

Verbs of perception are receiving increased attention in cognitive and linguistic research. In cognitive terms, the relationship between perception and linguistic structure reveals clues about the neurolinguistic aspects of language production (Shipley and Zucks, 2008). The polysemy of perception verbs has been studied extensively, in Cooper (1974), Viberg (1984) Lehrer (1990), Sweetser (1990), Whitt (2010), and Gisborne (2010). These studies mainly focus on three aspects perception verbs: 1. The pattern that governs the derivation of sub meanings of these verbs from a typological (Viberg, 1984, Sweetser, 1990) and universal perpectives (Jackendoff, 1997; Langacker, 1991; Gisborne, 2010) 2. The semantic change of these verbs and their meaning extensions (Sweetser, 1990; Diewald & Smirnova, 2010; Whitt, 2010; Nakamura, 2010). 3. Their relation to evidentiality and epistemic modality (Aijmer, 2009; Grossman and Tutin, 2010).

There are many challenges of studying on historical texts such as lack of native speakers and sometimes discourse contexts. While studying the meaning of perception verbs, one may also face the diffuculty of determining the meaning extension of a given verb in a specific text type. This study will try to provide some linguistic parameters to overcome these challenges.

Körmek and *eşitmek* refer to [-control] verb with a patient subjects. *Bakmak* and *tınglamak*, on the other hand, refer to a [+control] verb with a agent subject. Köktürk inscriptions address the köktürk people and consist of the advises of Bilge Kagan (the sultan of köktürk People) and his commander Kül Tigin. They can be classified as a narrative of personal experience, as in Labov (1972), with first person narration. They are meant to counsel Turkic people from Chinese attacks. Hence each narratice segment ends with a deontic conclusion. We found that the verb *körmek* has two meaning extensions as ‘to grasp a vision without control’ and ‘to take care of people with control’ as a physical state verb most frequently in narrative discourse. However, the verb has a mental extension in the deontic mood as ‘grasping a fact’ corresponding to ‘realize’ in English. The verb *bakmak* does not occur in the inscriptions. We observe the same pattern in auditory verbs: *eşit* is used in the beginning of a narrative addressing the audience as ‘sözlerimi iyi ***eşidin***’ (listen to my words carefully). Considering the context (the sultan advices his people) we conclude that it is used as a control verb ‘listen to’ and has a meaning extension for obedience. We also found the linguistic pattern that when these verbs have sentential complemets they refer to mental state meanings. When there is a noun phrase as a complemet than we find the physical state meanings such as to see, to come across, to take care, etc.

Studying polysemy becomes more diffucult in prose form. Kutadgu Bilig is a poetic text. There are stylistic considerations such as metrics and rhym. There are lines that the writer used the verb *körmek* just for metrical reasons. (Tezcan, 1981; Ata, 1993). Arat translated *körmek* as ‘to pay attention’ and ‘to look’ into Turkish. Another challenge is that it is easier to find the distributive features, collacational patterns, and complementation patterns in narrative texts (whether it is the predicate of the subordinate clause etc.). However, the poet uses a non default syntax and it becomes harder to find the default word order and collocational relations.

REFERENCES

Aijmer, K. (2009). 'Seem' and evidentiality.. Functions of Language. 16 (1) s. 63-88.
Arat, R. Rahmeti (1979). Kutadgu Bilig III-İndeks. İstanbul: TKAE Yay.
_____, (1997). Kutadgu Bilig I-Metin. Ankara: TDK. Yay.
_____, (1998). Kutadgu Bilig II-Çeviri. Ankara: TTK. Yay.
Arat, R. Rahmeti (1991). Eski Türk Şiiri. 3. Baskı, TTK, Ankara.
Ata, A. (1993). "Kutadgu Bi lig Üzerinde bir Düzenleme Denemesi: kör mü? kür mü?". Türkoloji Dergisi 11.

Atalay, B. (2006). Kaşgarlı Mahmud Divanü Lugati't-Türk I-III, IV (Dizin), Ankara: TDK Yay.
Ayazlı, Ö. (2012). Altun Yaruk Sudur VI. Kitap. İstanbul: TDK Yayınları.
Aydarov, G. (1971): Yazık Orhonskih pamyatnikov drevnetyurkskoy pis'mennosti VIII veka. Alma-ata: Akademiya Nauk Kazakskoy SSR.
Aydın, E. (2011). Uygur Kağanlığı Yazıtları. Kömen Yayınları
Aydın, E. (2012). Orhon Yazıtları (Köl Tegin, Bilge Kağan, Tonyukuk, Ongi, Küli Çor). Kömen Yayınları
Bazin, L. (1991). Les systèmes chronologiques dans le monde turc ancien. Budapest: Akadémiai Kiadó.
Bazin, L. (1993). Quelques remarques d"Epigraphie Turque ancienne. Türk Dilleri Araştırmaları 3, 33–41.
Clauson, S. G. (1972): An Etymological Dictionary of Pre-Thirteenth-Century Turkish. Oxford: The Clarendon.
Dankoff, R. - KellyY, J. (1985). Mahmud al-Kaşgari: Compendium of the Turkic Dialects (Dîvân Lugât at-Turk),I-lll, Sources of Oriental Languages and Literatures 7, Türkish Sources V II, Cambridge.
Dankoff, R. (1983). Yûsuf Hâs Hâcib, Wisdonı of Royal Glory (Kutadgu Bilig), A Turko-Islamic Mirror for Princes, Chicago & London.
de Haan, F. (2007). 'Raising as grammaticalization: the case of Germanic Seem verbs'. Rivista di Linguistica 19.1. pp 129-150.
Diewald, G. and E. Smirnova. (2010). Linguistic realization of evidentiality in European Languages. Berlin-New York: Walter de Gruyter. Gisborne, N. (2010). The event structure of perception verbs. Oxford: Oxford University Press.
Emeksiz, E. Z. and Uluyüz, B. (2014). 'Evolution of Perception verbs in Turkish: Körmek'. 16th ICTL, University of Rouen, France.
Hamilton, J. Russell (2011). İyi ve Kötü Prens Öyküsü (Çev. Vedat Köken). TDK, 2. Baskı, Ankara
Labov, W. (1972). 'Some further steps in narrative analysis'. Language in the Inner City. Philadelphia: Univ. of Pennsylvania Press.
Nakamura, F. (2010). Uncovering of rare or unknown usages: a history of seem meaning to pretend. In M. Kytö (Ed.), Language change and variation from Old English to Late Modern English: A festschrift for Minoji Akimoto (pp. 217-238). Berne: Peter Lang.
Ölmez, M. (2012): Orhon-Uygur Hanlığı Dönemi Moğolistan'daki Eski Türk Yazıtları, Metin-Çeviri-Sözlük, BilgeSu yay., Ankara.
Renata, E. 2012. Les modalités de perception visuelle et auditive : Différences conceptuelles et répercussions sémantico-syntaxiques en espagnol et en français. Berlin-New-york: Walter de Gruyter.
Röhrborn, K. (1977-1996). Uigurisches Wörterbuch, Sprachmaterial der vorislamischen türkischen Texte aus Zentralasien: a-ärŋäk 1-6. Wiesbaden.
Rybatzki, V. (1997): Die Tonukuk-Inschrift. Szeged: Studia Uralo-Altaica.
Sweetser, E. E. (1990). From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure
Tekin, Ş. (1976). Uygurca metinler II, Maytrısimit, Burkancıların mehdisi Maitreya ile buluşma Uygurca iptidai bir dram (Burkancılığın Vaibhāşika tarikatine ait bir eserin Uygurcası), Ankara.
Tekin, T. (1994): Tunyukuk Yazıtı. Ankara: Simurg.
Tezcan, S. (1981). "Kutadgu Bilig Dizini Üzerine". TTK Belleten 178: 23-78.
Traugott, E. and Dasher, R.B. (2002). Regularity in semantic change. Cambridge: Cambridge University Press
Vasil'ev, D. D. (1983): Korpus Tyurkskih runiçeskih pamyatnikov basseyna Yeniseya. Leningrad: Akademiya Nauk SSSR.
Viberg, A., 1983. The verbs of perception: a typological study. Linguistics 21 (1), 123–162.
Whitt, R.J. (2010). Evidentiality and perception verbs in English and German. Frankfurt: Peter Lang Verlag.
Willett, T. (1988). A cross-linguistic survey of the grammaticalization of evidentiality. Studies in Language 12(1): 51–97.
Yakup, A. (2010), Prajñāpāramitā. Literature in Old Uyghur. (Berliner Turfantexte XXVIII.) Turnhout: Brepols, 2010, 319s. + 23 levha
Yılmaz, E. (2010): Sözlerimi İyi Dinleyin…, Türk ve Uygur Runik Yazıtlarının Karşılaştırmalı Yayını, (Árpád Berta, Szavaimat jól halljátok, Szeged 2004)

Using a parallel corpus of biblical translations to control for structural and stylistic factors in morphosyntactic variation: The case of Old Spanish possessives

Andrés Enrique-Arias
Universitat de les Illes Balears

This paper aims to demonstrate the methodological advantages of the *Biblia Medieval* corpus (www.bibliamedieval.es) in the historical study of stylistic variation in Medieval Spanish. Containing over five million words, *Biblia Medieval* is a freely accessible online tool that enables linguists to consult and compare side-by-side Old Spanish biblical translations. An interesting feature of the Bible is that it encompasses texts of varied textual typology: narrative, legislative, lyrical poetry, wisdom literature, epistles and dialogues. As a result, the *Biblia Medieval* corpus is particularly appropriate to explore register variation, as it is possible to examine how the same translator selects language options that are appropriate for each one of the genres represented in the Bible.

In order to illustrate the possibilities that *Biblia Medieval* offers for the analysis of stylistic variation, this paper uses as a case study the variation in the distribution of the definite article preceding the possessive marker (*la mi casa* 'the my house', henceforth ART+POSS) as opposed to possessive alone (*mi casa* 'my house', henceforth POSS) in Old Spanish. This structure, which was quite common throughout the Middle Ages declined and all but disappeared from the standard variety in the 1400s.

The appearance of ART+POSS as opposed to POSS has been attributed to a considerable number of structural factors, such as person and number of the possessor entity, and animacity of both the possessor and the possesed entity. At the same, the use of this structure is favored by stylistic factors: because it is a structure that emphasizes possession, it is used with stylistic functions such as expressivity, solemnity, poeticality or reverence (Lapesa 2000 [1971]: 422). As a consequence, when using a conventional corpus, it is rather complicated to control for all the possible factors that may be conditioning the variation observed in each particular text. In contrast, comparisons of this kind are relatively more straightforward in a parallel corpus like *Biblia Medieval*: as parallel texts put the discourse contextual factors largely in control, the behaviour of the elements used to express possession can be observed and compared in a focused manner.

This study demonstrates the usefulness of *Biblia Medieval* by providing an in-depth analysis of the developments in the expression of possession in Old Spanish. Over 4000 examples of possessive structures are extracted from the corpus and are analyzed quantitatively taking into account all of the structural factors considered in the literature; in addition the analysis focuses on how the behavior of ART+POSS varies across different stylistic parameters such as text types (narrative vs. lyrical) or type of discourse (direct speech vs. narration). The results show that the use of ART+POSS is more probable in stylistically marked registers such as lyrical passages or direct speech, and that in the 15th century, when ART+POSS is falling into disuse, stylistic parameters replace structural parameters as important predictors of the opposition between POSS and ART+POSS.

The use of computational and statistical methods as applied to the analysis of data from multi-and single-genre diachronic corpora

Fabrizio Esposito
University of Naples Federico II

Pierpaolo Basile
University of Bari

Marco Venuti
University of Catania

Multi-genre diachronic corpora are nowadays widespread across the (corpus) linguistics community. These resources, developed both in English (The Helsinki Corpus[1],

[1] Kytö M.(1991), *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*, Department of English, University of Helsinki.

COHA[2]) and in other languages (Onelli et al. 2006, Sánchez-Martínez 2013), represent "collection[s] of texts that vary along the parameter of time" (Hilpert & Gries 2009: 386). The intrinsic nature of this kind of corpora led to put the quantitative approach into the foreground (Biber & Jones 2008). Evert and Baroni (Evert & Baroni 2005, Baroni & Evert 2008) applied and estimated statistical models of word frequency distributions. Hilpert and Gries, focusing on "multistage diachronic corpora" (Hilpert & Gries 2009: 385), proposed analytic strategies to interpret ambiguous and messy data on objective grounds and identify stages in diachronic data (Gries & Hilpert 2008).

While the aforementioned statistical approaches investigate the purely lexical level, in the present contribution we instead propose the use of the lexical-semantic approach of Distributional Semantic Models (DSMs) to analyse multi-genre diachronic corpora. The core idea behind any distributional semantic model is based on Zelig Harris' *distributional hypothesis* (Harris 1954), stating that the lexical meaning of a word depends on its distributional properties. This notion of word meaning has survived in corpus linguistics endorsed by J.R. Firth's famous quotation "You shall know a word by the company it keeps" (Firth 1957: 11) and it has often been regarded as "the unique possible source of evidence for the exploration of meaning" (Lenci 2008: 6).

DSMs, as computational tools, build geometrical spaces of concepts known as *Word Spaces* where words are represented as vectors and their semantic similarity is expressed by the spatial proximity between them. In this contribution we use a specific framework developed by Basile et al. (Basile et al. 2014), known as Temporal Random Indexing (TRI), that allows one to analyse word meaning change over time (Sahlgren 2005, Sahlgren 2006). The TRI is employed on the CompWhob (Computational White House press Briefings) Corpus, currently being developed at the University of Naples Federico II. The CompWhob is a diachronic corpus spanning from 1993 to 2014 collecting the transcripts of the White House Press Briefings. It focuses on political and media communication genre, considering the briefings as an "informal institutional genre" (Spinzi andVenuti 2013: 183; Partington 2003, Partington 2006). Our aim is to prove that TRI is a robust "easily adaptable [tool]" (Basile et al. 2014: 39) applicable to the analysis of a multi-stage and multi-genre corpus, as recent studies have also revealed (Brigadir et al. 2015; Rule et al. 2015). In more detail, we will focus on those moments of political life where the administration has to cope with risky and serious scenarios, the so-called *crisis communication management*. After exploring word meaning variation across these specific time-spans, word neighbourhood will be then analysed in order to understand how the detected change relates to this specific phenomenon.

REFERENCES

Baroni M., Evert S.(2008), Statistical methods for corpus exploitation, in Lüdeling K., KytöM. (eds.), *Corpus Linguistics: An International Handbook.* Vol. 2, Berlin/New York: Mouton de Gruyter, pp 777-803

Basile P., Caputo A., Semeraro G. (2014), Analysing Word Meaning over Time by Exploring Temporal Random Indexing, in Basili R, Lenci A., Magnini B. (eds.), *First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa: Pisa University Press

Biber D., Jones J. K.(2008), Quantitative methods in corpus linguistics, in LüdelingK., KytöM. (eds.), *Corpus Linguistics: An InternationalHandbook.* Vol. 2, Berlin/New York: Mouton de Gruyter, pp 1287-1304

Brigadir I., Greene D., Cunningham P. (2015), Analyzing Discourse Communities with Distributional Semantic Models, in *ACM Web Science 2015 Conference*, 28 June –1 July 2015, Oxford, UnitedKingdom, available at http://websci15.org/, Accessed August 7, 2015

Evert S., Baroni M.(2005), Testing the Extrapolation Quality of Word Frequency Models, in *Proceedings from the Corpus LinguisticsConference Series*, Vol. 1, no. 1

Firth J. R. (1957), Papers in Linguistics, London: Oxford University Press

[2] Davies M.(2010-), *The Corpus of Historical American English: 400 million words, 1810-2009.*

Gries S. Th., Hilpert M. (2008), The identification of stages in diachronic data: variability-based neighbour clustering, in *Corpora*, Vol. 3, Issue 1, pp 59-81

HarrisZ.S. (1954), Distributional Structure, *Word*, 10/2-3, pp 146-162, reprinted in Harris Z. S. (1970),Papers in Structural and Transformational Linguistics, Dordrecht: Reidel, pp 775-794

Hilpert M., Gries S. Th.(2009), Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition, in *Literary and Linguistic Computing*, Vol. 24, no. 4

Lenci A. (2008), Distributional semantics in linguistic and cognitive research, in *Rivista di Linguistica*, Vol. 20, no. 1, pp 1-31

Onelli C., Proietti D., Seidenari C., Tamburini F.(2006), The DiaCORIS project: a diachronic corpus of written Italian, in *Proceedings of the 5thInternational Conference on Language Resources and Evaluation-LREC 2006*, Genova, pp 1212-1215

Partington A. (2003), The linguistics of Political Argument: The Spin-doctor and the Wolf-pack at the White House, London: Routledge

Partington A. (2006). The Linguistic of Laughter: A Corpus-Assisted Study of Laughter-talk, London: Routledge

Rule A., Cointet J., Bearman P. S.(2015), Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790-2014, in *PNAS September 1, 2015*, Vol. 112, no. 35, pp 10837-10844, available at http://www.pnas.org/cgi/doi/10.1073/pnas.1512221112, Accessed September 15, 2015

Sahlgren M. (2005), An Introduction to Random Indexing, in *Methods and Applications of Semantic Indexing Workshop at the 7thInternational Conference on Terminology and Knowledge Engineering*, TKE, Vol. 5

Sahlgren M. (2006), The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces, *Ph.D. Thesis*, Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics

Sánchez-Martínez F., Martínez-Sempere I., Ivars-Ribes X., Carrasco R. C. (2013), An open diachronic corpus of historical Spanish, in *Language Resources and Evaluation*,Vol.47,Issue 4, pp 1327-1342

Spinzi C., Venuti M. (2013), Tracking the change in an Institutional Genre: A Diachronic Corpus-based study of White House Press Briefings, Newcastle upon Tyne: Cambridge Scholars Publishing, pp 182-197

Proclaiming the Royal Prerogative: creating the Tudor Royal Proclamations (1509-1603) Corpus

Melanie Evans
University of Leicester

Tudor royal proclamations convey the royal prerogative, capturing the socio-political concerns of their age. As texts they appear to constitute a distinct genre: written documents printed for display and read aloud by an official, designed to construct and propagate the sovereign's power. But despite their social and linguistic significance, proclamations have received relatively little attention since Hughes & Larkin's (1964-9) print edition and the monographs by Heinze (1976) and Young (1976). This paper introduces and seeks feedback on a new single-genre corpus of Tudor Royal Proclamations, currently in preparation, which will enable linguists and historians to investigate this political genre using corpus techniques. I give a taster of some early linguistic findings pertaining to their key stylistic properties and highlight evidence of diachronic stylistic changes over the sixteenth century.

