Analysis of Language Variation
Using a Large-Scale Corpus of Spontaneous Speech*

Kikuo Maekawa

The National Institute for Japanese Language

Large-scale corpus of spontaneous speech can be a powerful tool for the study of language variation. Moreover, given that the corpus is publicly available, corpus-based analysis could open up the possibility of follow-up analysis in this area of linguistic study. Generally speaking, follow-up study is highly desirable in sciences but so far it has been virtually impossible in the area of socio-linguistics due to the lack of shared corpus. In this paper, I will present some results of the analyses of the Corpus of Spontaneous Japanese (CSJ) that we developed in the years 1999-2003.

CSJ is a large, richly annotated corpus of spontaneous speech of present-day Japanese (http://www2.kokken.go.jp/~csj/public/index.html), containing more than 660 hours of speech uttered by more than 1400 speakers. This corpus was designed primarily for statistical machine learning of acoustic- and language-models for automatic spontaneous speech recognition, but it was also designed for the study of language variation.

So far, we have analyzed variations at different levels of language structures including, vowel devoicing, pitch-accent location in adjectives, coalescence of particle succession, moraic nasalization of particles, diffusion of the new potential verb forms, choice of phrase-final boundary pitch movements (BPM), and strength of the prosodic boundary preceding accented particle. In addition to these, analysis of word-form variation was conducted. The last analysis was concerned not only with individual lexical items, but also with the lexicon as a whole.

Key words: spontaneous speech, corpus, variation, intonation, Japanese

*I am very grateful to my colleagues and former colleagues at the NIJL who participated in the CSJ project and analysis of language variation; Hanae Koiso, Hideki Ogura, Masaya Yamaguchi, Hideaki Kikuchi, Takayuki Kagomiya, Wataru Tsukahara, Kiyoko Yoneyama, Masako Fujimoto, Kenya Nishikawa, Yoko Mabuchi, Yohichi Maki, Kenji Yamazumi, Takehiko Maruyama and Yosuke Igarashi.