New developments in corpus linguistics and the study of language variation and change

Kristin Bech and Gisle Andersen

University of Oslo and NHH Norwegian School of Economics

It is the task of the corpus users to evaluate whether the corpus is suitable for their research questions. All corpora are not suitable for all types of research questions.

This mild admonition stems from Stig Johansson (2012: 54), in whose honour ICAME 32 in Oslo was organised, and it is a response to, or rather, a wholehearted agreement with, Meyer’s words of caution in his 2004 article ‘Can you really study linguistic variation in linguistic corpora?’. In this article, Johansson writes about social and situational factors in textual distribution, and the close relation between the research question(s) and the interpretation of corpus data. He further laments the apparent gulf between corpus linguistics and sociolinguistics, as evidenced in the fact that the Handbook of language variation and change (Chambers et al. 2002) only contained one article dealing with corpora. However, Stig Johansson, whose motto was jobba på ‘soldier on’, was not one prone to resignation, and his suggestions for the future include going beyond mere corpus linguistics to develop new methods and new theoretical frameworks for the study of language variation.

Recent developments in corpus linguistics do in, in fact, give reasons for optimism for the scholar looking to study language variation and change (Andersen 2010). One important advance is the availability of a new generation of comparable corpora, sometimes called ‘snapshot corpora’ (Claridge 2008; Gabrielatos et al. 2012), that use the same corpus design and sampling frame for corpora recorded at different points in time. From the Brown family of corpora, compiled in 1961 and 1991, the methodology known as (short-term) ‘diachronic comparable corpus linguistics’ has emerged, of which Leech et al. (2009) is probably the most substantial contribution to date. Elsness’ study (this volume) gives an example of how this research method can be used to study gender and cultural differences. Recent additions of new Brown family members, specifically the BLOB-1931 corpus (Leech 2012) and the BE06 corpus (Baker 2009), now allow for research that tracks changes in English language use that go beyond the original 30-year span in both directions. Similarly, this method has recently been applied within spoken language by the comparison of the COLT and the LIC/MLE corpora (Torgersen et al. 2011; Andersen forthcoming) representing London adolescent speech. Two measures are crucial in diachronic comparable
corpus linguistics, notably time span and sampling points, and granularity is a function of the number of time points within a given time span. The time span may vary to suit particular research objectives, from centuries, as in the LAEME/LALME comparison by Stenbrenden (this volume), via the more typical 30-year span of the Brown family of corpora, to the much narrower time span of contemporary monitor corpora such as COCA (Davies 2009) and the Norwegian Newspaper Corpus (Andersen and Hofland 2012). In monitor corpora, the number of sampling points, and thereby the granularity, is generally much higher than in other corpora, and these may even enable the day-to-day monitoring of lexical innovation and other features (Renouf 1993, 2007; Andersen and Hofland 2012).

Further, more advanced statistical techniques add to the usefulness of the wide range of corpus resources open for investigation. Although most studies rely on traditional means of testing the significance of frequency differences at different time points, methods such as cluster analysis (Gries and Hilpert 2008), regression analysis (Millar 2009) and frequency development analysis (Gabrielatos et al. 2012) offer innovative data-driven techniques for revealing fine-grained patterns and identifying stages in diachronic developments in which a particular language feature is prominent. Although more exploratory work is needed, these techniques have a potential for establishing the extent of correlation between the variables of time and frequency with other sociolinguistic variables.

Moreover, we believe that corpus linguistics and sociolinguistics have a lot to gain from inspecting new data that have become available online. This applies not only to large collections of written language such as COHA, Google n-grams, newswire archives etc., but also to new media such as video blogs and other self-recordings on the internet. These may be valuable, especially perhaps to studies in pragmatics and discourse, but also to the study of phonology and prosody variation. It is reasonable to believe that these new data may at least complement, although not replace, data gathered through carefully designed elicitation tasks, sociolinguistic field work and interviews.

Against this background, it is to be hoped that Johansson would have shared this optimism and been satisfied with the research questions, selections of corpora and interpretations of data found in the eleven studies that make up the present volume. Some of them indeed bridge the gap between corpus linguistics and sociolinguistics, and are thus evidence of a development that Johansson wished to see. As the book title suggests, we have collected papers that use corpora to study variation along three dimensions – time, place and genre. The volume is accordingly divided into three sections. The first section contains three rather diverse studies whose common feature is that they deal with language variation in time, i.e. in the history of English. The second section contains five studies that investigate variation in space, most of them by comparing particular varieties of English, including some “New Englishes” (Koch and Bernaisch this volume; Suárez-Gómez and Seoane this volume). In the third section are found three studies on variation in genre, as represented by medical articles, social reports and academic English.