In the first part of the presentation, I explain how the data for the corpus has been collected using a combination of print editions, EEBO-TCP transcriptions, and transcripts of archived manuscript drafts, and reflect on the successes and difficulties encountered when using these resources. I describe the preparation process, including the mark-up protocols we have developed for this text-type (which is not, unsurprisingly, listed in the TEI guidelines), and the approximate size, diachronic span, and associated meta-data of the corpus when complete. In particular, I discuss the considerations that arise when developing a new corpus

that aims to serve multidisciplinary audiences, from linguists to literary scholars and historians. I give an account of the elements we have so far identified as essential for satisfying the diverse needs of this broad audience.

In the second part, I present some initial findings on the stylistic markers of royal proclamations, using stylometric (*Intelligent Archive*) and concordance (*AntConc*) software. I provide an account of the genre's key features (linguistic and typographical), and suggest how these show variation in accordance with particular monarchic periods (i.e. Henry VIII or Mary I), subject-matter, and royal printer. I then provide a broad account of how the stylistic features compare with other royal texts (e.g. letters, dramatic dialogue) in the sixteenth century, and how their style fits with existing narratives of the developmental history of Early Modern English.

REFERENCES

Anthony, L. 2014. AntConc (Version 3.4.4) Windows. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Centre for Literary and Linguistic Computing (CLLC) at the University of Newcastle Australia. 2015. Intelligent Archive 2.0 Corella. Available from http://www.newcastle.edu.au/research-and-innovation/centre/education-arts/cllc/intelligent-archive

Heinze, R. 1976. *The Proclamations of the Tudor Kings*. Cambridge: CUP

Hughes, P. and J. Larkin. 1964-9. *Tudor Royal Proclamations* (3 voles). New Haven and London: Yale University Press

Youngs, F. 1976. *The Proclamations of the Tudor Queens*. Cambridge: CUP

VP fronting in Old Icelandic: A case of genre-specific syntax

Thorhallur Eythorsson & Sirry Sigurdardottir

University of Iceland / *Ghent University*

This paper deals with VP fronting, which occurs to a limited extent in Old Icelandic poetry. A corpus-based study reveals that it is a genre-specific syntactic phenomenon in Old Icelandic. It is argued that it is nevertheless important for a more general linguistic analysis of the language, involving a complex interplay of syntax, meter and information structure.

VP fronting is well known from modern Germanic languages, e.g. German, Mainland Scandinavian and English, as in (1).

(1) John wanted to read a book and **read a book** he did

The VP, consisting of a main verb (V) and an object or some other type of complement or adjunct (O), does not occur in its base position at the end of the clause; rather, it is fronted (topicalized) to the beginning of the clause, to the left of the subject and a finite auxiliary. On the other hand, VP fronting does not occur in Icelandic at all, neither in the modern language nor in Old Icelandic prose (Rögnvaldsson 1995, Thráinsson 2007). Various other types of fronting occur in Icelandic, however, e.g. Stylistic Fronting, targeting subparts of the VP (Maling 1980).

Furthermore, VP fronting is almost totally absent from Old Icelandic alliterative poetry, both the Poetic Edda and skaldic verse, the latter being notorious for its "free" word order. The only exception is found in poetry composed in a type of traditional meter called *ljóðaháttr* ("ballad meter"), consisting of two-line stanzas with two lifts and one line with three lifts. Recent corpus-based research (*Greinir skáldskapar*, *IcePaHC*) shows that there are more than twenty examples of VP fronting in the Poetic Edda, as in (2), and more cases are found other *ljóðaháttr* poetry outside of the eddic poems.

(2) a. H**öfuð höggva** / ek mun þér hálsi af (Skírnismál 23,5)

head-ACC hew I will you-DAT neck off
'I will cut the head of your neck'

b. **Upp líta** / skal-at-tu í orrosto (Hávamál 129,6)
up look shall-not-you in battle
'You shall not look up in battle'

Although very limited, VP fronting is clearly a characteristic of this particular type of poetry, a special genre-specific device of poetic syntax. Thus, even though VP fronting is not a feature of "normal" Old Icelandic syntax, its presence in the poetry constitutes indipendent evidence in favor of a VP consitutent in the language, a matter that has been of longstanding debate (Rögnvaldsson 1995, Faarlund 2004).

Finally, the question arises if a motivation for the VP fronting in Old Icelandic poetry can be identified. It is concluded that it is indeed possible to pinpoint the empirical conditions under which VP fronting occurs, as they involve the interaction of syntax, meter and information structure which are general characteristics of the *ljóðaháttr* poetry.

Pragmatic features in two single-genre corpora of Classical French

Annette Gerstenberg & Carine Skupien-Dekens
Freie Universität Berlin / Université de Neuchâtel

Our contribution concerns two corpora comprised of non-fictional texts representing genres that, until now, have not been sufficiently studied from a diachronic linguistic perspective. We will discuss the notion of "genre" applicable to the represented text types and the characteristic pragmatic features. The aim of our contribution is to show the evolution of a genre through a longitudinal approach and examine how the comparison of different genres from the same period can help to develop a more fine-grained analytical scheme applicable with the methods of corpus linguistics.

The corpora presented, aside from their differences, have the similarity of being persuasive in nature and originating from highly performative contexts, such as politics and religion. They are deeply embedded in traditions of learned elocution, but at the same time they are linked to specific communicative contexts. First, the SERMO-corpus is composed of sermons dating from the middle of the sixteenth century to the middle of the eighteenth century, allowing for the application of a longitudinal approach. Second, APW-corpus is composed of diplomatic letters that were exchanged between diplomats in Münster (Germany) and the French government in Paris during the negotiations preceding the Peace of Westphalia (1643–1648; 2,4 Mio tokens), which allows for an in-depth synchronic analysis.

We will present both corpora in order to focus on the genre-marking features of "sermons" and "diplomatic letters" respectively, that is, on features of situation, content, and pragmatic function. The detailed overview of these high-level features will be integrated by an exemplification of the internal heterogeneity of both corpora, which is due to differences in the social and geographical origins of the various authors of the studied sermons and letters. Furthermore, on the level of discourse functions, the sermons and diplomatic letters are variegated (Kohnen, 2010; Palander-Collin, 2010).

In our analysis, we will show which analytical tools of the corpus linguistic framework can help to identify the portions within the corpora that correspond to these different functions. We will focus on personal pronouns:
(1) the frequency of personal pronouns (1. and 2. sg. and pl.) is linked to the less or more direct situational reference realized in the text of the corpus.

(2) distributional features help to identify portions within the corpus with a stronger presence of deictic references.
(3) the analysis of the predicates with personal pronouns in the subject position strengthens the notion of formulaic vs. individualistic genre-specific features. The analysis revealed various benefits from the tools for morphosyntactic tagging and lemmatization applied to our corpora. In the work chain of semi-automatic annotation (manual disambiguation, cleaning), we respected the limits of the existing state-of-the-art tools used for speech tagging and lemmatization. As a result of the analysis, we will propose a tagset on the level of pragmatics.

In the summary, we discuss the results of the longitudinal approach (SERMO-corpus) and the results from the comparison of both corpora (SERMO- and APW-corpus) in order to underline the similarities in the two non-fictional genres of persuasion.

REFERENCES

Beacco, Jean-Claude. 2004. Trois perspectives linguistiques sur la notion de genre discursif. *Langages* 153(2). 109–119.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language* 65(3). 487–517.

Diwersy, Sascha, Achille Falaise, Marie-Hélène Lay, Gilles Souvay & Denis Vigier. [2014]. *Modèle panchronique d'étiquetage morphosyntaxique pour le français (16e–20e siècles)*. Köln, Nancy.

Gerstenberg, Annette. 2014. Diskursive Spezialisierung infiniter Verbformen in den französischen Texten der APW. In Annette Gerstenberg (ed.), *Verständigung und Diplomatie auf dem Westfälischen Friedenskongress. Historische und sprachwissenschaftliche Zugänge*, 172–194. Köln, Weimar, Wien: Böhlau.

Gerstenberg, Annette. in print. Les sources françaises des Acta Pacis Westphalicae : Approches linguistiques et la construction du corpus APWCF. In Wendy Ayres-Bennett (ed.), *L'histoire du français: Nouvelles approches, nouveaux terrains, nouveaux traitements*. Paris: Garnier.

Herring, Susan C. Pieter van Reenen & Lene Schøsler. 2000. On Textual Parameters and Older Languages. In Susan C. Herring, Pieter T. v. Reenen & Lene Schøsler (eds.), *Textual parameters in older languages* (Amsterdam studies in the theory and history of linguistic science), 1–31. Amsterdam, the Netherlands, Philadelphia, PA: John Benjamins Pub. Co.

Jucker, Andreas H. 2010. Historical Pragmatics. In Mirjam Fried, Jan-Ola Östman & Jef Verschueren (eds.), *Variation and change*: *Pragmatic perspectives* (Handbook of pragmatics highlights v. 6), 110–122. Amsterdam, the Netherlands, Philadelphia, PA: John Benjamins Pub. Co.

Kohnen, Thomas. 2010. Religious Discourse. In Andreas H. Jucker & Irma Taavitsainen (eds.), *Historical Pragmatics* (Handbook of pragmatics 8), 524–547. Berlin, New York: Mouton de Gruyter.

Palander-Collin, Minna. 2010. Correspondence. In Andreas H. Jucker & Irma Taavitsainen (eds.), *Historical Pragmatics*, 2nd edn. 651–678. Berlin, New York: Mouton de Gruyter.

Skupien-Dekens, Carine. in print. Un genre sous-exploité en histoire du français pré-classique et classique: le sermon. In Wendy Ayres-Bennett (ed.), *L'histoire du français: Nouvelles approches, nouveaux terrains, nouveaux traitements*. Paris: Garnier.

Skupien-Dekens, Carine. 2014. Reste-t-il des marques de l'oral dans les sermons de Calvin? In Dorothée Aquino-Weber, Federica Diémoz, Laure Grüner & Aurélie Reusser-Elzingre (eds.), *Toujours langue varie…*: *Mélanges de linguistique historique du français et de dialectologie galloromane offerts à M. le Professeur Andres Kristol par ses collègues et anciens élèves* (Recueil de travaux publiés par la Faculté des lettres et sciences humaines de l'Université de Neuchâtel 59), 83–97. Genève: Droz.

Tischer, Anuschka. 1999. *Französische Diplomatie und Diplomaten auf dem Westfälischen Friedenskongress. Außenpolitik unter Richelieu und Mazarin*. Münster: Aschendorff.

Studying genres in order to understand Greek diglossia in the 20th century

Dionysis Goutsos & Georgia Fragaki
University of Athens

The treatment of Greek diglossia – the parallel use of a High (katharevousa) and a Low (demotic) linguistic code in the language's recent past – has been characterized by an

almost exclusive focus on linguistic attitudes rather than the investigation of actual use. The lack of evidence from use seriously impinges on the question of standardization of Modern Greek, as well as on the broader question of its actual diachronic development. This paper aims at studying evidence from a number of genres in a diachronic corpus of Greek in order to gain a more informed understanding of the language's history in the 20th century.

The corpus consulted for this purpose is the *Diachronic Corpus of Greek of the 20th century* (*Greek Corpus 20*), developed at the University of Athens for the study of recent language change in Greek. *Greek Corpus 20* includes a variety of genres of 20th century Greek from the 1900s to the 1980s, designed to be integrated with the existing synchronic 30 million word *Corpus of Greek Texts* (Goutsos 2010), which includes data from 1990 to 2010.

In this paper we analyze approximately 3.5 million words from seven different genres of the corpus, namely spoken news (newsreels), public speeches, film scripts, literature, academic texts, private letters and legal-administrative texts. We compare the frequencies of High vs. Low variants of several grammatical items across these seven genres and along the nine decades.

Our preliminary findings suggest that recent language change in Greek largely depends on genre. Specifically, in film scripts and literature there is steady preference for Low variants across the century. By contrast, in academic texts and public speeches High variants are preferred in most decades before the 1960s, when there is a sudden rise of Low variants. Newsreels show a haphazard pattern, conforming to the expected rise of Low variants only after the 1960s, whereas private letters are the only genre in which the expected gradual rise of Low variants across all decades is found.

These findings have important implications for both the study of Greek diglossia and the role of genres in recent language change, as has been found in other languages, e.g. with regard to the contribution of speech-like genres (Culpeper & Kytö 2010) and especially private letters (cf. Dossena & Del Lungo Camiciotti 2012), as well as with regard to diachronic corpora in general (Taavitsainen et al. 2015).

REFERENCES

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

Dossena, M. & Del Lungo Camiciotti. G. 2012. *Letter Writing in Late Modern Europe*. Amsterdam/Philadelphia: Benjamins.

Goutsos, D. 2010. The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora* 5 (1), 29-44.

Taavitsainen, I., Kytö, M., Claridge, C. & Smith, J. (2015). *Developments in English: Expanding Electronic Evidence*. Cambridge: Cambridge University Press.

"You can't control a thing like that": Modern English human impersonal pronouns and changing 'oral' genres

Florian Haas
Friedrich-Schiller-Universität Jena

Human impersonal pronouns have the function of introducing non-specific, often generic, participants, avoiding the identification of specific individuals (often the speaker him-/herself) for different reasons. In this paper, I will report on corpus-based research tracing changes in the distribution of impersonal uses of Modern English *one*, *you* and *they* (cf. [1]-[3] and the example in the title above [ARCHER, 1938mccr.d7b]).

(1) Also when sleeping in woods **one** wakes very soon, and I woke when the dawn came through the chinks of the tent on to my face. [ARCHER, 1957maca.f8b]

(2) **You** can't understand a thing the bloody man says. [ARCHER, 1973trev.f8b]

(3) The workhouse where **they** put me. **They** beat you there like a drum. [ARCHER, 1979pomr.d8a]

There is substantial work on impersonal reference in Old, Middle and Early Modern English (Fröhlich 1951; Meyer 1953; Jud-Schmid 1956; Rissanen 1997; Seoane 2000), and it is clear that the demise of the dedicated human impersonal pronoun *man* in late Middle English has led to some degree of reorganization (Jud-Schmid 1956; van Gelderen 1997; Los 2006). In order to find out more about the long-term consequences of this turning point, all instances of *one*, *you* and *they* were extracted from the ARCHER corpus (*A Representative Corpus of Historical English Registers* 3.1) and manually coded as impersonal or personal. In addition to information about registers and diachronic stages, more specific lexical and morphosyntactic features of the data points have been coded in order to explore changes in usage conditions. Results indicate that, whereas some patterns of distribution and the general availability of these forms for impersonal reference have been well-established for the entire period covered by the corpus (i.e. Modern English), the overall frequency of impersonal *one* and *you* has risen after 1850. In addition, those subtypes of impersonal *you* that deviate more from its canonical, deictic use ('simulation', as opposed to 'generalization') become more common in these later stages, too. These developments may be based on "intimization and personification processes", as has been suggested for Danish second person impersonals by Nielsen et al. (2009; cf. also Traugott 2003 on 'intersubjectification'). Yet, whether these data demonstrate a reorganization of impersonalization strategies in English in the first place crucially depends on what we know about the diachrony of English registers in general. Biber and colleagues have shown that the degree to which characteristics of spoken language are represented in writing has increased for 'spoken' registers (cf. Biber & Finegan 1989, 1997). I will discuss attempts at distinguishing the observed changes in the distribution of impersonals from developments that English registers as such have gone through. This includes testing in how far the shifting usage conditions of impersonals are statistically and conceptually independent of those features that for Biber et al. indicate changes in registers like drama, letters, fiction and diaries.

REFERENCES

Biber, Douglas and Edward Finegan (1989). Drift and the evolution of English style: A history of three genres. *Language* 65: 487-517.

