The Corpus of Spontaneous Japanese and Its Application to the Study of Japanese Phonetics

Kikuo Maekawa (Dept. Corpus Studies and the Center for Corpus Development, NINJAL)

Since 1999, the author has been responsible for several corpus compilation projects including the Corpus of Spontaneous Japanese (CSJ, 7.5 million word, publicly available since 2004), the Balanced Corpus of Contemporary Written Japanese (BCCWJ, 100 million word, available since 2011), and the Ultra-Large Scale Corpus of Japanese (10 billion word, compilation underway). Among these corpora, the focus will be placed on the CSJ in the rest of this talk.

CSJ is a large-scale corpus of spontaneous speech consisting of about 662 hours of speech spoken by more than 1400 speakers. It was developed jointly by the NINJAL, NICT (National Institute of Information and Communications Technology), and Tokyo Institute of Technology. CSJ was designed primarily as a resource for the development of a next generation automatic speech recognition system that could handle more-or-less spontaneous monologues. In addition to this, the corpus had a secondary application area: the phonetics of spontaneous Japanese speech.

To meet these two requirements, CSJ is reasonably large and richly annotated. In particular, a 44 hour subset of the corpus called CSJ-Core was the most richly annotated spontaneous speech corpus of the world when it was released, and, presumably, remains so to this day. In addition to standard annotations like transcription, POS analysis, and clause-boundary classification, the CSJ-Core is annotated with respect to both segmental and prosodic characteristics using the X-JToBI labeling scheme. Dependency structure and topic structure annotations are also applied.

The last half of the talk is devoted to a demonstration of the CSJ's contributions to the phonetics of spontaneous Japanese. The first topic is the variation of /z/ between [z] and [dz], which is a long-standing issue in Japanese phonetics. Statistical analyses of the CSJ-Core revealed clearly that the variation was not an allophonic variation. It was rather a coarticulatory variation. More than 70% of the variation could be explained by a new phonetic measure called the TACA, which had nothing to do, in principle, with the location of the phoneme in a word. It turned out also that TACA was a good predictor of the weakening of voiced stops (i.e. [b]~[ß], [d]~[ð], and, [g]~[ɣ]).

CSJ is also useful for the study of prosody. Analysis of the PNLP (penultimate non-lexical prominence, a special variant of the LHL boundary pitch movement) provides a typical example. Statistical analyses revealed that PNLP occurred typically in the penultimate accentual phrase of an utterance, thereby predicting the termination of the utterance. This is a finding that can hardly be achieved without the analysis of spontaneous speech corpus. This and other examples show convincingly that there are many phonetic events that cannot be properly analyzed in experimental settings.

The last topic is the evaluation of the X-JToBI prosodic annotation scheme from the viewpoint of paralinguistic and/or non-linguistic information. Comparison of traditional JToBI and new X-JToBI annotations revealed the clear superiority of the X-JToBI both in terms of the prediction of speech registers (i.e., academic speech, public speaking, dialogue, and reading around), and the prediction of speakers' gender and age-groups. These results show convincingly that annotation of linguistic contrasts alone is not enough for the full understanding of the message involved in the speech signal.