Phonetic corpora and big data

Martine Adda-Decker

During the last years, 'big data' has emerged as a trendy, highly promising portmanteau term in economics and high-tech domains, such as information technology and speech processing. Big data are often described using a 3V scheme: volume, variety, velocity: a huge volume of data, a large variety of possibly unstructured, heterogeneous data sources, a high frequency or velocity of data generation over time. In this Glasgow ICPhS 2015 discussant session, we will question the 'big data' term with respect to phonetics and speech sciences at large. In this context, big data typically refer to huge, generally unstructured collections of speech or audio-visual data, pre-existing any phoneticians' investigation hypotheses. Can such data become beneficial to phonetic sciences?

Speech is known to be highly variable across time and space, speakers and communities, discourse situations and recording conditions. The quest for invariants or rules of systematic variation is one of the holy grails of speech scientists. Phonetic research thus makes extensively use of speech corpora. Getting easy access to large varieties of speech collections certainly represents a positive perspective. However, big data generally come unstructured, whereas phonetic research corpora tend to be very carefully designed, collected, annotated and processed by the phonetician in view of specific investigations. Speech is enriched with metadata describing the speakers' linguistic competences and socioprofessional backgrounds: variation needs to be related to hypothesized originating or facilitating factors. Unstructured data complicate the scientific usability of big data, especially in phonetics.

Recently, new technologies dealing with unstructured data became available, including data mining, text and speech analytics, machine learning... For example, the last decades witnessed tremendous progress in automatic speech processing. Very smart speech transcription systems are nowadays available in everyday smartphones. This progress entails at least two interesting outcomes with respect to phonetics research: (i) very large transcribed speech corpora become at reach in number of languages and (ii) automatic transcription or alignment systems may provide time-stamped linguistic annotations with limited human effort.

No papers of this session can be qualified as big data papers in the above exposed 3V sense. However, the papers all deal with large speech corpora, most of them collected out of realm of phoneticians' laboratories. Corpus content was not controlled with respect to a priori linguistic criteria, rather they were collected in somewhat controlled production situations reflecting spoken language usage in these conditions (public journalistic speech, broadcast and telephone conversations…).

The first paper by Bartkova addresses a methodological question: how to deal with very short sound segments in automatically aligned speech segmentations which are increasingly used in phonetic and linguistic studies. The paper investigates the effect of changes in acoustic analysis frame rates on phonetic segmentations using large French public radio and TV broadcast news and debates. Results show that especially for fast speech a frame shift impacts corpus-based phonetic analyses.

The second paper investigates phonological processes in Tokyo Japanese using a large corpus of broadcast speech. More particularly, Kilbourn-Ceron studies the effect of prosodic information on the known high vowel devoicing phenomenon in Tokyo Japanese across word boundaries. A special focus is given to the role of pauses, which may be part of the trigger for an alternation, or may block it.

The third paper by Chodroff examines voice onset times (VOT) in American English plosives (Mixer 6 corpus) in 130 speakers living in the Philadelphia region. Talkers were found to vary considerably in their production of VOT in word-initial stop consonants, extending previous results to more realistic speech. Results suggest talker-specific characteristics of phonetic realization to generalize across stop categories.

The last paper revisits the question dealing with the effect of word frequency on production. The proposed study by Sherr-Ziarko examines word usage frequency effects on the production of homophone words in Mandarin Chinese. The author makes use of 30 hours of Mandarin broadcast news to question the validity of previous results cross-linguistically. Results tend to indicate that lemma categorization plays a more important role in the organization of the lexicon than phonemic structure in Mandarin Chinese speakers.

Big data certainly rise high expectations and offer amazing opportunities in the next future to investigate almost any of the world's spoken languages. However, their successful use in phonetic sciences is not straightforward and even questionable. To take the best out of big data we have to think about new methodologies and need to include processing instruments relying on machine learning, data mining, automatic speech transcription and metadata production (language, dialect, speaker, emotion...) into the phonetician's toolbox.