Building the "corpus production line" for speech research

Mietta Lennes

CLARIN ERIC aims at providing easy and sustainable access to digital language data and the tools for finding and using them, independent of where the materials are located (CLARIN = Common Language Resources and Technology Infrastructure, ERIC = European Research Infrastructure Consortium; http://www.clarin.eu/). The Language Bank of Finland participates in the CLARIN cooperation through the national FIN-CLARIN consortium.

The goal of FIN-CLARIN is to support the various stages in the researcher's workflow, the so-called "Corpus Production Line". I will provide a brief overview of the main issues in each stage. CLARIN already has some working solutions for legal issues as well as user authentication and authorization. META-SHARE (http://www.meta-share.org/) provides a distributed repository for creating and searching metadata. The Language Bank of Finland provides the LAT platform, developed by MPI in the Netherlands, for storing, searching and distributing speech and multimedia corpora. LAT includes a variety of useful tools and functionalities, although some problems remain unsolved.

Many valuable speech materials do not yet come with time-aligned annotation. Automatic speech recognition (ASR) can help in situations where audio files exist with no transcripts. However, the corpus may consist of audio files whose transcriptions are available in some format but have not been aligned with the sound files. The recordings may include several, partly overlapping speakers whose turns need to be separated in different annotation tiers. Moreover, the language variety may be "non-standard", for instance a dialect, and the recording may be old and/or of bad technical quality. Currently, these are problems for ASR systems. I will demonstrate a semi-automatic tool for Praat that can be used to roughly align a conversation transcript with the audio.