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: Studies in the changing English language in honour of Matti Rissanen*, Berlin: de Gruyter, 253–275.

Fröhlich, Jürg (1951). *Der indefinite Agens im Altenglischen, unter besonderer Berücksichtigung des Wortes* man. Winterthur-Töß: Paul Gehring.

Jud-Schmid, Elisabeth (1956). *Der indefinite Agens von Chaucer bis Shakespeare. Die Wörter und Wendungen für* man. Meisenheim am Glan: Anton Hain.

Los, Bettelou (2006). *The Rise of the* To*-Infinitive*. Oxford: Oxford University Press.

Meyer, Hans-Heinrich (1953). *Der indefinite Agens im Mittelenglischen (1050-1350). Wörter und Wendungen für* man. Bern: Francke.

Nielsen, Søren, Christina Fogtmann and Torben Juel Jensen (2009). From community to conversation -- and back: Exploring the interpersonal potentials of two generic pronouns in Danish. *Acta Linguistica Hafniensia* 41: 116-142.

Rissanen, Matti (1997). Whatever happened to the Middle English indefinite pronouns? In: Jacek Fisiak (ed.), *Studies in Middle English Linguistics*, Berlin: de Gruyter, 513-529.

Seoane Posse, Elena (2000). Impersonalizing strategies in Early Modern English. *English Studies* 81: 102-116.

Traugott, Elizabeth Closs (2003). From subjectification to intersubjectification. In: Raymond Hickey (ed.) *Motives for Language Change*, Cambridge: Cambridge University Press, 124-139.

van Gelderen, Elly (1997). *Verbal Agreement and the Grammar behind its 'Breakdown': Minimalist feature checking*. Berlin: de Gruyter.

Textual distribution of verbal free adjuncts in the recent history of English

Carla Bouzada-Jabois
University of Vigo

Free adjuncts, like (1) below, are subjectless, generally, nonfinite constructions occurring at the periphery of the clause or even interrupting the main clause at some point. Syntactically, free adjuncts are independent due to their lack of integration in the main clause. Semantically, they usually establish referential links to the main clause and express some kind of adverbial meaning affecting the clauses to which they are attached.

(1) He walk in and out among the people, and lay the skin at the feet of the witch, **singing all the while**. (READE-1863,213.272)

Free adjuncts have been described as stylistically marked constructions occurring in formal, written and narrative genres more often than in spoken language (Thompson 1983: 45, Kortmann 1991: 2, Río-Rey 2002: 313). Likewise, these constructions are said to be common in discourse types that aim at describing events (Thompson 1983: 46).

This paper investigates verbal *–ing* and *–ed* free adjuncts, which constitute the most frequent categorial subtypes (Kortmann 1991, 1995), and analyses their textual distribution from Late Modern English to Present-day English in an attempt to corroborate their textual preference for formal narrative text types. The Penn Parsed Corpus of Modern British English and the British component of the International Corpus of English have been selected as main sources of exploration. Comparison with previous studies on the topic (Kortmann 1991 for Present-day English, and Río-Rey 2002, 2004 for Early Modern English) will also be established for a more detailed analysis. Problematic issues such as genre classification and comparison in the corpora will be neutralized as far as possible by establishing a more general taxonomy based on Culpeper and Kytö (2010). The data show that free adjuncts evince genre dependency and confirm the trends suggested in the relevant literature. In particular, this study confirms that the productivity of free adjuncts in written-related texts overtakes, by far, the use of these constructions in speech-related genres. As regards specific text types, those genres of a more narrative type seem to accommodate most of the instances of FAs in the database.

REFERENCES

Culpeper, Jonathan and Merja Kytö. 2010. *Early modern English dialogues: spoken interaction as writing*. Cambridge: Cambridge University Press.

Kortmann, Bernd. 1991. *Free adjuncts and absolutes in English: problems of control and interpretation*. London: Routledge.

Kortmann, Bernd. 1995. Adverbial participial clauses in English. In Martin Haspelmath and Ekkehard König eds. *Converbs in cross-linguistic perspective: structure and meaning of adverbial verb forms – adverbial participles, gerunds*. Berlin: Mouton de Gruyter, 189-237.

Río-Rey, Carmen. 2002. Subject control and coreference in Early Modern English free adjuncts and absolutes. *English Language and Linguistics* 6/2: 309-323.

Río-Rey, Carmen. 2004. Free adjuncts and absolutes in Early Modern English: some text type considerations. In Mª Luisa Pascual Garrido, Pilar Guerrero Medina, Carmen Portero Muñoz and Antonio Ruíz Sanchez eds. *Estudios de Filología Inglesa: Actas de las IV Jornadas de Filología Inglesa*. Córdoba: Servocio de Publicaciones Universidad de Córdoba, 167-175.

Thompson, Sandra A. 1983. Grammar and discourse: the English detached participial clause. In Flora Klein-Andreu ed. *Discourse perspectives on syntax*. New York: Academic Press, 43-65.

Diachronic Collocations and Genre

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften

Abstract
This paper outlines some potential applications of the open-source software tool "DiaCollo" to multi-genre diachronic corpora. Explicitly developed for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus, DiaCollo – unlike conventional collocation extractors – is suitable for processing collocation pairs whose association strength depends on the date of their occurrence. By tracking changes in a word's typical collocates over time, DiaCollo can help to provide a clearer picture of diachronic changes in the word's usage, especially those related to semantic shift or discourse environment. Use of the flexible DDC search engine back-end allows user queries to make explicit reference to genre and other document-level metadata, thus allowing e.g. independent genre-local profiles or cross-genre comparisons. In addition to traditional static tabular display formats, a web-service plugin also offers a number of intuitive interactive online visualizations for diachronic profile data for immediate inspection.

1 Introduction
DiaCollo is a software tool for automatic *collocation profiling* (Church and Hanks, 1990; Evert, 2005) in diachronic corpora such as the *Deutsches Textarchiv*[3] (Geyken et al., 2011) or the Corpus of Historical American English[4] (Davies, 2012) which allows users to choose the granularity of the diachronic axis on a per-query basis (Jurish, 2015). Unlike conventional collocation extractors such as DWDS Wortprofil[5] (Didakowski and Geyken, 2013) or Sketch Engine[6] (Kilgarriff and Tugwell, 2002), DiaCollo is suitable for extraction and analysis of diachronic collocation data, i.e. collocate pairs whose association strength depends on the date of their occurrence and/or other document-level properties such as author or genre.

2 Implementation
DiaCollo is implemented as a Perl library, and provides both a command-line interface as well as a modular RESTful web service plugin (Fielding, 2000) with a form-based user interface for evaluation of runtime database queries and interactive visualization of query results. For finer-grained selection of profiling targets, DiaCollo supports the full range of the DDC[7] query language (Sokirko, 2003; Jurish et al., 2014) whenever the DiaCollo instance is associated with an underlying DDC server back-end. In particular, use of the DDC back-end allows explicit reference to all document-level metadata encoded in the corpus, including e.g. text genre for the *Deutsches Textarchiv* DiaCollo index accessible at http://kaskade.dwds.de/dstar/dta/diacollo/.[8]

[3] http://www.deutschestextarchiv.de
[4] http://corpus.byu.edu/coha
[5] http://zwei.dwds.de/wp
[6] http://www.sketchengine.co.uk
[7] http://www.ddc-concordance.org
[8] For faster processing, arbitrary token- and/or document-attributes from the source corpus can be selected for inclusion in DiaCollo's native index structure at index compilation time. The default configuration includes only the token attributes *Lemma* and *Pos* ("part-of-speech") in its native index.

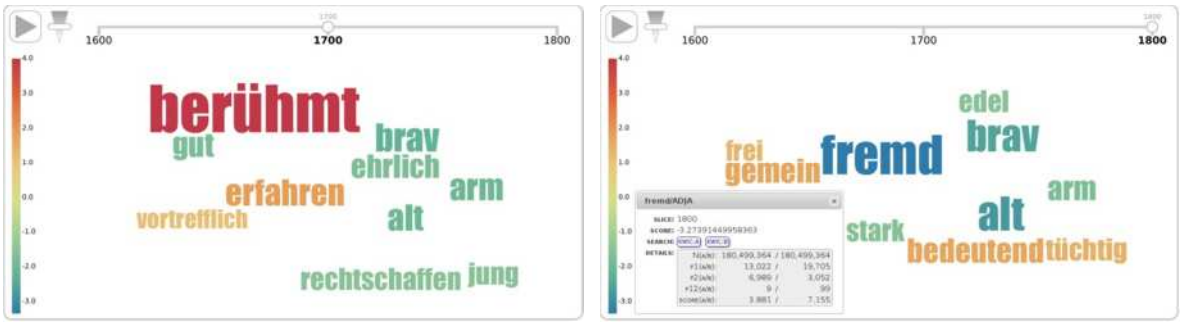


Figure 1: DiaCollo interactive tag-cloud visualization of the $k = 10$ most strongly divergent adjectives immediately preceding the noun *Mann* (“man”) in the genres “science” (warm colors) and “*belles lettres*” (cool colors) over the *Deutsches Textarchiv* corpus for the epochs 1700–1799 (left) and 1800–1899 (right).

3 Example

Figure 1 contains example tag-cloud visualizations for a cross-genre comparison over the *Deutsches Textarchiv* corpus. The DDC back-end was used to acquire raw frequency counts over 100-year epochs for all adjectives in immediately preceding the noun *Mann* (“man”) in the genres *Wissenschaft* (“science”) and *Belletristik* (“belles lettres”), respectively. After computing association scores (Evert, 2008; Rychlý, 2008) for each such candidate collocate, the DiaCollo engine extracts and returns the *k* collocates in each epoch with the greatest absolute score differences. In the tag-cloud visualization mode, absolute score differences are mapped to tag font-size, and the signs of the score differences are mapped to an intuitive color-scale, with warm tones indicating a relative preference for the “science” genre and cool tones indicating a preference for “belles lettres”. As Figure 1 shows, men in scientific texts are more likely to be described as *berühmt* (“famous”), *erfahren* (“experienced”), *bedeutend* (“significant”), or *tüchtig* (“capable”); while men in belles lettres are more likely to be designated *brav* (“wellbehaved”), *rechtschaffen* (“righteous”), *arm* (“poor”), *alt* (“old”) – assumedly reflecting the properties considered most salient in the context of the respective genres.

4 Conclusion

A new software tool “DiaCollo” for the efficient extraction, comparison, and interactive online visualization of collocations was introduced. In its top-level incarnation as a modular web service plugin, DiaCollo provides a simple and intuitive interface for assisting linguists, lexicographers, and humanities researchers to acquire a clearer picture of variation in a word’s usage over time and/or corpus subset. Use of either the flexible DDC search engine back-end or compile-time index-attribute selection allows user queries to make explicit reference to genre and other document-level metadata, thus allowing cross-genre diachronic comparisons, as demonstrated on the basis of a simple example. Publicly accessible DiaCollo web-service instances exist for a number of German corpora hosted by the DWDS project at the Berlin-Brandenburg Academy of Sciences, and the DiaCollo source code itself is available via CPAN at http://metacpan.org/release/DiaColloDB.

REFERENCES

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

M. Davies. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157, 2012. URL http://davies-linguistics.byu.edu/ling450/davies_corpora_2011.pdf.

J. Didakowski and A. Geyken. From DWDS corpora to a German word profile – methodological problems and solutions. In A. Abel and L. Lemnitzer, editors, *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*, (OPAL X/2012).

IDS, Mannheim, 2013. URL http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikografie_2012_final.pdf.

S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut f¨ur maschinelle Sprachverarbeitung, Universit¨at Stuttgart, 2005. URL http:// elib.uni-stuttgart.de/opus/volltexte/2005/2371/.

S. Evert. Corpora and collocations. In A. L¨udeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, 2008. URL http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_ manuscript.pdf.

R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000. URL https://www.ics.uci.edu/˜fielding/pubs/dissertation/top.htm.

A. Geyken, S. Haaf, B. Jurish, M. Schulz, J. Steinmann, C. Thomas, and F. Wiegand. Das deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In S. Schomburg, C. Leggewie, H. Lobin, and C. Puschmann, editors, *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pages 157–161, 2011. URL http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_ Wissenschaft.pdf# age=159.

B. Jurish. DiaCollo: On the trail of diachronic collocations. In K. De Smedt, editor, *CLARIN Annual Conference 2015 (Wro law, Poland, October 14–16 2015)*, pages 28–31, 2015. URL http://www.clarin.eu/sites/default/files/book%20of%20abstracts% 202015.pdf.

B. Jurish, C. Thomas, and F. Wiegand. Querying the deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, and C. Gurrin, editors, *Proceedings of the Workshop "Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities" (MindTheGap 2014)*, pages 25–30, Berlin, Germany, 4th March 2014. URL http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf.

A. Kilgarriff and D. Tugwell. Sketching words. In M.-H. Corr´eard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, EURALEX, pages 125–137, 2002. URL http://www.kilgarriff.co.uk/Publications/2002-KilgTugwell-AtkinsFest.pdf.

P. Rychl´y. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9, 2008. URL http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf.

A. Sokirko. A technical overview of DWDS/Dialing Concordance. Talk delivered at the meeting *Computational linguistics and intellectual technologies*, Protvino, Russia, 2003. URL http://www.aot.ru/docs/OverviewOfConcordance.htm.

Genre influence on word formation change: The development of adjectival derivation in the early stage of New High German

Luise Kempf
Johannes Gutenberg-Universität Mainz

This paper presents results from a comprehensive study on adjective forming suffixes in German. Two corpora are surveyed, the BONN EARLY NEW HIGH GERMAN CORPUS and the GERMANC, together spanning 1350−1800 CE. Combined with previous studies that cover the 13th, 15th, and 20th century respectively ([1],[2],[3]), the present results provide a long term diachronic survey of productivity changes in (competing) adjectival suffixation patterns (such as in *König-lich* "king-ly"= 'royal', *stein-ig* 'stony', *tör-icht* 'fool-ish').

The GERMANC (1650−1800) allows disentangling the factors region, time, and genre − with striking results: While there is only minor variation among the regions, or the three periods covered, the factor genre accounts for highly significant differences. Expectations about the character of the genres are confirmed, if not exceeded: Sermons in the GerManC are highly conservative or even archaic, in that they show a distribution of suffixes typically found in texts dating to ca. 1500. This holds for the share each suffix has, for the specific inventory of suffixes used, and also for the small size of the inventory. Narrative prose texts

show the highest type-token ratio, indicating a broader vocabulary spectrum, and also the highest hapax-token ratio, signifying an above-average word formation activity (cf. [4]). Both, arguably, hint at a more creative way of employing the suffixes in general use at that time. Newspapers and scientific texts are particularly innovative and ahead of their time in terms of suffixation patterns: They contain the highest percentage of both foreign suffixes (*pyramid-al* 'pyramid-shaped') and foreign bases (*elektr-isch* 'electric(al)'), evidencing increased language contact at the dawn of the New High German period.

Additionally, scientific texts exhibit two more features that are particularly illuminating about language change at the time: They not only utilize a rather large number of different suffixes, but they are also the principal breeding ground for a novel class of suffixes that rapidly gains frequency in the 19th century. In contrast to the older, monomorphemic and semantically opaque suffixes (*-lich*, *-ig*, *-isch*), these new suffixes are complex in form and transparent in meaning, e.g. *-artig* 'X-like', lit. "X-kind-y"; *-förmig* 'X-shaped', lit. "X-shape-y"; *-haltig* 'containing X', lit. "X-content-y", to name but a few. Both the larger suffix inventory and the higher usage of explicit suffixes are a testament to the increasing importance of written language: Speech is highly contextualized, its communicative contents typically are supported by gestures, facial expressions, prosody, shared knowledge about participants and situation, and so forth. Written texts, in contrast, lack this rich extra- and paralinguistic context and also do not allow immediate callback for clarification. To compensate for this relative contextual poverty, writing relies on constructions that are more explicit, transparent and unambiguous.

Overall, the study unearths which genres were most innovative, unveils specific properties of different genres, and uncovers how these properties connect with the rise of written language at the dawn of New High German. Comparison with similar studies on other languages should permit to identify general tendencies of the text-types in which language change happens first.

CORPORA AND REFERENCES

BONN EARLY NEW HIGH GERMAN CORPUS http://www.korpora.org/Fnhd/

GERMANC CORPUS http://www.ota.ox.ac.uk/desc/2544

[1] Ganslmayer, Christine (2012): Adjektivderivation in der Urkundensprache des 13. Jahrhunderts. Eine historisch-synchrone Untersuchung anhand der ältesten deutschsprachigen Originalurkunden. Berlin, New York.

[2] Thomas, Barbara (2002): Adjektivderivation im Nürnberger Frühneuhochdeutsch um 1500. Eine historisch-synchrone Analyse anhand von Texten Albrecht Dürers, Veit Dietrichs und Heinrich Deichslers. Berlin, New York.

[3] Kühnhold, Ingeburg/Putzer, Oskar/Wellmann, Hans (eds.) (1978): Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Teil 3: Das Adjektiv. Eine Bestandsaufnahme des Instituts für Deutsche Sprache, Forschungsstelle Innsbruck. Berlin.

