Publications Abstract

Campbell W., Campbell J., Torres-Carrasquillo, Reynolds, Yaeger-Dror Advantages of Large Dialect Corpora: Data Sources for Linguists and Engineer Linguistics Society of America San Francisco, CA 6-9 January 2005.

Abstract

Research area(s): sociolinguistics, speech analysis, corpus linguistics
Paper 'type': 15 minute paper

Corpora for sociolinguistic study are generally gathered from a demographic cross-section of speakers in a given community (Boberg 2004; Labov/Ash/Boberg online; Sankoff&Sankoff 1973). For most of these studies there is one dialect speaker being spoken to (or better, listened to) by an interviewer trained to create a casual interactive frame for their talk. Interviewers are trained to maximize the sound quality of the recordings in a consistent way, to create a focus on specific topics, to have similar interactions with all interviewees, and to speak as little as possible. Good interviewers vary their speech to accommodate to the apparent dialect characteristics of the interviewee (e.g., Trudgill 1986, Paradis 1996); nevertheless, it is generally acknowledged that 'group interviews' (where both speakers are already acquainted and speak the same dialect) are less hampered by the observer's paradox (Labov 1994), but there is a dearth of corpora which fulfill the need to fill the sociodemographic cells for such a study.

Independently, the speech-engineering community has discovered the need for a great deal of conversational speech in order to improve the recognition strategies of their systems (J.Campbell et al, 2004b, W.Campbell et al, 2004); since phone conversations obviate the need for interviewers, and minimize participants' focus on speech, large 'group interview' corpora have been collected and made generally available (J.Campbell et al, 2004a, Cieri et al 2002, 2004, Martin et al 2004).

Recent studies have used specific LDC corpora, like CallFriend, both to analyze dialect variables (Strassel 2002, Yaeger-Dror et al, 2004), and to train a system for dialect recognition (Torres et al 2004, Zissman et al 1996). The paper will discuss the advantages of CallFriend and CallHome for dialect studies, and answer the following questions: To what extent do these corpora provide a regionally balanced sample? What is the actual regional breakdown of the speech collected for the corpora? What are the limitations on these corpora, and how can they be avoided in future data-collections?

The paper will provide sufficient information to permit linguists to make best use of the corpora available, and to assist in improvement of the 'coverage' within a given corpus, to maximize the usefulness of these corpora. Hopefully, the paper will stimulate useful discussion between linguists and speech engineers.

Websites of interest for this paper
www.ling.upenn.edu/phono_atlas/
wave.ldc.upenn.edu/Catalog/project_index.jsp
www.ldc.upenn.edu/Mixer/
This work was sponsored by the United States Government Technical Support Working Group under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Government.