Martin Hilpert* and Hubert Cuyckens

How do corpus-based techniques advance description and theory in English historical linguistics? An introduction to the special issue

1 Introduction

This special issue brings together six contributions that showcase different corpus-based approaches to the study of historical developments in English. Each of the studies offers new empirical results on a given phenomenon of language change, but when viewed in their mutual contexts, the papers serve to illuminate the unifying question that is given in the title of this introduction. It is clear that during recent years, both corpus-linguistic resources and analytical techniques have been evolving at a remarkable rate. What is perhaps less clear is how the use of new resources and the application of new techniques can be put into the service of transforming our knowledge of how the English language changes. Beyond giving us more depth and precision, what do larger corpora and more sophisticated methodologies bring to the table in terms of description and theory?

   Despite all innovations, it is important to remember that the current developments in English historical corpus linguistics form part of a tradition that has been on-going for some time, and that owes much to the creation of the Helsinki corpus (Kytö 1991), and also to the long and fruitful connection between corpus linguistics and grammaticalization studies (Lindquist and Mair 2004). More and more diachronic resources have become available in the meantime, among them ARCHER (Biber et al. 1994), the Penn Parsed Corpora (Kroch et al. 1997), the

*Corresponding author: Martin Hilpert, Université de Neuchâtel, Espace Louis-Agassiz 1, CH-2000 Neuchâtel, Switzerland, E-mail: martin.hilpert@unine.ch
Hubert Cuyckens, KU Leuven, Blijde-Inkomststraat 21 - Box 3308, 3000 Leuven, Belgium, E-mail: hubert.cuyckens@arts.kuleuven.be

Corpus of Early English Correspondence (Nurmi et al. 1996), the Corpus of Late Modern English Texts (De Smet 2005), Mark Davies' suite of diachronic corpora (Davies 2007, 2010), and the Old Bailey Corpus (Huber 2007). With regard to theory, a growing number of diachronic studies have adopted ideas from construction grammar (Traugott and Trousdale 2013), often paired with a corpus-based methodology. As a consequence, diachronic corpus linguistics has developed into a topic of considerable methodological and theoretical interest. Studies on the basis of these resources hold many theoretical implications that as yet have not been fully explored. With regard to these implications, on-going work has chiefly been applied in theoretical frameworks such as usage-based linguistics (Bybee 2007), grammaticalization theory (Hopper and Traugott 2003), construction grammar (Goldberg 2006), and quantitative variationist linguistics (Labov 2001). As the following brief descriptions will reveal, also the papers in this special issue can be broadly situated in these frameworks.

2 The contributions in this special issue

Setting the tone for the special issue, the first paper by Marie José López-Couso illustrates the state of the art with regard to corpus-based studies of grammaticalization. She discusses three empirical case studies that demonstrate the benefit of using corpus-based methodologies for the investigation of questions that are relevant to grammaticalization theory. The first of these addresses existential *there*, compares its development in the history of English against patterns of usage during language acquisition, and finds intriguing similarities. The second case study probes the question how corpus data can guide the detection of incipient grammaticalization. An analysis of parenthetical clauses with *like* (*He's scared to debate, it looks like*) in recent American English points to several processes that are indeed indicative of on-going grammaticalization. Thirdly, the paper tackles the grammaticalization of low-frequency constructions (Hoffmann 2004, Mair 2004), which is a phenomenon that is inherently problematic for current standard views of grammaticalization. The example that is chosen to illustrate this is the use of *namely* as a marker of an apposition (*He carried an offensive weapon, namely a crowbar*). Diachronic corpus data reveal that this use of *namely* is a relatively recent phenomenon, and that other functions of *namely* are on the decline.

The contribution by Christopher Shank, Koen Plevoets, and Julie van Bogaert illustrates how diachronic corpus studies can benefit from the adoption of current variationist methodology. The authors present an account of

the alternation between *that* and zero following the verbs *think*, *believe*, and *suppose*. The analysis covers a time span from the sixteenth to the twenty-first century and takes into account eleven structural features that previous research has identified as conditioning factors in speakers' choices between *that* and zero. Among the empirical observations that Bogaert et al. offer, a result with particular theoretical significance is that diachronically, the zero variant is not on the rise, but rather on the decline. This finding casts doubt on accounts of zero complementation as the end result of a reductive grammaticalization cline, and provides support for alternative explanations (e.g. Brinton 1996, 2008).

Britta Mondorf studies verbal constructions that contain a non-referential pronoun *it* in the position of an object, as in *leg it*, *snuff it*, or *beat it*. Constructions with a dummy *it* exhibit several idiosyncratic traits with regard to their argument structure, the pronoun fails standard tests for objecthood. Mondorf argues that the rise of constructions with dummy *it* needs to be understood against the background of more general diachronic processes that are currently transforming the verbal grammar of English, notably processes of transitivization and detransitivization. The function of dummy *it* in this regard is to increase the transitivity of verbs that are normally used intransitively.