[4] Baayen, R. Harald (2009): Corpus linguistics in morphology. Morphological productivity. In: Herbert E. Wiegand, Merja Kytö & Anke Lüdeling (Hgg.): Handbücher zur Sprach- und Kommunikationswissenschaft, 29. Bd. 2. Berlin, 899–919.

Compiling the *Going Dutch Corpus*: A multi-genre approach to eighteenth- and nineteenth-century Dutch

Andreas Krogull
Leiden University

The early nineteenth century can be regarded as a fundamental phase of Dutch nation building and nationalist language planning. One of the key instruments during this process

was the introduction of the first official orthography and grammar of Dutch in 1804 and 1805, respectively. The language norms codified in Matthijs Siegenbeek's orthographic treatise (*Verhandeling over de Nederduitsche spelling ter bevordering van eenparigheid in dezelve*) and Petrus Weiland's grammar (*Nederduitsche spraakkunst*) became the national standard, intended for use in administration and education. Surprisingly, the effectiveness of the so-called *schrijftaalregeling* 'written language regulation' on actual language usage has not yet been studied systematically.

In order to gain new insights into language variation and change in eighteenth- and nineteenth-century Dutch, a multi-genre diachronic corpus has been compiled as part of the research project *Going Dutch. The Construction of Dutch in Policy, Practice and Discourse, 1750-1850*, currently conducted at Leiden University. Taking the publication of Siegenbeek's spelling and Weiland's grammar as its main point of departure, the *Going Dutch Corpus* (approx. 420,000 words) provides access to language use before and after the codification of Standard Dutch. The corpus comprises data from three genres: (1) private letters, (2) personal diaries and travelogues, and (3) regional newspapers. They represent two different types of historical text sources: On the one hand, the first two sub-corpora (letters and diaries) include handwritten, conceptually more 'oral' ego-documents, which are considered to be the most valuable sources for historical sociolinguistic research. The sub-corpus of newspapers, on the other hand, represents the language used in published, printed texts, typically associated with writing. Also taking into account the regional diversity of the Northern Netherlands in the late eighteenth and early nineteenth centuries, the corpus contains texts from seven regions, covering the urbanised centre as well as more peripheral provinces of the language area. Furthermore, the ego-documents in the corpus were written by both men and women. The *Going Dutch Corpus* can thus be used for the study of language change on several variational dimensions (i.e. genre, temporal, spatial, social, individual).

In my presentation, I will briefly outline the socio-historical context of the *Going Dutch* project. Then, I will focus on the compilation of the *Going Dutch Corpus*, discussing its genres and structure, as well as some methodological considerations and challenges. Finally, I will present the corpus-based analyses of two orthographic variables: (1) the representation of /x/+*t* in etymologically different positions by either ⟨gt⟩ or ⟨cht⟩ (e.g. in *klagt*/*klacht*), and (2) the representation of Wgm. *$\bar{\imath}$ by either dotted ⟨ij⟩ or undotted ⟨y⟩ (e.g. in *mijn*/*myn*). By comparing the quantitative results found in private letters and newspapers, I will highlight the importance of a multi-genre approach for historical sociolinguistics in order to assess language variation and change in eighteenth- and nineteenth-century Dutch.

The evolution of SUCCESS: A diachronic study of conceptual metaphors in American success books

Jolanta Łącka-Badura
University of Economics in Katowice, Poland

The American myth of success is an extremely influential cultural belief in the collective American mind. Importantly, the way Americans understand and pursue success determines not only much of the daily life in the US, but also, fortunately or not, shapes the lifestyles of millions of people in other parts of the globe. Given the unchanging popularity of self-help and self-improvement publications in America and elsewhere, providing advice and inspiration for success-hungry people, it seems both interesting and worthwhile to investigate whether the metaphorical conceptualisation of SUCCESS reflected in best-selling American success books has evolved during the last century.

The study is based on a comparative analysis of utterances in which the lexical entry SUCCESS is regarded as constituting part of a metaphorical expression (Pragglejaz Group 2007). The utterances have been extracted from two corpora, each comprising approximately 1,550 pages of continuous text. The first ('old') corpus includes SUCCESS utterances found in six success books whose first editions were published in the early decades of the 20th century (1906 – 1928); the second ('new') corpus comprises utterances of the same kind extracted from seven books published in the years 2002 – 2011. The majority of both the 'old' and 'new' books analysed are widely recommended in the US as valuable success and self-improvement resources.

The analysis is methodologically grounded in the *Conceptual Metaphor Theory* (Lakoff and Johnson 1980, 1999). Despite the frequently expressed views that the significance of *conceptual metaphor* as an explanatory construct is sometimes overstated in cognitive linguistic research, the CMT remains to be considered by many as the dominant perspective on metaphor, influencing a vast array of studies in humanities and cognitive sciences (e.g. Evans 2010, 2013; Gibbs and Perlman 2006; Gibbs 2013). In the first instance, the study seeks to investigate which metaphorical source domains, as understood within the framework of CMT, prove to be most productive and influential in both corpora. Secondly, the most frequent metaphors found in both sets of texts are compared, with a view to determining which recurring mappings seem characteristic of each corpus, and which have remained unchanged during the last hundred years.

The analysis demonstrates that SUCCESS is metaphorically conceptualised in a variety of ways, encompassing structural, orientational and ontological metaphors reflecting our everyday experience. Notwithstanding the wide range of author-specific metaphors, there are certain recurring patterns of conceptualisation observed in both corpora. Some appear to be relatively timeless metaphorical mappings, including the metaphors SUCCESS IS A DESTINATION (DESIRED LOCATION AT THE END OF A PATH/ROAD), SUCCESS IS A PHYSICAL OBJECT, SUCCESS IS KNOWLEDGE/SCIENCE, with a variety of sub-mappings of the overarching conceptualisations. A more detailed analysis reveals some interesting differences between the metaphors found in the two corpora that may be interpreted as reflecting a shift in the American reality, culture and values during the last century. A more extensive study would, however, be required to confirm the findings.

REFERENCES

Evans, V. 2010. "Figurative language understanding in LCCM theory". *Cognitive Linguistics* 21 (4). 601-662.

Evans, V. 2013. "Metaphor, lexical concepts and figurative meaning construction". *Journal of Cognitive Semiotics* V (1-2). 73-107.

Gibbs, R.W. and M. Perlman. 2006. "The contested impact of cognitive linguistic research on the psycholinguistics of metaphor understanding". In: Kristiansen, G., M. Achard, R. Dirven and F.J. Ruiz de Mendoza Ibáñez (eds.), *Cognitive Linguistics: Current Applications and Future Perspectives*. Berlin: Mouton de Gruyter. 212-228.

Gibbs, R.W. 2013. "Why do some people dislike conceptual metaphor theory?" *Journal of Cognitive Semiotics* V (1-2). 14-36.

Lakoff, G. and Johnson, M. 1980. *Metaphors We Live by*. Chicago: University of Chicago Press.

Lakoff, G. and M. Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: basic Books.

Pragglejaz Group. 2007. "MIP: A method for identifying metaphorically used words in Discourse". *Metaphor and Symbol* 22. 1-39.

Tracing recent change in two multi-genre corpora of advanced non-native English

Mikko Laitinen & Magnus Levin
Linnaeus University

This presentation investigates how recent and ongoing grammatical change is adapted in advanced non-native English. This concept of advanced non-native points to English used as an additional linguistic resource alongside people's L1s in countries with no colonial links to Britain. The first part of this presentation shows that English in such contexts has for long been conceptualized as learner English and evidence has been drawn from single-genre corpora. The results in that paradigm have shown genre interference and over/under-representation of certain characteristics in writing (Granger et al, eds. 2015). More recently, the English as a lingua franca paradigm has investigated patterns of spoken interactions (Mauranen 2012). We show that both of these approaches have focused less on variability and have therefore by and large overlooked genre as a possible factor in change and variation. We suggest that the increased globalization and expansion of English highlights the possible role of non-native speakers/writers as agents of present-day language change and calls for new empirical approaches of how the native varieties and the various new English(es) overlap.

After that, we present a research project that sees advanced non-native English(es) as one stage in the long continuum of varieties. Our research charts what role diachronic processes of language change play in shaping advanced non-native use and studies quantitative and qualitative patterns of how the English language is shaped in multilingual settings (Laitinen & Levin in press, Laitinen forthcoming). We are currently working with collecting two multi-genre corpora of advanced non-native English texts. The objective is to incorporate a broader perspective of advanced non-native English than in the learner corpus and the ELF paradigm by adding written evidence which covers a broad range of texts on the informative–interactive continuum of genres. These corpora are synchronic, but similarly to Collins (2009) it is assumed that the patterns of variability in the channels (spoken-written) and genre could be used to draw diachronic conclusions of how what happens to recent change in advanced non-native English(es). After illustrating the corpus work, we present two case studies which focus on patterns of variability emerging in non-native contexts. The first one focuses on core and emergent modal auxiliaries expressing obligation and necessity, and the second on phraseology of discourse organizers (e.g. WHEN IT COMES TO / IN TERMS OF). Both of these are currently undergoing change in the native varieties, and we present evidence of the patterns emerging in non-native contexts. The topics discussed are useful to those interested in the role of genre in language change and in particular to those interested in recent and ongoing change.

REFERECNES

Collins, Peter. 2009. Modals and quasi-modals in world Englishes. *World Englishes* 28:3, 281–292.

Granger, Sylviane, Fanny Meunier & Gaetanelle Gilquin (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Laitinen, Mikko. Forthcoming. Ongoing changes in English modals: On the developments in advanced L2 use of English. In Olga Timofeeva, Anne Gardner, Alpo Honkapohja (eds.). Amsterdam: John Benjamins.

Laitinen, Mikko & Magnus Levin. In press. On the globalization of English: Observations of subjective progressives in present-day Englishes. In Elena Seoane & Cristina Suárez-Gómez (eds.) *World Englishes: New Theoretical and Methodological Considerations* (*Varieties of English around the World*). Amsterdam: John Benjamins.

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

The quotative use of the impersonal third person plural pronoun in English versus the agentless passive of *say* – a diachronic genre comparison

Eliane Lorenz
Friedrich-Schiller-Universität Jena

Impersonalization can be defined as an operation where the subject is not formally represented (Kibort 2004: 15-16) or where it is encoded by an impersonal expression (Creissels 2007: 4). Concerning the latter, one way of expressing impersonality is via lexical items, such as personal pronouns (Kibort 2004: 61). Pronouns are used impersonally when they have one or more human referents that cannot be identified in the real world:
(1) *The large grain'd (**they say**) is better for the Landlord, and the Land [...]*. (1675hook.s2b)
One particular impersonal pronoun, the third person plural pronoun *they*, is typically divided into five subtypes: 'universal', 'corporate', 'specific', 'inferred', and 'vague' (Cabredo Hofherr 2003; Siewierska 2011; Siewierska & Papastathi 2011). In this talk, I will describe a sixth type, the quotative use. So far, this construction, which consists of a 3pl-IMP and the speech-act verb *say*, has been subsumed under the five types (van der Auwera et al. 2012: 44; Siewierska & Papastathi 2011: 585). On the basis of a historical corpus, I will first describe and classify this quotative use and argue for it being a specific type. These 3pl-IMPs report the thoughts of others; the referents can be either universal (1) or existential (2); and the sentences can refer to generic (1) or episodic (3) situations.
(2) *"**They say** she's handsomer than Gal Sullivan," said Hanlon; "and I think myself she is.* (ARCHER 3.1; 1847carl.f5b)
(3) ***They say**, that they saw there a Whale, upon whose back stuck a Harking-Iron of Gascony.* (ARCHER 3.1; 1675ano2.s2b)
Second, I will compare the quotative 3pl-IMPs to the agentless passive with the speech-act verb *say*. The passive, which is another strategy of expressing impersonality, has been argued to be used as an equivalent to the 3pl-IMPs (Siewierska 2008:31; Myhill 1997: 802-803).
(4) ***It is certainly said**, that the Prince of Conde will be with an Army here about the beginning of May.* (ARCHER 3.1; 1672on1.n2b)
All instances of both the quotative 3pl-IMPs and the passive have been manually extracted from the ARCHER corpus and manually coded as impersonal or personal. The corpus includes nine different genres and covers the years between 1650 and 1999 – this allows for a thorough investigation in order to find genre differences and to compare different periods. Various linguists have observed that the genre affects the choice between an agentless passive and an active 3pl-IMP construction (Kitagawa & Lehrer 1990: 746; Myhill 1997: 839). This seems to hold for the quotative use and the equivalent passive structure: *they* appears particularly frequent in 'drama', 'fiction', and 'letters', whereas the passive is favored in 'news' and 'science', for instance. The level of formality and informality seems to play the decisive role, because the meanings both constructions express are fundamentally similar. Another observation is the decrease of the agentless passive with *say* and the increase of the quotative 3pl-IMPs throughout the periods.

REFERENCES

ARCHER 3.1 (Bamberg) (1990–1993/2002/2007/2010/2013). *A Representative Corpus of Historical English Registers.* Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Bamberg, Freiburg, Heidelberg, Helsinki, Lancaster, Leicester, Manchester, Michigan, Northern Arizona,

Santiago de Compostela, Southern California, Trier, Uppsala, Zurich. Examples of usage taken from ARCHER were obtained under the terms of the ARCHER User Agreement (available on the Documentation page of the ARCHER website, http://www.manchester.ac.uk/archer/).
van der Auwera, Johann, Volker Gast & Joroen Vanderbiesen (2012). "Human impersonal pronoun uses in English, Dutch and German." *Leuvense Bijdragen* 98(1): 27-64.
Cabredo Hofherr, Patricia (2003). "Arbitrary readings of 3pl pronominals." In: *Proceedings of the Conference "sub7 — Sinn und Bedeutung", 7th Annual Meeting of the Gesellschaft für Semantik*, Ed. Matthias Weisgerber. Arbeitspapiere des Fachbereichs Sprachwissenschaft, Vol. 114. Konstanz University, FB Linguistik. 81-94.
Creissels, Denis (2007). "Impersonal and anti-impersonal constructions: a typological approach." Extended version of the paper presented at Alt-7, Paris, September 24-28, 2007.
Kibort, Anna (2004). "Passive and passive-like constructions in English and Polish." Ph.D. thesis, University of Cambridge. Available online at <http://ak243.user. srcf.net/pdfs/AnnaKibortThesis.pdf>. 12 May 2015.
Kitagawa, Chisato & Adrienne Lehrer (1990). "Impersonal uses of personal pronouns." *Journal of Pragmatics* 14: 739-759.
Myhill, John (1997). "Toward a functional typology of agent defocusing." *Linguistics* 35: 799-844.
Siewierska, Anna (2008). "Ways of impersonalizing: Pronominal vs. verbal strategies." In: *Current Trends in Contrastive Linguistics.* Eds. Maria de los Angeles Gomex Gonzales, Lachlan Mackenzie, Elsa Gonzalez Alvarez. Amsterdam: John Benjamins. 3-26. Available online at <http:// http://eprints.lancs.ac.uk/27999/1/WaysofImp.pdf>. 3 June 2015.
Siewierska, Anna (2011). "Overlap and complementarity in reference impersonals: man-constructions vs. third person plural-impersonals in the languages of Europe." In: *Impersonal Constructions. A cross-linguistic perspective.* Eds. Andrej Malchukov, Anna Siewierska. Amsterdam, Philadelphia: John Benjamins Publishing Company. 57-90.
Siewierska, Anna & Maria Papastathi (2011). "Towards a Typology of Third Person Plural Impersonals." *Linguistics* 49(3): 575-610.

Genres, audience and thematic organisation in the recent history of English

Ana Elina Martínez-Insua & Javier Pérez-Guerra
University of Vigo

This paper reports on our ongoing investigation of Theme in a number of genres in the recent history of English. The goal of this project is to investigate the relation between the target audience of a given genre (or a text type within a genre) and the thematic choices made by the writer, Theme being conceived as in, among others, Berry (1995, 2013) (i.e. the opening lexical material with some function in the experiential structure of the clause and coming up to, and including, the Subject). In particular, we focus on the syntactic function and the meaning of (Subject) Themes. As regards meaning, we will follow Berry's (2013) distinction between contentful and contentlight Subject Themes, and take as a basis Berry's (2013: 259) hypothesis that most Subject Themes are contentful (New Topics and Resumed Topics) in formal written texts, while most of them are contentlight (Given Topics) in informal spoken texts. In this vein, this study tests to what extent the target audience may be also a factor affecting the category and the content weight of the thematic position or what the Subject Themes refer to.

Framed within a larger project on the variation and textual characterisation of English in its recent history, the present study takes data from two genres: medical texts (from the electronic corpus of Early Modern English Medical Texts; Taavitsainen & Pahta 2011) and newspaper discourse (from the Zurich English Newspaper corpus; Lehmann et al. 2006) in Early and Late Modern English. Two research questions will be addressed here: (i) the way in which texts represent the world and organise the messages, depending on whether they are addressed to learned, unlearned or intermediate target audiences, and (ii) the connection between the target audience and the type of information conveyed by the Subject Themes. On

the one hand, the data reveal that the high frequency of textual elements in texts addressed to unlearned audiences is the result of the writer's choice to use their linking and explanatory nature so as to make the message as clear as possible. On the other hand, the significant differences between the informative load of Subject Themes addressed to a learned audience (more contentful) and of those directed to an unlearned/intermediate audience (more contentlight) has been interpreted as the writers' attempt to ease processing in texts addressed to the latter audiences.

