Bringing back the past: Towards a diachronic corpus of spoken English

Jose A. Mompean
University of Murcia

Written corpora have traditionally been the main tool in corpus studies while spoken corpora are fewer in number and more limited in data. Moreover, most spoken corpora are synchronic. Thus, an adequate study of various phenomena related to the evolution of the spoken language is also limited. It seems, therefore, that more longitudinal or diachronic spoken corpora should be compiled and studied by the research community.

The current paper describes the process of building a longitudinal corpus of spoken English, the Diachronic Corpus of Spoken English (DIACSEN), which is an ongoing project at the University of Murcia. It provides an overlook of the steps in this process, including sampling criteria. A description is provided of the target population, including data such as the number of speakers, gender, dates of birth, ages at the dates of production, etc. The population is made up of speakers of Received Pronunciation (RP) or Standard Southern British Pronunciation (SSBP). This variety is, historically, the most widely available one in broadcasting in the United Kingdom, from which most materials are obtained. A description is also provided of the time covered (i.e. 1910s-2010s), the text types (scripted and non-scripted; dialogic and monologic; etc.) as well as the modality (spoken and multimodal), the sample per speaker, and the approximate size of the corpus (ca. 550,000 words).

DIACSEN may prove useful for researchers working, among other aspects, on sound change from a real-time perspective. This perspective is based on the analysis of data collected from different periods of time as opposed to the apparent-time approach of most sociolinguistic studies, which compares data from individuals of different ages at a specific moment in time. An example of the use of material from the DIACSEN corpus is then provided: the use of /r/-sandhi by Queen Elizabeth II. /r/-sandhi refers to the pronunciation of an r-sound between two adjacent heterosyllabic vowels, the first of which is [-high]. The most common type of /r/-sandhi is referred to as 'linking' /r/ when there is <r(e)> in the spelling (e.g., *here in* [ˈhɪəɹ‿ɪn]). Linking /r/ coexists with a non-etymological, non-orthographic type referred to as 'intrusive' /r/ (e.g. *idea in* [aɪˈdɪəɹ‿ɪn]).

Potential contexts of /r/-sandhi were identified and analysed for the presence or absence of rhoticity and glottalisation in a corpus of Christmas speeches over seven decades. The results show that the Queen avoids intrusive /r/ altogether but that she uses linking /r/ in most potential cases, that glottalisation is common when /r/-sandhi is not used, and that linking /r/ and glottalisation can also co-occur. A comparison with a longitudinal corpus of speakers also shows that the Queen converges towards group-level trends in the case of linking /r/ but diverges from those trends in the case of intrusive /r/.