Guest Editors' Introduction

Haruo Kubozono*, Kikuo Maekawa and Timothy J. Vance

Corpus-based approaches to the phonological analysis of speech

This special issue is a collection of selected papers from the 14th Conference on Laboratory Phonology, which was held at NINJAL (National Institute for Japanese Language and Linguistics) in Tachikawa, Tokyo, on July 25–27, 2014. The conference was sponsored by NINJAL with cooperation from four academic societies related to language and speech in Japan: the Acoustical Society of Japan, the Linguistic Society of Japan, the Phonetic Society of Japan, and the Phonological Society of Japan.

The general theme of the conference was "Laboratory phonology beyond the laboratory: Quantitative analyses of speech produced outside the phonetics laboratory". The papers selected for this special issue were drawn from the conference thematic session on corpus-based approaches, and present corpus studies of spontaneous speech, endangered languages, and L1 phonology/prosody.

Seven articles were selected and reviewed for inclusion in the present volume. **Mazuka et al.** and **Zellou and Scarborough** analyzed corpora of infant-directed speech. **Mazuka et al.**'s analysis is based on a corpus of Japanese mothers' spontaneous speech directed to their infant children and on a comparison with adult-directed speech and read speech. They use the corpus to demonstrate that a phonologically-informed analysis of infant-directed speech can reveal specific ways in which segmental and supra-segmental aspects of phonology are modulated dynamically to accommodate the specific communicative needs of speakers and hearers. **Zellou and Scarborough**, on the other hand, compare two corpora of English utterances, infant-directed in one and adult-directed in the other, and explore the extent to which age-of-acquisition and neighborhood density predict phonetic variability in the two sets of data.

The papers by **Den** and **Hasegawa-Johnson et al.** deal with issues of statistical modeling. Using a large-scale corpus of spontaneous Japanese, i.e., the

*Corresponding author: Haruo Kubozono, NINJAL, Tachikawa, Tokyo, E-mail: kubozono@ninjal.ac.jp
Kikuo Maekawa: E-mail: kikuo@ninjal.ac.jp, Timothy J. Vance: E-mail: tjvance@ninjal.ac.jp, NINJAL, Tachikawa, Tokyo

Corpus of Spontaneous Japanese (CSJ), the paper by **Den** analyzes phrase-final lengthening from a psycholinguistic point of view. Special attention is given to two cognitive factors, namely clause complexity and boundary depth, in addition to ordinary linguistic and phonological factors. The complex interactions among the two types of factors are modeled by means of general linear mixed-effect models. It turns out that the influences of cognitive factors are much more complex than those of linguistic/phonological factors.

The paper by **Hasegawa-Johnson et al.** discusses issues involved in corpus development using crowd-sourcing and mismatched annotation, and demonstrates the use of statistical learning to control for labeling error in phonological databases. The techniques discussed in this article are essential to the understanding of automatic corpus building and annotation and will be useful for readers who want to understand the technical aspects of corpus building and want to apply that knowledge to the construction of their own corpora. The article also includes appendices that review important concepts and results related to machine learning theory and conditional entropy.

The remaining three papers present quantitative analyses of spontaneous speech in individual languages. **Niebuhr and Hoekstra** investigate the intonation of North Frisian, based on a spontaneous interview corpus of a dialect spoken on the island of Föhr off the western coast of Germany. They report strong evidence for a phonological pitch-accent distinction relying on the difference between a pointed and a plateau-shaped F_O peak. The paper by **Kang et al.** uses tokens from a read-speech corpus and points out frequency effects on the ongoing vowel length contrast merger in Seoul Korean. **Stuart-Smith et al.** look at VOT in Scottish English based on tokens of voiced and voiceless stops extracted from a corpus of spontaneous Glaswegian that includes two sets of recordings made about 30 years apart. Their results corroborate earlier claims that VOT has been lengthening in this variety and also suggest that the shifts show a kind of gradience that indicates subtle sociolinguistic control.