REFERENCES

Berry, Margaret. 1995. Thematic options and success in writing. In Mohsen Ghadessy ed. *Thematic development in English texts*. London: Pinter, 55–84.

Berry, Margaret. 2013. Contentful and contentlight subject themes in informal spoken English and formal written English. In Gerard O'Grady, Tom Barlett & Lise Fontaine eds. *Choice in language. Applications in text analysis*. Sheffield: Equinox, 243–268.

Lehmann, Hans Martin, Caren auf dem Keller & Beni Ruef. 2006. Zen Corpus 1.0. In Roberta Facchinetti & Matti Rissanen eds. *Corpus-based studies of diachronic English*. Bern: Peter Lang, 135–155.

Taavitsainen, Irma & Päivi Pahta. 2011. *Medical writing in Early Modern English*. Cambridge: Cambridge University Press.

Automatic genre identification in EEBO-TCP: A multidisciplinary perspective on problems and prospects

Seth Mehl
The University of Sheffield

This paper addresses issues in automatic genre identification in Early English Books Online (specifically, EEBO-TCP), which contains approximately 55,000 keyed texts. The paper is presented in relation to a major research project mapping semantic and conceptual change in Early Modern English (EModE). Central to the paper is the major methodological question for exploring genre in very large historical datasets like EEBO-TCP: How might genre be identified automatically in such datasets, in a relatively inductive, bottom-up way?

The paper first briefly defines *genre* in relation to EModE studies (cf. Schmidt 2013, Richetti 2005), corpus linguistics and corpus stylistics (cf. Stockwell and Whiteley 2014, Biber and Conrad 2009), and Natural Language Processing (NLP; cf. Cleuziou and Poudat 2007, Kessler *et al*. 1997, Karlgren and Cutting 1994, Mehler *et al*. 2010). *Genre* can be construed in many different ways in these different fields: in a shallow, formal fashion (e.g. prose, verse), or in a deeper, more rhetorical fashion (e.g. instructional writing, recipes, medical manuals). Genre-specific change is forwarded as crucial to the project's goals, which include investigating to what extent particular genres relate to semantic and conceptual change. For example, if a particular meaning of *modern* is first attested in military manuals, how does this context relate to the word's semantics? What was its rate of adoption and adaptation in other genres? The project is identifying semantic characteristics of thousands of words in EEBO using a bottom-up computational approach and distributional semantic analysis. Because of the immense amount of data in EEBO-TCP, manual analysis of every word and every text, including manual genre analysis, is not feasible. The problem, then, is to identify genres automatically. Although EEBO-TCP texts have been previously tagged for genre categories, this project's bottom-up approach is not served by reliance on the top-down (and sometimes incomplete) genre identification previously presented.

Next, research in automatic genre identification is reviewed. Previous research has been conducted mainly in two contexts: NLP and corpus linguistics. NLP research in genre identification generally involves information retrieval using Present Day English (PDE),

whereas EModE has not been extensively studied in this regard (though Underwood, 2014, is an important exception). Studies of PDE may identify genre based on morphosyntactic characteristics of texts (cf. Cleuziou and Poudat 2007), or a combination of morphosyntactic, lexical, and punctuation characteristics (Kessler *et al.* 1997). In NLP, lexical characteristics alone are not generally used for genre identification, but are used for topic-modelling (cf. Cleuziou and Pudat 2007, Rehurek and Sojka 2010), though topic may be seen to relate to genre. In corpus linguistics, experiments are commonly undertaken linking genre (or *text type*, or *register*) with an array of linguistic features, including strictly lexical, as well as phrasal, grammatical, and discursive (cf. Biber 1989).

Finally, preliminary findings on genre identification in EEBO-TCP are reported. Questions about the usefulness and limitations of automatic genre identification are discussed in detail, with particular attention to the problems presented by EModE and EEBO-TCP. Key questions are highlighted for the project moving forward, including how the chosen methods might be further developed, and the possibility of providing further insight into previous, assignments of genre to EEBO texts.

REFERENCES

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27, 3-23.

Biber, Douglas and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Cleuziou, Guillaume and Celine Poudat. 2007. On the impact of lexical and linguistic features in genre- and domain-based categorization. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing*. Berlin: Springer. 599-610.

Karlgren, Jussi and Douglas Cutting. 1994. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*.

Kessler, B., G. Nunberg, and Hinrich Schutze. 1997. Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. 32-38.

Mehler, Alexander, Serge Sharoff and Marina Santini. 2010. *Genres on the web: Computational models and empirical studies*. Berlin: Springer.

Rehurek, Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop: New challenges for NLP frameworks*. 46-50.

Richetti, John (ed.). 2005. *The Cambridge history of English literature, 1660-1780*. Cambridge: Cambridge University Press.

Schmidt, Gary A. 2013. *Renaissance hybrids: Culture and genre in Early Modern England*. London: Ashgate.

Stockwell, Peter and Sarah Whiteley (eds). 2014. *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press.

Underwood, Ted. 2014. Understanding Genre in a Collection of a Million Volumes, Interim Report. http://dx.doi.org/10.6084/m9.figshare.1281251. Accessed 8 October, 2015.

Alien invaders and handsome little villains: a diachronic collocates analysis of the grey and red squirrel in news texts 1825 – 2005

Emma McClaughlin
Lancaster University

A current Leverhulme-funded project (2013 – 2016) is investigating the discursive representation of animals in contemporary Britain across a range of spoken and written discourse types (Cook & Sealey, 2013). My own research attached to this project uses a modern-diachronic corpus assisted discourse studies (MD-CADS) (Partington, 2010) approach to investigate how a small number of key wildlife species in Britain are represented in the genre of news discourse over a 220-year period.

As well as identifying changes in linguistic variation, MD-CADS can be used to study the influence of social, political and cultural perspectives on language change. I am interested in the influence that such changes have had on news discourse about British

wildlife across the late-modern period of English. Unlike traditional MD-CADS, I am not using temporally parallel corpora for diachronic language comparison but rather a (relatively) continuous dataset gathered specifically for this project. This has had a bearing on the methods open to me for carrying out a diachronic analysis.

I will present a short analysis of a 320,000-word specialised news corpus containing letters and articles about red and grey squirrels published between 1825 and 2005. First, I will demonstrate how I have adopted the waves, peaks and troughs (WPT) statistical method (Gabrielatos, McEnery, Diggle, & Baker, 2012) to segment the corpus for contrastive linguistic analysis. I will then present a short diachronic collocates analysis (McEnery & Baker, 2015) for both species ('red squirrel(s)' and 'grey squirrel(s)') across the corpus to establish how the representations of these animals can be seen to have changed over time. A small number of illustrative findings will be discussed including: (i) that the language used to report about these animals is motivated by text-external factors such as legislation and campaigns; (ii) place of origin and national identity is an important part of how these creatures are reported about; and (iii) collocates describing characteristics of the red and grey squirrel, combined with the findings above, contribute to a familiar pattern of positive in-group and negative out-group presentation present in racist discourse. There are indications that human-human relationships may influence such language choices. I discuss the findings in the context of the broader social, cultural and political history of Britain in line with the discourse-historical approach (Reisigl & Wodak, 2009).

REFEREMCES

Cook, G., & Sealey, A. (2013). 'People', 'products', 'pests' and 'pets': the discursive representation of animals. Retrieved 16th August 2014, from http://animaldiscourse.wordpress.com/

Gabrielatos, C., McEnery, T., Diggle, P. J., & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics, 17*(2), 151-175.

McEnery, T., & Baker, H. (2015). *The corpus and the historian: using corpora and corpus linguistics in historical investigations*. Paper presented at the Corpus Linguistics Conference, Lancaster University.

Partington, A. (2010). Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: an overview of the project. *Corpora, 5*(2), 83-108.

Reisigl, M., & Wodak, R. (2009). The discourse historical approach (DHA). In R. Wodak & M. Meyer (Eds.), *Methods of Critical Discourse Analysis* (2nd ed., pp. 87-121). London: Sage.

The Bilingual Public Notices from the City of Luxembourg (1795-1920): Functional Specifics and Linguistic Developments

Olivier Moliner & Rahel Beyer
Universität Duisburg-Essen, Germany / Université du Luxembourg, Luxembourg

This paper presents findings of the binational project „Language standardization in Diversity: The case of German in Luxembourg (1795-1920)", which is funded by the National Research Fund Luxembourg and the German Science Foundation. With a long history of multilingualism, Luxembourg constitutes a prime case for studying the impact of language contact on language standardization, i.e. language contact between Germanic varieties (i.e. Moselle-Franconian/emerging Luxembourgish, colloquial German) and between German and French.

To investigate the process of standardization, the project mainly draws on a corpus of 2.348 predominantly bilingual German/French public notices issued by the municipality of Luxembourg (namely the mayor and the councillors) during the so-called long 19th century. These notices constitute a specific type of posters, which were mostly put up on doors of

churches and town halls, thus leading to a "colonialization of the public space" (Habermas 1990). For the study, the texts were chronologically and representatively sampled, image-scanned and text-digitized (cf. Gilles/Ziegler 2013).

They all belong to one and the same medium as a major part of the posters were published as parallel texts in German and French as two equivalent texts of same content and an important number of the public notices were produced by four different Luxembourgish printers. However, this medium encompasses lots of different genres, e.g. laws, enactments, regulations or auctions. They, in turn, can be related to different levels of normativity. Besides these findings concerning the entirety of the corpus, three notices are analysed in detail concerning functional (Brinker/Cölfen/Pappert 2014) and pragmatic aspects (Klein 2000). From the two text clusters "strongly normative" and "weakly normative" we focus on the most frequent genres respectively, i.e. regulation (réglement) and auction (adjudication). Concerning the temporal dimension, differences (regarding e.g. length, amount of words, number of paragraphs, and titles) between the French and the German versions are more important at the beginning of 19th than at the end of 19th century. The choice of documents also takes different political regimes, or rather phases, of frenchification and germanification into account. They reveal that the priority given to German or French, for instance the language choices in the titles, in many cases correlates with the political regimes.

Linguistically, certain constructions can be identified as typical for specific genres and for specific periods. In the text cluster with strongly normative texts, for instance, we find canonical formulas from (French) legal language. Extensive corpus analysis of explanation-introducing *considérant que* shows that their realization in the German column is variable at the beginning of the 19th century and also exhibits structural and lexical similarities with the French original. In the course of the century the formula is firstly expressed more uniformly, secondly adapted to German characteristics (towards nominal structures) and finally subject to lexical changes. After its emergence from the influence of French, the formula thus manages to develop autonomously.

REFERENCES

Arnold, Judith (2007): *Das Abstimmungsplakat als Textsorte – Kontext und Merkmale eidgenössischer Abstimmungsplakate*. Zürich. [http://www.arsrhetorica.ch/Abstimmungsplakate-02.htm]

Brinker, Klaus/Cölfen, Hermann/Pappert, Steffen (2014): *Linguistische Textanalys*e. Eine Einführung in Grundbegriffe und Methoden. Berlin: Erich Schmidt.

Gilles, Peter/Ziegler, Evelyn (2013): "The "Historical Luxembourgish Bilingual Affichen Database"". In: P. Bennett/M. Durrell/S. Scheible and R.J. Whitt (eds.): *New methods in Historical Corpus Linguistics*. Tübingen: Narr: 127-138.

Habermas, Jürgen (1990): *Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der Bürgerlichen Gesellschaft*. Frankfurt am Main: Suhrkamp.

Klein, Josef (2000): „Textsorten im Bereich politischer Institutionen". In: Brinker, Klaus (ed.): *Text- und Gesprächslinguistik: ein internationales Handbuch zeitgenössischer Forschung* (*Linguistics of text and conversation*). 2nd vol. Berlin: de Gruyter: 732-755.

Rickards, Maurice (1973): *The public notice*. Newton Abbot: David & Charles.

Genre and change in the Corpus of English History Texts

Isabel Moskowich & Begoña Crespo
University of A Coruña

The aim of this presentation is to offer an overview of the *Corpus of History English Texts* (CHET) as part of the *Coruña Corpus of English Scientific Writing* (CC) (Sinclair, 2004; Moskowich and Crespo, 2012). Two of the defining characteristics of the *Coruña*

Corpus are, on the one hand, that it is a diachronic corpus; on the other, it may be considered either as a single- or multi-genre corpus (depending on the theoretical tenets adopted) (Kytö, 2010; McEnery and Hardie, 2013). This corpus has been designed as a tool for the study of language change in English scientific writing in general and, in particular, in the different scientific disciplines which have been sampled in each separate sub-corpus as we will see. All texts in this compilation were published between 1700 and 1900, thus offering a thorough view of late Modern English scientific discourse, a period traditionally neglected in English historical studies. The analysis of this variety of English may also contribute to the description of the origins of English as "the language of science".

The purpose of the corpus is to facilitate investigation at all linguistic levels, although, in principle, it does not seem appropriate for the study of phonology. The novelty it offers is the possibility to use these scientific texts for socio-linguistic research as well. Metadata files containing some personal details about the author (age, sex, place of education) of each sample and about the work (date of publication, genre/text-type) from which the sample has been extracted have been incorporated (Crespo and Moskwich, 2010). This applies to all the sub-corpora in the *CC (Pahta and Taavitsainen, 2010)*, both published as is the case with *CETA, A Corpus of English Texts on Astronomy* (Benjamins, 2012) and *CEPhiT, A Corpus of English Philosophy Texts* (Benjamins, forthcoming) and those under compilation, such as *CHET* (*Corpus of English History Texts*), the specific sub-corpus we want to introduce here. On this occasion, we would like to explore the constraints discipline imposes upon genre diversity in corpora by comparing *CHET* with its predecessors. Similarly, we will try to ascertain to what an extent the variable time/period also exerts any influence in the selection of genres on the author's part.

From a technical point of view, all the texts have been keyed in following the Text Encoding Initiative conventions and saved in XML format. Although some editorial decisions had to be made due to the peculiarities found in the samples, the use of an extended mark-up language has made and will continue making wide distribution and exploitation possible. We also decided to create a *corpus* management tool in order to retrieve information from the compiled data, both linguistic and non-linguistic. Thus, the *Coruña Corpus* Tool is an Information Retrieval system, where the indexed textual repository is a set of compiled documents that constitutes the *CC* (Lareo, 2009). Evidence of the searches that can be carried out is found in the pilot studies that will be briefly sketched (Moskowich et al, forthcoming).

REFERENCES

Crespo, Begoña and Moskowich, Isabel. 2010. *CETA* in the Context of the *Coruña Corpus* . Literary and Linguistic Computing, 25/2: 153-164.

Kytö, Merja. 2010. Data in historical pragmatics. In Jucker, Andreas and Taavitsainen, Irma (eds.). Historical Pragmatics. Berlin: Mouton de Gruyter, 33-67.

Lareo, Inés. 2009-. *El Coruña Corpus. Proceso de compilación y utilidades del Corpus of English Texts on Astronomy (CETA). Resultados preliminares sobre el uso de predicados complejos en ceta.* In Cantos, Pascual and Sánchez, Aquilino (eds.) *A Survey on Corpus-based Research Panorama de investigaciones basadas en corpus.* Murcia: Asociación Española de Lingüística de Corpus. 267-280. http://www.um.es/lacell/aelinco/contenido/ pdf/19.pdf.

McEnery, T. & Hardie, A. 2013 *The Oxford handbook of the history of linguistics.* Allan, K. (ed.). Oxford: Oxford University Press

Moskowich, Isabel and Crespo, Begoña (eds.). 2012. *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900.* Amsterdam/Philadelphia: John Benjamins. Moskowich, Isabel; Camiña, Gonzalo; Crespo, Begoña; Lareo, Inés. Forthcoming. *The Conditioned and the Unconditioned´: Late Modern English Texts on Philosophy* . Amsterdam/Philadelphia: John Benjamins.

Pahta, Päivi and Taavitsainen, Irma. 2010. Scientific Discourse. In In Jucker, Andreas and Taavitsainen, Irma (eds.). Historical Pragmatics. Berlin: Mouton de Gruyter, 549-586.

Sinclair, J. (2004). Corpus and Text — Basic Principles. In Martin Wynne (ed). Developing Linguistic Corpora: a Guide to Good Practice Produced by AHDS Literature, Languages and Linguistics, University of Oxford, UK. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm.

A Diachronic corpus of sermons: Tracing grammatical change throughout the history of German

Simon Pickl
Universität Salzburg

In this contribution, a single-genre diachronic corpus of German sermons spanning nearly one millennium will be presented. It is being compiled in order to be able to study grammatical change focussing on the diachronic aspect, keeping other factors of variation as uniform as possible, especially genre.

