Building Corpus of Spoken Bahasa Indonesia for Phonetic and Phonological Research

Diana Stojanovic (stojanov@hawaii.edu)
Dept. of Linguistics-University of Hawai´i at Manoa

Abstract

Corpus of Spoken Language is invaluable resource in research on phonetic and phonological phenomena. While transcribed corpora of spoken language are easily analyzed, they contain limited amount of information. Thus having the raw materials (sound recordings) accessible is necessary in order to be able to search for the information not coded in the transcripts.

This research is a pilot project in building a Corpus of Spoken Bahasa Indonesia for the purpose of phonetic and phonological research. Specifically, the objectives were: (1) to design a small balanced expandable corpus; (2) to transcribe the pilot corpus for the chosen variables; and (3) to create tools for automatic generation of easily searchable text-based documents as well as analysis tools tailored to phonetic and phonological questions.

In the full version of the corpus, the following genres will be present: 1) controlled sentences, 2) read speech, and 3) naturally occurring speech (following experiences of building IViE corpus for English dialects http://www.phon.ox.ac.uk/~esther/ivyweb/). In the pilot version of the corpus (present status), materials consist of 4 interviews (naturally occurring speech) conducted with 2 female speakers (one speaker of Jakarta Indonesian, one considering herself representative of Javanese intolect of Indonesian). Praat platform is used for transcribing variables of interest in two forms: (1) using direct transcription into textgrids, and (2) using Praat scripts to extract the values of acoustic variables (pitch, intensity) and write them into acoustic-data textfiles. Several sets of variables are chosen to be coded in the corpus: (1) pitch, intensity, duration; (2) phonemes, syllables, prominent syllables, words, English gloss, and (3) pitch movements and break indices (loosely based on ToBI transcription system for English intonation).

So far, this pilot project has produced several outputs in the form of computer programs: (1) MakeWords, a program to convert format from textGrid to easily searchable *word* format that contain information about word duration (used for word frequency analysis and durational phenomena involving words); (2) MakeTones, a program to convert format from textGrid to easily searchable *tone* format containing information about tones and breaks (used for analysis of the intonational sequences); (3) MakeSyllables, a program to convert format from textGrid to format containing information about syllables (used for quantifying phenomena involving syllables); (4) MakeSegments, a program to convert format from textGrid to easily searchable format containing information about segments; and (5) WorkingCorpus, a program to build the corpus consisting of particular set of utterances.

This project was motivated by the interest in doing phonetic and prosodic research on Bahasa Indonesia. Some applications of the output include computing various segment statistics; formulating a framework for Indonesian for the purpose of studying intonational meanings and for typological comparisons; and examining cross-dialectal variation in spoken Indonesian.

Presentation will involve an overview and demonstration of the tools.