Javier Perez Guerra offers a diachronic study of word order that examines the relative placement of modifiers and complements in English verb phrases and noun phrases. Drawing on the work of Hawkins (1994, 2004), the analysis focuses on the dynamics between two determinants of relative word order, namely the syntactic principle of 'complements first' and the processing-related principle of end weight. Contrasting verb phrases and noun phrases in their respective usage patterns across Middle, Early Modern and Late Modern English, it becomes apparent that the principle of end weight shows a strong effect throughout. In the verbal data, a historically increasing effect of the 'complements first' principle makes itself felt; the nominal data fail to show a systematic tendency. One theoretical conclusion that can be drawn from these observations pertains to the relative prototypicality of modifier-head constructions. Despite structural parallelisms, verbs appear to be more typical syntactic heads than nouns.

Tanja Säily investigates diachronic changes in the productivity of the English suffixes *–ness* and *–ity* during the eighteenth century. The study draws on and extends a method for the study of productivity changes (Säily and Suomela 2009), determining whether social factors such as gender or social rank correlate with greater or lesser use of the respective word formation processes in the Old Bailey Corpus. The results partly contradict earlier findings about gendered differences in the use of *–ity*, and they raise issues with regard to corpus periodization and the practice of multiple hypothesis testing.

Benedikt Szmrecsanyi makes the argument that frequency changes in diachronic corpus data should not automatically be taken as evidence for grammatical change. An alternative explanation that should always be considered is that changes in the cultural environment of the respective texts may have boosted or suppressed the use of certain linguistic structures. The example that is used to illustrate the problem is the English genitive alternation, that is, the variability between the *s*-genitive and the *of*-genitive. As is well-known, the *s*-genitive has a marked preference for animate possessors (*John's watch*, *the captain's office*). An analysis of Late Modern English data reveals that if fewer animate referents appear in a text, the frequency of use of the *s*-genitive is trivially depressed. That is, despite the absence of any grammatical change, cultural developments may have a tangible effect on corpus frequencies.

3 Towards new questions

To summarize, the studies in this special issue showcase the broad range of corpus-based approaches to diachronic English linguistics that are currently available, including several techniques that represent genuinely new methodological developments. In keeping with the overall aim of this special issue, the papers translate the insights gained by such empirical methodologies into a fruitful discussion of theoretical issues. The availability of new resources and new methods thus yields an added value; we are not only able to answer old questions with more precision, we can actually begin to ask – and answer – new questions that simply could not have been asked in this way only a few years ago.

References

Biber, Douglas, Edward Finegan & Dwight Atkinson 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In Udo Fries, Gunnel Tottie & Peter Schneider (eds.), *Creating and using English language corpora. Papers from the fourteenth international conference on English language research on Computerized Corpora, Zürich 1993*, 1–13. Amsterdam: Rodopi.

Brinton, Laurel J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: Mouton de Gruyter.

Brinton, Laurel J. 2008. *The comment clause in English: Syntactic origins and pragmatic development*. Cambridge: Cambridge University Press.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Davies, Mark. 2007. TIME Magazine Corpus (100 million words, 1920s-2000s). Available online at http://corpus.byu.edu/time.

Davies, Mark. 2010. The corpus of historical American English. Available online at http://corpus.byu.edu/coha.

De Smet, Hendrik. 2005. A corpus of Late Modern English texts. *ICAME Journal* 29. 69–82.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hoffmann, Sebastian. 2004. Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition. In Hans Lindquist & Christian Mair (eds.), *Corpus approaches to grammaticalization in English*, 171–210. Amsterdam: John Benjamins.

Hopper, Paul J. & Elizabeth C. Traugott. 2003. *Grammaticalization*, 2nd edn. Cambridge: Cambridge University Press.

Huber, Magnus. 2007. The old bailey proceedings, 1674–1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English. *Studies in variation, contacts and change in English* 1. http://www.helsinki.fi/varieng/series/volumes/01/index.html (accessed July 12, 2015).

Kroch, Anthony, Beatrice Santorini & Lauren Delfs. 2004. *The Penn-Helsinki parsed corpus of early modern English*. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/ (accessed 12 July 2015).

Kytö, Merja. 1991. *Manual to the diachronic part of the Helsinki corpus of English texts: Coding conventions and lists of source texts*, 3rd edn. Helsinki: Department of English, University of Helsinki.

Labov, William. 2001. *Principles of linguistic change. Volume II: Social factors*. Oxford: Blackwell.

Lindquist, Hans & Christian Mair (eds.). 2014. *Corpus approaches to grammaticalization in English*. Amsterdam: John Benjamins.

Mair, Christian. 2004. Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond. In Hans Lindquist & Christian Mair (eds.), *Corpus approaches to grammaticalization in English*, 121–150. Amsterdam: John Benjamins.

Nurmi, Arja, Ann Taylor, Anthony Warner, Susan Pintzuk & Terttu Nevalainen. 2006. *Parsed corpus of early English correspondence, tagged version. Compiled by the CEEC Project Team*. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

Säily, Tanja & Jukka Suomela. 2009. Comparing type counts: The case of women, men and – *ity* in early English letters. In Antoinette Renouf & Andrew Kehoe (eds.), *Corpus linguistics: Refinements and reassessments* (Language and Computers: Studies in Practical Linguistics 69), 87–109. Amsterdam: Rodopi.

Traugott, Elizabeth C. & Graeme Trousdale. 2013. *Constructionalization and constructional changes*. Oxford: Oxford University Press.