There are two main reasons for the decision to design a diachronic corpus that is based on the single genre 'sermons'. Firstly, the received history of grammatical change in German is characterized by a heterogeneous data basis. For instance, Old High German has been studied especially on the basis of translations of Latin religious texts, Middle High German has typically been portrayed as the language of courtly poetry, and studies on Early New High German have focussed predominantly on the language of chancery and trade (cf. Fleischer/Schallert 2011, 26–27). It is not always clear, therefore, if the differences identified between these periods are to be attributed exclusively to diachronic change or at least partly to inter-genre variation. Narrowing down the view by focussing on a single genre that is well documented throughout the history of German is a way of mitigating this problem. A second problem lies in the type of source material typically used when studying grammatical change in the past. Usually the sources come from widely available material, which is often representative of registers detached from orality. This restriction has been tried to overcome by using sources 'from below' (cf. Elspaß 2005) or 'ego-documents' (cf. Rutten/van der Wal 2013), which exhibit more oral features than the written registers. Sermons offer a different take on historical orality: they represent a reproduction or imitation of physically oral (albeit monologic) communication from the pulpit (cf. e.g. Mertens 1991) and retain a typical set of oral features even in their written form (cf. e.g. Mertens 1992).

The corpus is designed to include prints and manuscripts of sermons from different German-speaking regions. The texts are transcribed using TEI and manually annotated to make the corpus searchable for grammatical phenomena to be investigated, especially inflection. This presentation will sketch the outline of the project and its current status along with some first results.

REFERENCES

Elspaß, Stephan (2005): *Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert.* Tübingen: Niemeyer.

Fleischer, Jürg / Oliver Schallert (2011): *Historische Syntax des Deutschen. Eine Einführung.* Tübingen: Narr.

Mertens, Volker (1991): "‚Texte unterwegs'. Zu Funktions- und Textdynamik mittelalterlicher Predigten und den Konsequenzen für ihre Edition". In: Danielle Buschinger / Wolfgang Spiewok (eds): *Mittelalterforschung und Edition*. Amiens: Université de Picardie, 75–85.

Mertens, Volker (1992): "Predigt oder Traktat? Thesen zur Textdynamik mittelhochdeutscher geistlicher Prosa". In: *Jahrbuch für Internationale Germanistik* 24/2, 41–43.

van der Wal, Marijke J. / Gijsbert Rutten (eds) (2013): *Touching the Past. Studies in the historical sociolinguistics of ego-documents.* Amsterdam/Philadelphia: Benjamins.

Projecting and developing the Sardinian Medieval Corpus

Nicoletta Puddu
University of Cagliari

Coding and annotating corpora for historical languages is a difficult task. Consequently, as Viana, Zyngier and Barnbrook (2011) underline, this may often lead to the creation of text archives rather than corpora.

The present contribution will examine the methodological and theoretical problems encountered in planning and realizing the Sardinian Medieval Corpus (SMC).

The SMC is made up of documents written in Sardinian, a Romance language spoken in the island of Sardinia, whose first attestation dates back to the 11th century. Consequently, we considered the 11th century as the upper chronological limit of our corpus, while the 14th century is to be considered as the lower chronological limit, when Sardinia underwent the Aragonese domination and its language started to be strongly influenced by both Catalan and Castilian.

The corpus consists of about 200.000 words, and is not balanced nor representative, since the genre of SMC texts is predominantly legal. However, unprincipled corpora (as historical corpora usually are) can also be fruitfully used for linguistic analysis (see Curzan and Palmer 2006). The oldest and most interesting Sardinian Medieval texts are the *Condaghes*, documents recording acts of donation, transactions and income of churches and monasteries, often containing transcriptions of legal disputes, called *kertos*. According to Schneider (2013), court transcriptions are precisely those texts most closely representing spoken language. Thus, they can be particularly useful in the study of sociolinguistic variation (see Rissanen 2008).

The SMC shares many typical features with other Medieval corpora, such as orthographic variation and code mixing with Latin. We decided to use TEI compliant tagging at different levels, not only to preserve as much variation as possible (as Lass 2004 recommends), but also to make the corpus more accessible. At the philological level, editorial choices (like supplied texts or expansions of abbreviations) are constantly signaled. However, since the corpus is lemmatized and POS-tagged, it is also searchable through "normalized" forms, which are the equivalent of "glosses" in the philological tradition.

Moreover, the TEI header of each document contains contextual information on both text and participant description. This makes the SMC searchable by relevant sociolinguistic features which, as Vazquez and Marques Aguado (2012) point out, must be kept in due consideration when studying Medieval texts. Finally, we decided to mark the *kertos* in the Condaghes with specific tags, in order to identify the peculiar features which could mirror spoken language in these "conversations".

REFERENCES

Curzan, Anne, Palmer, Chris C. (2006), "The Importance of Historical Corpora, Reliability, and Reading", in Facchinetti, Roberta, Rissanen, Matti (a cura di), *Corpus Based Studies of Diachronic English*, Bern, Peter Lang, 17-34.

Lass, Roger (2004), "Ut custodiant litteras: Editions, Corpora and Witnesshood", in Dossena, Marina, Lass, Roger (a cura di), *Methods and Data in English Historical Dialectology*, Bern, Peter Lang, 21-48.

Puddu, Nicoletta (2015) "Costituzione del Sardinian Medieval Corpus: prime proposte per la codifica e l'annotazione", in Molinelli, Piera, Putzu, Ignazio E. (eds.) *Modelli epistemologici, metodologia della ricerca e qualità del dato. Dalla linguistica storica alla sociolinguistica storica*, Milano, FrancoAngeli, 282-299.

Rissanen, Matti (2008), "Corpus linguistics and historical linguistics", in Lüdeling, Anke, Kytö, Merja (eds.), *Corpus Linguistics. An International Handbook*, vol. 1, Berlin, Mouton de Gruyter, 53-68.

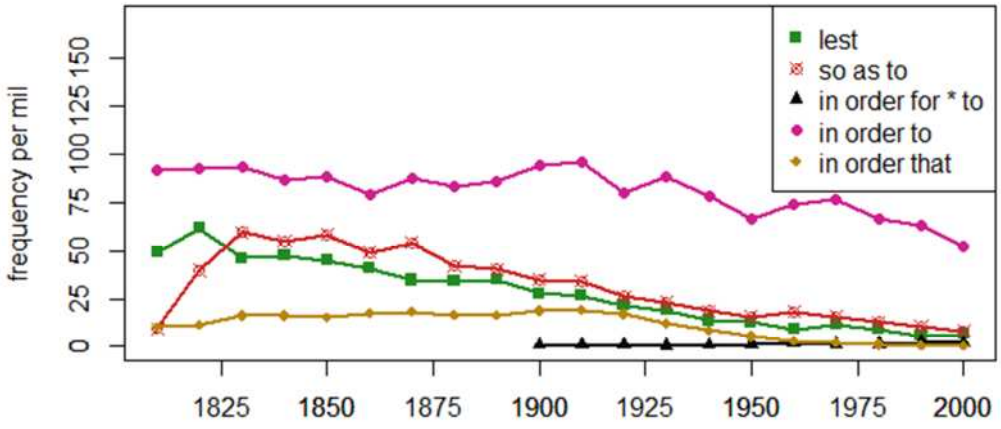
Schneider, Edgar W. (2013[2]), "Investigating Historical Variation and Change in Written Documents: New Perspectives", in Chambers, J.K., Schilling, Natalie (eds.), *The Handbook of Language Variation and Change*, Oxford, Blackwell, 57-82.
Vazquez, Nila, Marques-Aguado, Teresa (2012), "Editing the Medieval Manuscript in its Social Context", in Hernandez-Campoy, Juan Manuel, Conde-Silvestre, J. Camilo (eds.), *The Handbook of Historical Sociolinguistics*, Malden (MA)/Oxford, Wiley-Blackwell, 123-139.
Viana, Vander, Zyngier, Sonia, Barnbrook, Geoff (eds), (2011)"Synchronic and diachronic uses of corpora. Interview with Mark Davies", in Viana, Vander, Zyngier, Sonia, Barnbrook, Geoff (eds.), *Perspectives on Corpus linguistics, Amsterdam/New York, John Benjamins*, 63-80.

How to investigate grammatical obsolescence?: Some of the interesting things diachronic multi-genre corpora can tell us

Karolina Rudnicka
Albert-Ludwigs-Universität Freiburg

Based on first results from my ongoing PhD project, the present paper discusses an approach to investigating the under-researched topic of grammatical obsolescence. Grammatical obsolescence in a nutshell can be defined as a notion describing a situation when a previously productive construction is losing its productivity and popularity, often in a very gradual and long process that follows until there are only residues.

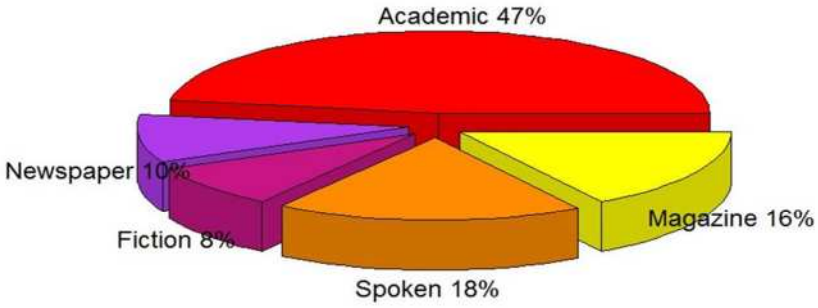
This chart shows the correlation of time and frequency per million words in the case of 5 variants of the first investigated variable – subordination of purpose (Data extracted from COHA):



The hypothesis that the decline observed in the above figures is not a mere fluctuation is further confirmed by correlation testing. But are all of these constructions actually obsolescent? What other criteria, apart from negative correlation between time and the frequency of use, should one consider to fully grasp the evolution of this broad constructional network?

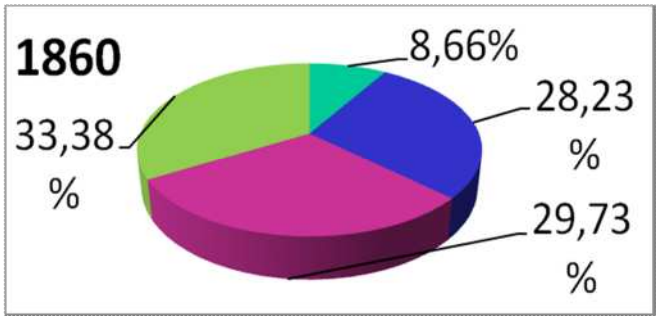
As Hundt (2014) claims "another symptom of syntactic decline may be that a construction specialises to certain kinds of discourse". Leech et al. (2009) refer to this phenomenon as distributional fragmentation: "instead of being dispersed in different varieties of texts in a corpus, the form tends to be increasingly restricted to certain genres and, within those genres, to certain texts".

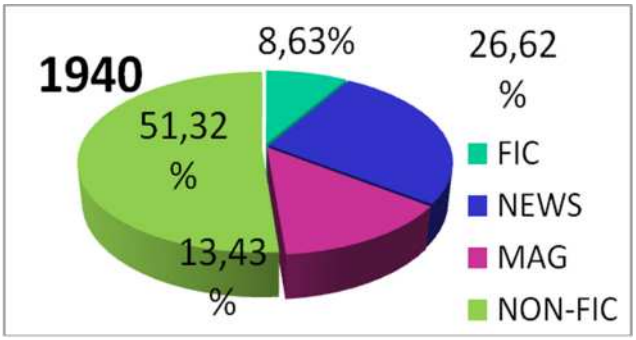
This pie chart shows the distribution of all cases of the purpose subordinator *in order that* across different genres in COCA:

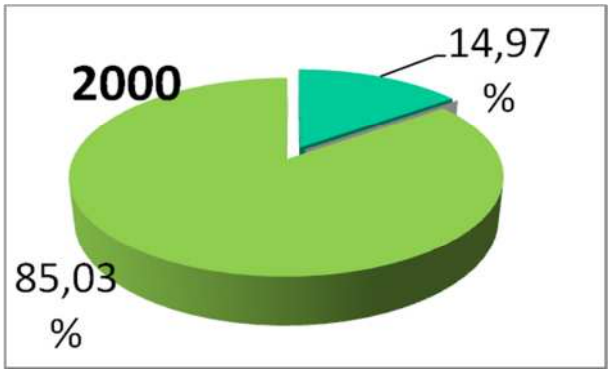


As we can see, the majority of cases are found in the academic genre. Still, to assume that what we see is distributional fragmentation, one should clarify if the construction was in the past (more or less) evenly spread across different genres and has, only for some time, been now squeezed into one particular genre. This is important to differentiate between a development in progress and a situation in which a given construction was born into a certain genre and then simply continued to stick to this genre.

The availability of big diachronic multi-genre corpora made it possible to trace changes in the distribution of grammatical constructions. The pie charts below show the ongoing change in the distribution of the already mentioned purpose subordinator *in order that*. The data was extracted from COHA and later processed to visualise a genre-related distribution of all the instances of *in order that* if each genre in COHA had an equal size (contrary to COCA, genres in COHA are of different sizes):







To sum up, we can see there seems to be a dynamic trend – a movement of nearly all instances of *in order that* towards the non-fiction genre. Still, what other interesting and useful pieces of information can we derive from these findings? Do some differences between genres play a role? The paper answers these questions and presents the findings and their potential implications in the framework of usage-based approach.

REFERENCES

Hundt, M. 2014. *The demise of the being to V construction*. Transactions of the Philological Society 112(2): 167-187.

Davies, Mark. (2010-) COHA: The Corpus of Historical American English: 400 million words, 1810- 2009. Available online at http://corpus.byu.edu/coha/.

Davies, Mark. (2013) COCA: Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. Available online at http://corpus.byu.edu/glowbe/.
Leech, G., Hundt, M., Mair, C., Smith, N., 2009. *Change in contemporary English. A grammatical study*. Cambridge: Cambridge University Press.

A diachronic study of the ABOUT/OF variation in communication verbs

Mateusz Sarnecki
University of Warsaw

Some English verbs are characterized by variation of their complementing preposition which introduces the topic of a communication event. For example, speakers can choose between the alternative forms *talk about* and *talk of*, both of which might be described as nearly synonymous.

Given this semantic similarity, it is interesting to try and account for the existence of the variation. One hypothesis states that the *about* and *of* constructions represent different views of the topic they introduce, with *about* indicating various aspects and *of* implying a more limited perspective, with the speaker focusing exclusively on the topic; cf. Vorlat (1982: 27), Dirven (1982: 60, 62), Lindstromberg (2010: 207).

Yet another explanation might be that the *about*/*of* variation is an instance of linguistic change in progress, with one of the topic-introducing prepositions becoming more widespread, and possibly ultimately replacing the other. While the semantic hypothesis has already been statistically tested on corpus data (cf. Krawczak — Glynn 2015), it appears that the diachronic account has not yet been examined in a similar way.

This study will target the topic-introducer variation in the verbs *speak*, *tell*, *talk*, and *write*. It will be based on the ARCHER corpus (ARCHER-3.2: 2013), which contains texts from 1600 until 1999. The procedure will involve performing statistical analyses of frequency counts of the *about*/*of* constructions to investigate the diachronic aspect of the variation.

As the corpus used here represents a cross section of registers and contains texts from British and American English, these sociolinguistic variables will also be included in the analysis. Additionally, the findings will be compared with analogous results for the mental verbs *dream*, *know* and *think* in order to provide a more complete account of the *about*/*of* variation.

REFERENCES

ARCHER 3.2 (2013) *A Representative Corpus of Historical English Registers* version 3.2. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. http://www.manchester.ac.uk/archer/.
Bretones Callejas, C.M. (2015) *Construals in language and thought: What shapes what?* Amsterdam: John Benjamins.
Dirven, René (1982) "*Talk*: linguistic action perspectivized as discourse". In: R. Dirven — L. Goossens — Y. Putsey — E. Vorlat (eds.), 37–84.
Dirven, René — L. Goossens — Y. Putsey — E. Vorlat (eds.), (1982) *The Scene of Linguistic Action and its Perspectivization by* speak, talk, say, *and* tell. Amsterdam: John Benjamins.
Krawczak, Karolina — Dylan Glynn (2015) "Operationalising construal. *Of / about* prepositional profiling for cognitive and communicative predicates". In: C.M. Bretones Callejas (ed.).
Lindstromberg, Seth (2010) *English Prepositions Explained*. Amsterdam: John Benjamins.
Vorlat, Emma (1982) "Framing the scene of linguistic action by means of *speak*", in R. Dirven — L. Goossens — Y. Putsey — E. Vorlat (eds.), 9–36.

The changing syntax and semantics of SPITE and THOUGH in written American English

Ole Schützler
Otto-Friedrich-Universität Bamberg

Concessive constructions are multi-faceted linguistic objects. First, different linkers can be selected (conjunctions, prepositions, conjuncts); second, at least three semantic (or pragmatic) types can be identified (Sweetser 1990): content (1), epistemic (2) and speech act (3a + b), with the possible addition of textual concessives (Crevels 2000). These types can be ranked according to their degree of subjectification (content → epistemic → speech-act → textual; Crevels 2000; cf. Traugott 1989), content concessives being the least subjectified. In content and epistemic concessives, there is an underlying presupposed mechanism of frustrated causality or conditionality, while in speech-act concessives, two pragmatic stances are contrasted.

