A new corpus of spoken English: extending the diachronic scope of the DCPSE

As part of my PhD project, I am currently compiling a new corpus of spoken English designed to match the *Diachronic Corpus of Present-Day Spoken English* (DCPSE) (Aarts and Wallis 2006) and containing up-to-date spoken data from 2012 to 2016. Existing corpora of spoken British English either date back nearly 20 years (like the BNC or the spoken part of ICE-GB) or cover only individual genres (like the Hansard corpus). Thus, if one wants to investigate language variation and ongoing change, there is a clear need for a new corpus containing up-to-date spoken material covering different genres. Like the DCPSE, the corpus will consist of six genres ranging from conversational to non-conversational data: face-to-face conversations, broadcast discussions, broadcast interviews, spontaneous commentary, parliamentary debates and prepared speech. In the first stage of compilation, it will comprise roughly 500,000 words, which can later be expanded to provide a full match for the DCPSE.

For written English, the development of the BE06 (Baker 2009) and the Lancaster1931 corpus (Leech and Smith 2005) – both diachronic extensions to the British English corpora of the Brown family (LOB and FLOB) – has proven to be highly valuable for the study of recent language change. While some previously observed developments such as the increasing use of the progressive construction with modal auxiliaries could be corroborated, others, such as the frequency increase of the progressive passive were revealed as not being part of linear changes (Smith and Leech 2013:84ff.).

In my talk, I will address the question whether the frequency increase of the progressive construction (e.g. Mair and Hundt 1995; Leech et al. 2009; Aarts et al. 2010) is still under way in 21st-century spoken English. If yes, in which genres? And, most importantly, what are the driving forces behind potential frequency changes? Finally, I will address the question of how to deal with the implications of the relatively small size of the corpus and demonstrate how larger corpora can and should be used to complement the analysis of specific aspects of the progressive's development and use (cf. Mair 2006).

References