(1) **Although** Carl had worked very hard, he failed the exam. [content]
(2) Carl had worked very hard, **although** he failed the exam. [epistemic]
(3) a. Keep working, Carl! – **although** I hardly need to say this. [speech-act]
b. Carl is a hard worker, **although** he's not very bright. [speech-act]

Finally, subordinate structures can occur in different positions relative to the matrix clause ('contrastive sequencing'; Altenberg 1986). Hardly anything is known about the diachronic and genre-related variation of those formal and semantic/pragmatic parameters.

This paper is based on data from the diachronic *Corpus of Historical American English* (COHA; Davies 2010–) which contains written material from the early 19th to the early 21st century, stratified into four genres (fiction, non-fiction, newspapers and popular magazines). The investigation focuses on the two macro variables SPITE (*despite* and *in spite of*) and THOUGH (*although*, *though* and *even though*). Specific questions are: (i) How frequent are the variables (or their variants); (ii) which semantic or pragmatic types of concession do they typically encode; and (iii) which syntactic position do subordinate clauses assume? The theoretically most interesting question is whether or not the answers to (i)–(iii) interact with diachronic and genre-related factors.

First results suggest that there is competition between *in spite of* and *despite* early in the period, with *despite* becoming the dominant form in the 20th century; moreover, *although* and *even though* generally increase in frequency. Based on findings from an earlier study of British English, the paper will test the hypothesis that the increase in *although* mainly reflects an increase of subjectified types associated with this linker, while *even though* and *despite*/*in spite of* come to be associated with content and epistemic types. Speech-act (and of course textual) concessives are predominantly found to the right of the matrix clause, while content and epistemic concessives are considerably more likely to be in non-final position. Indices of specialisation and stylistic distance are used to quantify genre-based differences in COHA. However, these differences prove to be rather subtle, probably due to the narrow stylistic range of the four genres in the corpus. More importantly, the paper shows that the semantic (or pragmatic) makeup of a concessive construction has an influence on the choice of connective and the syntactic arrangement of the sentence, and that the interplay of factors changes over time.

REFERENCES

Altenberg, Bengt. 1986. Contrastive linking in spoken and written English. In: Gunnel Tottie & Ingegerd Bäcklund (eds.) *English in Speech and Writing. A Symposium.* Stockholm: Almqvist & Wiksell. 13–40.

Crevels, Mily. 2000. Concessives on different semantic levels: A typological perspective. In: Couper-Kuhlen, Elisabeth & Bernd Kortmann (eds.) *Cause – Condition – Condition – Contrast. Cognitive and Discourse Perspectives.* Berlin: Mouton de Gruyter. 313–339.

Davies, Mark. 2010–. *The Corpus of Historical American English: 400 Million Words, 1810–2009*. Available online at http://corpus.byu.edu/coha/.

Sweetser, Eve E. 1990. *From Etymology to Pragmatics*. *Metaphorical and Cultural Aspects of Semantic Structure.* Cambridge: Cambridge University Press.

Traugott, Elisabeth C. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 65(1), 31–55.

Phraseology, lexical bundles and level of formality in American journalism

Ramón Martí Solano
Université de Limoges

Lexical bundles as "recurrent expressions" (Biber *et al.* 1999: 990) or "convenient routines" (Altenberg 1998: 122) are pervasive in natural languages in general and in all language registers and genres in particular. These multi-word units display not only pragmatic or discourse functions (Chen & Baker 2010: 30) but they play the overarching role of stylistic markers.

Contrary to Biber's Multi-dimensional approach (Biber 1995: 34), the goal of this paper is the analysis of individual phraseological strings, and more particularly lexical bundles and other word combinations labelled as 'formal' in dictionaries, as used diachronically in American journalism.

It is a well known fact that a large number of written registers, including newspaper language, has evolved towards a more popular speech-like style (Biber 2003: 169) but there is a type of phraseology, mainly used in formal written style and conveying particular stylistic connotations (Gläser 1998: 129), that has clearly declined as part of the popularisation process of the media at large.

It should be noted that phraseology has not traditionally been included among the linguistic features when classifying language change in general written registers (Biber & Conrad 2009: 116-117) or even in newspaper editorials (Westin 2002: 150). However, the use and frequency of various types of phraseology can change dramatically from genre to genre (Moon 1998: 68) and as "(p)hraseological units may contribute to the stylistic quality of texts in various registers" (Gläser 1986: 41), its inclusion appears to be fundamental for analysing stylistic changes and levels of formality.

For this study both the *Corpus of Historical American English* (COHA) and the *Time Magazine Corpus* (TMC) have been used. The former is a 400-million-word general corpus covering a 200-year time span from 1810 to 2009 and thus reflecting language change in American English. There exist four distinct genres, namely fiction, newspapers, popular magazines and non fiction books. The latter is a 100-million-word single-genre diachronic corpus spanning almost nine decades.

This paper addresses the issue of statistically-significant decrease (but also increase) over time of lexical bundles as style markers in journalism. There has been a steady and marked decrease in fixed word combinations such as *be that as it may, by virtue of, it is understood that, with one accord, in the bosom on, with one voice, through the agency of, by the name of, be at liberty to, by dint of, see fit to* and *have occasion to.* Nevertheless, *call into question, in sum, with dispatch, of every stripe* and *be no stranger to* have been clearly on the increase.

By way of illustration, the clausal lexical bundle *it is understood* yields 63 tokens in the 1920s compared to none between the 1970s and the 2000s in the TMC, which corroborates the decline of this type of stance markers as part of the "drift" towards more "oral" linguistic characterizations (Biber & Finegan 1989, 1992).

REFERENCES

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent wordcombinations. In A. P. Cowie (ed) *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press. 101–122

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

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

Biber, D. (2003). Compound noun-phrase structures in newspaper discourse: the competing demands of popularization vs. economy. In J. Aitchison & D. M. Lewis (eds) *New Media Language*. London: Routledge. 169-181.

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

Biber, D. & Finegan, E. (1989). Drift and the evolution of English style. A history of three genres. *Language* 65. 487-515.

Biber, D. & Finegan, E. (1992). The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. In M. Rissanen, O. Ihalainen, T. Nevalainen & I. Taavitsainen (eds) *History of Englishes: New methods and interpretations in historical linguistics*. Amsterdam: Mouton de Gruyter. 688-704.

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

Gläser, R. (1998). The Stylistic Potential of Phraseological Units. In A. P. Cowie (ed) *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press. 125-143.

Moon, R. (1998). *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Clarendon Press.

Westin, I. (2002). *Language change in English newspaper editorials*. Amsterdam: Rodopi.

Left-dislocated noun phrases in Modern English epistolary prose: a comparison with contemporary spoken Left Dislocation

David Tizón-Couto
University of Vigo

The aim of this paper is to offer a corpus-based analysis of the variation between the Modern English correlates of Left Dislocation [LD] in texts from genres that have been defined as speech-like (cf. Culpeper & Kÿto 2010), and to compare the results with the conclusions reported in previous studies of spoken LD (cf. Gregory & Michaelis 2001, Snider 2005). Example (1) is from contemporary spoken English, while (2) and (3) are from the letters and diaries included in the Penn-Helsinki Parsed corpora of Early Modern English (PPCEME; Kroch et al. 2004), Modern British English (PPCMBE; Kroch et al. 2010) and the Parsed Corpus of Early English Correspondence (PCEEC, 2006).

(1) *This girl* this morning **she** threw a wobbly. (Biber et al. 1999: 956)

(2) For *the privilege of Convocation* they intend not to infring **it**, (PCEEC, EModE2)

(3) But *those that are in a nearer alliance to the Divinity by a livening sense of the divine life in them*, there is a more speciall provision for **them** then for the ordinary sway of man-kinde. (PCEEC, EModE3)

A taxonomy of four left-dislocated [LDed] strings, which exhausts the types found in the data at hand (cf. (4) to (8)), constitutes the categorical dependent variable within a number of statistical tests that aim at describing both the grammatical and usage features of the different

types of ModE LDed strings in detail so that the results can be consistently compared to those reported for spoken LD. In analogy with the features reported for contemporary spoken LD (namely that it is short and nominal), the shortest and less postmodified LDed NPs are employed as the control group for the comparison. A linear model is also employed in order to test the effects of other predictive variables on the length/complexity (i.e. a continuous dependent variable) of the LDed constituent. The results suggest that shorter items are closer to the general behavior formerly reported for contemporary spoken LD. In contrast, LDed relative clauses or NPs postmodified by a relative clause (cf. (3)) differ in several respects from the features of the shorter control type. Despite the apparent distinction, LDed items that contain a *wh*-clause share two central features with the control type: they are recoverable in terms of information status; they also promote continuity in that they are often picked up as pronouns or full NPs in the immediately subsequent text. This correspondence highlights that no particular given information status might clearly define the LD construction, and that its discourse functions should be established by means of qualitative analysis (cf. Netz & Kuzar 2007).

On the diachronic plane, LDed items follow a statistically significant decline in epistolary prose, in line with previous reports covering a wider range of genres (Pérez-Guerra & Tizón-Couto 2009, Los & Komen 2012, Tizón-Couto 2015). The low percentage of tokens in letters and diaries, in comparison with other speech-purposed texts in the same corpora, such as sermons or drama, suggests that LDed NPs do not constitute unplanned devices that might signal a higher degree of communicative immediacy in a genre that has been defined as speech-like (Culpeper & Kÿto 2010).

REFERENCES

Baayen, Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Culpeper, Jonathan & Merja Kytö. 2010. *Early Modern English dialogues. Spoken interaction as writing*. Cambridge: Cambridge University Press.

Gregory, Michelle & Laura Michaelis. 2001. Topicalization and left-dislocation: a functional opposition revisited. *Journal of Pragmatics* 33. 1665-1706.

Kroch, Anthony Beatrice Santorini and Lauren Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, http://www.ling.upenn.edu/hist-corpora/, (30 June, 2014).

Kroch, Anthony Beatrice Santorini and Ariel Diertani. 2010. The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, http://www.ling.upenn.edu/hist-corpora/ (30 June, 2014).

Los, Bettelou & Erwin Komen. 2012. Clefts as resolution strategies after the loss of a multifunctional first position. In *The Oxford Handbook of the History of English*, Terttu Nevalainen and Elizabeth Closs Traugott (eds.). New York: Oxford University Press, 884-898.

Netz, Hadar & Ron Kuzar. 2007. Three marked theme constructions in spoken English. *Journal of Pragmatics* 39. 305-335.

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

Pérez-Guerra, Javier & David Tizón-Couto. 2009. On left dislocation in the recent history of English: Theory and data hand in hand. In *Dislocated Elements in Discourse: Syntactic, Semantic, and Pragmatic Perspectives*, Benjamin Shaer, Philippa Cook, Werner Frey and Claudia Maienborn (eds.). London: Routledge, 31-48.

Snider, Neal. 2005. A corpus study of left dislocation and topicalization. Stanford University: Linguistics Department, TS, http://www.stanford.edu/~snider/ pubs/qp1.pdf, (30 April, 2008)

Tizón-Couto, David. 2015. A corpus-based account of left-detached items in the recent history of English: Left Dislocation vs. Left Detached-sequences. *English Text Construction* 8 (1). 21-64.

A comparison of multi-genre and single-genre corpora in the context of contact-induced change

Carola Trips & Achim Stein
Universität Mannheim / Universität Stuttgart

This paper discusses results from a quantitative study of possible contact-induced change in Middle English in a multi-genre corpus (the Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2), Kroch and Taylor, 2000) and a single-genre corpus (the Penn Corpus of Early English Correspondence (PCEEC), Taylor et al., 2006) and shows that the data from the correspondence corpus is critical to understand the rise of the recipient passive (RP) in late ME. It further stresses the fact that this type of data reflects the active competence of writers, much more than other genres (at least in the PPCME2) do.

The phenomenon under investigation is the rise of the recipient passive in the history of English. Quite a number of authors see it in relation to or as a consequence of general changes on the level of (morpho)syntax (cf. Estival, 1989: 25, Los 2009, Seoane 2006). We will take a different stance and explore whether the rise of the RP may be an instance of contact-induced change with Old French (OF).

Allen (1995) provides a comprehensive account of this type (and other types) of passive. For her, clear cases of RPs are passives formed from ditransitive verbs where the recipient is the subject and the subject either occurs in the nominative or agrees with the finite verb. Allen found that in ME RPs scarcely occur, and she states that this cannot be explained by assuming that it spread from a particular subclass of ditransitives. Rather, 'bare objects directly following the verb were generally interpreted by the last quarter of the fourteenth century as direct objects, but these direct objects were not frequently passivized' (394-395). Interestingly, it is the low frequency of RPs that Allen considers as a contact phenomenon, since the model languages (Latin and French) do not allow RPs. For the same reason, 'the first examples of recipient passives are found in texts not aspiring to a polished literary style.' (395). The RP becomes more frequent in the course of the fifteenth century, and it is 'quite common' (395) by the early sixteenth century.

Our quantitative study of data from the PPCME2 confirms that RPs were not common before 1400 and even in the 15th century they rarely occur. According to Allen they are 'most common in less literary texts, such as letters' (Allen, 1995: 393), and this is why we include the single-genre corpus of the PCEEC (comprising 84 letter collections, 4970 letters, ca. 2.2 mil. words). It was searched for passive ditransitive verbs with the 'dative' argument in subject position by specifying that the passive verb governs a subject and a direct object.

We found that the RP occurred more frequently than in the PPCME2. Moreover, it occurred with a much higher frequency with French verbs than with native verbs (10.68% with French verbs, and 0.13% with native verbs). This is surprising because, as noted above, (Old) French does not have RPs. In our talk, we are going to argue that our data reflect a verb class specific passive construction that seems to be firmly established in the grammar of the writers. This construction is not calqued from the model language (OF) but the result of interpreting the French dative as different from the English 'dative'. We assume that this difference can be captured by the contrast between inherent vs. structural dative case (cf. Woolford, 2006).

REFERENCES

Allen, C. 1995. Case Marking and Reanalysis: Grammatical Relations from Old to Early Modern English. Oxford: Oxford University Press.

Estival, D. 1989. “A diachronic study of the English passive”. Diachronica 6 (1): 23–54.
Kroch, A. and Taylor, A., (eds). 2000. The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2). Philadelphia: University of Pennsylvania.
Los, B. 2009. “The consequences of the loss of verb-second in English: information structure and syntax in interaction”. English Language and Linguistics 13: 97–125.
Seoane, E. 2006. “Information Structure and word order change: the passive as an information-rearranging strategy in the history of English”. In The Handbook of the History of English, A. van Kemenade and B. Los (eds), 360–391. Blackwell.
Taylor, A., Nurmi, A., Warner, A., Pintzuk, S. and Nevalainen, T., (eds). 2006. Parsed Corpus of Early English Correspondence (PCEEC). York and Helsinki: Universities of York and Helsinki.
Woolford, E. 2006. “Lexical case, inherent case, and argument structure”. Linguistic Inquiry 37: 111–130.

The Corpus of Historical Low German: A tagged and parsed corpus of historical Low German

Sheila Watts *Oxford University*
Anne Breitbarth *Ghent University*
George Walkden *University of Manchester*
Melissa Farasyn *Ghent University*

"Die Syntax des Mittelniederdeutschen ist weitgehend unerforscht. [...] Untersuchungen zur mnd. Syntax sind ein dringendes Desiderat"[1] (Peters 1973:105). While finally, this pressing need is beginning to be addressed (e.g. Mähl 2004; 2014, Sundquist 2007, Petrova 2011; 2012, Tophinke and Wallmeier 2011, Breitbarth 2013; 2014a, Wallmeier 2012), the study of historical Low German syntax more generally (including Old Saxon/Old Low German (OLG)) is still only in its infancy, and lags far behind comparable research on other Germanic languages. Exploratory studies with a Middle Low German (MLG) focus have highlighted the intermediate position of MLG within Continental West Germanic (CWG), identifying it as a crucial missing link (e.g. Breitbarth 2014b). Furthermore, the (lack of) continuity between OLG and MLG has been an issue of scholarly dispute: recent studies e.g. on the distribution of (different types of) referential null subjects in OLG and MLG (Walkden 2014, Farasyn/Breitbarth 2015, respectively) may help making a case for continuity.

One of the factors still holding back research is the lack of availability of parsed corpora facilitating reliable, reproducible research into the diachronic syntax of historical Low German. This gap will be filled by the Corpus of Historical Low German (CHLG)[2], currently under construction with support from the Hercules Foundation.[3] The CHLG covers both OLG (c. 800–1050) and MLG (c. 1250–1600). Most of the OLG subcorpus is already fully tagged and parsed, while c. 50% of the MLG subcorpus has been parts-of-speech (PoS) and morphologically tagged.

In this report, we will present our intermediate results from constructing the corpus, the text selection, tagging and parsing. We focus particularly on the choice of different genres to address variation within MLG. Texts were selected to meet three key criteria, namely that they are a) in prose, b) not translated and c) clearly dated and localized. Clearly, these criteria cannot be applied to OLG, where almost the entire corpus consists of alliterative verse texts, which cannot be securely dated and placed. For MLG, however, such texts belong to the key text types in the language, charters and legal documents, which have been selected alongside narrative texts including religious and medical prose. For this, CHLG is collaborating closely

---

1 "The syntax of Middle Low German is largely unexplored [...] Research into Middle Low German syntax is an urgent desideratum"
2 http://www.chlg.ac.uk
3 http://www.herculesstichting.be/in_English/index.php

with the *Referenzkorpus Mittelniederdeutsch/Niederrheinisch* (1200-1650) (ReN)[4], which is PoS and morphologically tagged but not parsed. The range of genres is intended to counter the effects of (supra)regional standardization and formulaic language characteristic of charters in particular. The texts come from selected data points, representing the more influential scribal languages of MLG. As the corpus is dated and localized, it will facilitate the study of diachronic change and diatopic variation in the Low German language area. Also, measurement of the linguistic impact of contact between the Germanic languages, most notably through the Hanseatic League, in particular on Dutch, is still a largely neglected research area, which the text selection of the CHLG will help to remedy.

REFERENCES

Breitbarth, A. 2013. Indefinites, negation and Jespersen's Cycle in the history of Low German. Diachronica 30,171–201.

Breitbarth, A. 2014a. The History of Low German Negation. Oxford: Oxford University Press.

Breitbarth, A. 2014b. Dialect contact and the speed of Jespersen's cycle in Middle Low German. Taal en Tongval 66(1), 1–20.

Farasyn, M. and A. Breitbarth. 2015. Null Subjects in Middle Low German. Paper presented at the Diachronic Generative Syntax conference (DiGS) 17, University of Iceland, Reykjavík, May 2015.

Mähl, S. 2004. Studien zum mittelniederdeutschen Adverb. Köln/ Weimar/Wien: Böhlau. (Niederdeutsche Studien 49).

Mähl, S. 2014. Mehrgliedrige Verbalkomplexe im Mittelniederdeutschen. Ein Beitrag zu einer historischen Syntax der Deutschen. Köln/ Weimar/Wien: Böhlau. (Niederdeutsche Studien 57).

Peters, R. 1973. Mittelniederdeutsche Sprache. In J. Goossens, *Niederdeutsch. 1: Sprache und Literatur: eine Einführung* (pp. 66–115). Neumünster: Wachholtz.

Petrova, S. 2011. The Syntax of Middle Low German. Habilitationsschrift, Humboldt-Universität zu Berlin.

Petrova, S. 2012. Multiple XP-Fronting in Middle Low German root clauses. Journal of Comparative Germanic Linguistics 15, 157–188.

Sundquist, J.D. 2007. Variable Use of Negation in Middle Low German. In J. Salmons and S. Dubenion-Smith (Eds.), Historical Linguistics 2005 (pp. 149–166). Amsterdam: Benjamins.

Tophinke, D. and N. Wallmeyer. 2011. Textverdichtungsprozesse im Spätmittelalter: Syntaktischer Wandel in mittelniederdeutschen Rechtstexten des 13.-16. Jahrhunderts. In S. Elspaß and M. Negele (Eds.), Sprachvariation und Sprachwandel in der Stadt der Frühen Neuzeit (pp. 97–115). Heidelberg: Winter.

Walkden, G. 2014. Syntactic Reconstruction and Proto-Germanic. Oxford: Oxford University Press.

Wallmeier, N. 2012. Uneingeleitete Nebensätze mit konditionaler Semantik im Mittelniederdeutschen. In R. Langhanke, K. Berg, M. Elmenthaler and Jörg Peters (Eds.), Niederdeutsche Syntax (pp. 33-55). Hildesheim: Olms.

Electronic Repository of Greater Poland Oaths 1386-1444 (ROThA): Lessons in mark-up design

Matylda Włodarczyk
Adam Mickiewicz University

Joanna Kopaczyk
Adam Mickiewicz University / University of Edinburgh

Elżbieta Adamczyk
Adam Mickiewicz University / University of Wuppertal

This paper presents a project in progress, the first *Electronic Repository of Greater Poland Oaths* (ROThA) covering the period 1386-1444 (National Science Centre grant, 2015-2018). The Greater Poland court oaths are the oldest extant collection of secular texts written largely in Polish, hence their central significance for the history of the language, in

[4] https://vs1.corpora.uni-hamburg.de/ren/

particular in its 'national' dimension. However, the prevailing monolingual perspective on the data (e.g. Czachorowska 1988, Trawińska 2009) disregards the profoundly multilingual nature of scribal practices in Europe at the time (see Adamska 2013, Kopaczyk 2013).
The main focus of the corpus under construction is to characterise multilingualism as a social and cultural phenomenon in late mediaeval Greater Poland and its representation in the court oaths, with specific focus on Polish-Latin and Latin-Polish code-switching (see Pahta 2012 for a recent overview of code-switching in historical texts). The starting point for text selection is the philological edition of the material which covers 6350 oaths from six different locations (Kowalewicz and Kuraszkiewicz 1959-1981). The structure of the corpus has already been established with regard to the potential research questions, representativeness and balance, so the next step in corpus design focuses on how to capture and categorize the instances of CS.

As no mark-up schemes exist to serve the specific purpose of the corpus, i.e. the representation of multilingualism in the mediaeval court oaths in Greater Poland, we first discuss the general principles and standards behind the design of corpus metadata, including mark-up and tagging systems. We go on to provide an overview of potentially relevant schemes developed for other corpora and their customisation for the purposes of ROThA. Then, we discuss multiple levels of mark-up and annotation following the recent advances in electronic text editions (Honkapohja et al. 2009, Marttila 2014) and how they apply to our material. Specifically, the discussion focuses on the mark-up scheme for default categories on the most general level of CS functions and presents some problematic cases with suggested solutions. We also touch upon the issue of multiple tagging, which is particularly sensitive in the case of specialised corpora serving multiple and diverse audiences (i.e. a multipurpose corpus) (Grund 2012).

REFERENCES

Adamska, Anna. 2013. Latin and three vernaculars in East Central Europe from the point of view of the history of social communication. In *Spoken and Written Language. Relations between Latin and the Vernacular Languages in the Earlier Middle Ages*, ed. by M. Garrison, A. Órban & M. Mostert, 325-364. Turnhout: Brepols.

Czachorowska, Magdalena. 1998. *System antroponimiczny Wielkopolskich rot sądowych* [The anthroponymic system of the Greater Poland court oaths]. Bydgoszcz: Wydawnictwo Uczelniane Wyższej Szkoły Pedagogicznej.

Grund, Peter. 2012. Textual history as language history? Text categories, corpora, editions, and the witness depositions from the Salem witch trials. *Studia Neophilologica* 84(1). 40–54.

Honkapohja, Alpo, Samuli Kaislaniemi and Ville Marttila. 2009. Digital editions for corpus linguistics: Representing manuscript reality in electronic corpora. In *Corpora: Pragmatics and discourse*, ed. Andreas H. Jucker, Daniel Schreier and Mariane Hundt, 451-475. Amsterdam: Rodopi.

Kopaczyk, Joanna. 2013. Code-switching in the records of a Scottish brotherhood in early modern Poland-Lithuania. *Poznań Studies in Contemporary Linguistics* 49(3). 281-319.

Kowalewicz, Henryk and Władysław Kuraszkiewicz (eds.). 1959–1981, *Wielkopolskie roty sądowe XIV–XV wieku* [The Greater Poland court oaths of the 14th-15th century], Vol. 1: Roty poznańskie [The Poznań oaths], Vol. 2: Roty pyzdrskie [The Pyzdry oaths], Vol 3: Roty kościańskie [The Kościan oaths], Vol. 4: Roty kaliskie [The Kalisz oaths], Vol. 5, A: Roty gnieźnieńskie [The Gniezno oaths], B: Roty konińskie [The Konin oaths]. Warszawa–Poznań–Wrocław–Kraków–Gdańsk: Państwowe Wydawnictwo Naukowe.

Marttila, Ville. 2014. Creating Digital Editions for Corpus Linguistics: The Case of *Potage Dyvers*, A Family of Six Middle English Recipe Collections. PhD dissertation, University of Helsinki. Available online in helda.helsinki.fi/bitstream/handle/10138/135589/Marttila_PhDThesis.

Pahta. Päivi. 2012. Code-switching in English of the Middle Ages. In *The Oxford Handbook of the History of English*, ed. by Terttu Nevalainen and Elizabeth C. Traugott, 528-537. New York: Oxford University Press.

Trawińska, Maria. 2009. Cechy dialektalne wielkopolskich rot sądowych w świetle badań nad rękopisem poznańskiej księgi ziemskiej [Dialect features of the Greater Poland court oaths: The analysis of the Poznań municipal book manuscript]. *Prace Filologiczne* LVI. 345-360.

Plenary Talks

Classical and Modern Arabic Corpora: Genre and Language Change

Eric Atwell
University of Leeds

Forthcoming

Digital corpora and the study of historical text types: A case study in 19th and early 20th century German technical texts (Dinglers "Polytechnisches Journal", 1820-1932)

Thomas Gloning
Justus-Liebig-Universität Gießen

During recent years an increasing number of electronic texts and corpora covering the whole range of the history of the German language from its Old High German beginnings to contemporary language use have become available. Moreover, linguistic tools, annotation software and metadata have become increasingly sophisticated, while infrastructure projects like CLARIN-D aim at providing interoperable, standardized and sustainable data and tools. The „Deutsches Textarchiv" (German text archive) of the Berlin-Brandenburg Academy of Sciences and Humanities comprises a large and balanced corpus of texts from the 17th to the 19th centuries. The core corpus has recently been enriched by a number of additions from other projects, among them the Dingler corpus, an electronic version of Dingler's „Polytechnisches Journal" (1820-1932), one of the most important sources in the fields of engineering, technology, industrialization and technical innovation.

While there are a number of reliable and fruitful applications of computational tools in historical linguistics, there is equally a broad range of fields of investigation where it is far from clear if and how electronic texts and computational tools can improve research in these fields. Disambiguation in old texts, analysis of the textual structure of historic documents, topic organization, forms of argumentation and changes in vocabulary structurre over time are examples. In addition, there is the problem how digital results gained from large historical corpora in an automatic or semi-automatic way can be controlled.

In my talk I will analyze the Dingler Corpus (1820-1932) by starting out from „traditional" research questions on diachronic text types for special purposes, e.g.:

(1) Textual organization

− Which kind of text types are used in the journal? How do they evolve over time?
− How can we describe these text types in respect of different parameters of textual organization?
− What are important functional units in different text types (e.g. descriptions of new machines)?
– How do typical word uses and syntactic patterns contribute to the communicative functions of different texts?
– …

(2) Word usage and vocabulary structure

− Which word usages serve specific purposes relative to the subject fields of engineering, technology and industrialization? How can specific usages from one sub-field be distinguished from other usages (e.g. *Schifflein* ‚shuttle in weaving' vs. ‚little ship').
− How can the internal structure of the technical vocabulary be organized? How does this structure evolve over time?
− In which way does word usage contribute to textual organization (e.g. use of connectives, words used in headlines for topic management)?
− Which words are of foreign origin?
− How does word formation contribute to serve specific functions in texts?
− How can frequency profiles of words and vocabulary sections be linked to the functions and the topics of technical texts? How do these frequency profiles evolve over time?
− etc.

(3) Syntactic organization and patterning

− Which syntactic patterns are use regularly for specific functional units (e.g. locating a part of a machine)?
− What kind of differences are there relative to non-technical texts of the time?
− What kinds of lexical items are used to fill different ‚slots' in syntactic patterns for specific purposes (e.g. adjectives of colour in complex noun phrases)?
− How does the valency of technical verbs contribute to the organization of sentences in different text types?
− etc.

In trying to answer these kinds of linguistic questions with the help of the Dingler Corpus and the DDC tool provided by the „Deutsches Textarchiv" it is my aim both to show some of the possibilities to investigate such a corpus that one simply cannot „read through". On the other hand, it is important do diagnose limitations and thereby to stimulate the development of new tools or new kinds of applications of the existing DDC tool.

Academic Writing as a Locus of Grammatical Change: The Development of Phrasal Complexity Features

Bethany Gray
Iowa State University

Based on large-scale corpus analysis, this talk challenges the notion that academic writing is conservative and largely resistant to change by documenting linguistic innovations that have emerged in academic writing over the past 200 years. Starting with the characteristic grammatical features of modern academic writing which distinguish this register[5] from other spoken and written registers, the talk explores the dramatic patterns of change that have resulted in the present-day discourse style of academic writing.

The nominal style of modern academic prose has been well-documented, with a focus on the importance and prevalence of nouns, nominalizations, and structures associated with noun phrases (e.g., Biber, 1988; Wells, 1960; Halliday, 2004; Biber et al. 1999). Yet, previous research on linguistic change in writing has often focused on features related to the

[5] I use the term *register* here in the sense of Biber & Conrad (2009), in which *register* refers to a variety of language that can be defined based on its non-linguistic or 'situational' characteristics, such as mode, communicative purpose, topic, degree of interactivity, relationships between participants, ability to edit, and so on. In a register perspective on language variation, differences in linguistic structure can be attributed to these non-linguistic features of a language or text variety.

'colloquialization' of writing (e.g., progressive verbs, semi-modals, etc.) and has found little change in academic writing compared to popular registers like fiction and news (e.g., Hundt & Mair, 1999; Mair, 2006; Leech et al., 2009).

For these reasons, the primary focus of the talk is on the analysis of phrasal complexity features associated with a nominal discourse style (e.g., attributive adjectives, nouns as nominal pre-modifiers, prepositional phrases as noun post-modifiers, appositive noun phrases). The talk demonstrates that modern academic writing regularly employs phrasal complexity features that were minimally used in earlier historical periods and which have largely not been adopted in other spoken and written registers. The results thus challenge not only the notion that academic writing is resistant to change, but also the claim that grammatical innovation originates primarily in speech.

Furthermore, the talk explores variations in the historical trajectories of these changes, showing that specialist science writing has adopted these phrasal complexity features to a much greater extent than humanities writing or non-specialist science writing. This research demonstrates the importance of considering the role of register in studies of linguistic change, as such considerations have implications for historical corpus research, including the selection of features to investigate and the design of diachronic corpora.

REFERENCES

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

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.Halliday, M.A.K. 2004. *The language of science.* London: Continuum.

Hundt, M. & Mair, C. (1999). '"Agile" and "uptight" genres: The corpus-based approach to language change in progress.' *International Journal of Corpus Linguistics, 4*, 221-242.

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

Mair, C. (2006). *Twentieth-century English: History, variation and standardization*. Cambridge University Press.

Wells, R. (1960). 'Nominal and verbal style', in T.A. Sebeok (ed.), *Style in language*, pp. 213-220. Cambridge University Press.

Genre changes in English medical writing 1375-1800

Irma Taavitsainen
University of Helsinki

Late medieval scientific and medical writing in English had several different genres and levels of writing from the beginning, as the remedy book tradition continued and new genres were introduced from Latin with the vernacularization boom. My presentation focuses on the developments of top genres of scholasticism, introduced after 1375, and their afterlives in the centuries that follow. In scholastic science the function of commentaries was to construct original knowledge of ancient authorities. The genre was created for use at universities as a research-based teaching aid that explicates a base text, cross-refers to other authors on the same topic, and gives the commentator's own opinion at the end. Commentaries acquired a conventionalized form early in Latin, but vernacular translations exhibit a looser and more practical form. Compilations were another academic genre that listed text passages of a topic under discussion, and the two top genres merged in the late medieval period. However, fully-fletched commentaries emerge in the early modern period when the heyday of logocentric scholastic science was already over. This shows that writing conventions were adopted with a time lag when the thought-style was already changing.

Research became increasingly based on observation and experimental reports, based on totally different principles, developed for the use of the new Royal Society discourse community. But changes are gradual, and the genre map becomes more complicated in the course of time as commentary/compilation features continue far beyond the early modern period. The new genres of empiricism have received a great deal of attention, but little has been written about the earlier top genres and their afterlives, and large areas are still uncharted.

I shall begin my presentation with a theoretical part of genre change. Relevant passages for an empirical corpus study are located with lexical searches of scholastic phrases and collocations. They are defined on the basis of earlier studies and by close reading of texts; thus both top-down and bottom-up corpus-driven approaches are used. An overall picture can be achieved with quantification, but my main approach is qualitative with contextual analyses of meaning-making practices in their sociocultural context. The sociolinguistic parameters of authors' education and levels of audience are taken into account. My research questions focus on generic features derived from scholasticism, whether they changed and in what way, and how they continue in later writings and other genres, such as textbooks and medical treatises. The aim is to achieve an in-depth view of the developments and cultural genre dynamics of medical writing in a diachronic perspective. The material comes from a large and systematically collected database of three corpora compiled by the Scientific thought-styles team at the University of Helsinki. These corpora provide a full scale of texts with background metadata from academic treatises to writings targeted at heterogeneous lay people.

REFERENCES

Middle English Medical Texts 1375-1500 (MEMT, 2005). Compiled by Irma Taavitsainen, Päivi Pahta and Martti Mäkinen. Amsterdam: John Benjamins. CD-ROM.

Early Modern English Medical Texts 1500-1700 (EMEMT, 2010). Compiled by Irma Taavitsainen, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr and Jukka Tyrkkö, with software by Raymond Hickey. Amsterdam: John Benjamins. CD-ROM.

Late Modern English Medical Text 1700-1800 (LMEMT, forthcoming). Compiled by Taavitsainen, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Päivi Pahta, Maura Ratia, Carla Suhr and Jukka Tyrkkö.