THE SLX CORPUS OF CLASSIC SOCIOLINGUISTIC INTERVIEWS: CORPUS OVERVIEW

Authors: Stephanie Strassel, Jeffrey Conn, Suzanne Evans Wagner, Christopher Cieri, William Labov, Kazuaki Maeda

Published by Linguistic Data Consortium - November 2003
[LDC2003T15]

http://www.ldc.upenn.edu/Projects/DASL/SLX

1 Background: The DASL Project

The project on Data and Annotations for SocioLinguistics (DASL) investigates best practices in the use of computer-based data and tools to support linguistic inquiry and documentation, with emphasis on the quantitative study of linguistic variation in society. Quantitative sociolinguistics is necessarily based upon empirical observation and statistical description of linguistic behavior. Collecting and annotating databases plays a crucial role. The current state of computing technology encourages the collection, annotation, analysis and even publication of results concerning linguistic behavior wholly within the digital domain. Digital data is easily shared, and that in turn encourages a whole range of positive practices.

The DASL Project hopes to encourage data sharing and the re-annotation and reuse of published data as an important complement to first-hand fieldwork. DASL first addressed these issues through a pilot study involving the analysis of a well-documented sociolinguistic variable (t/d deletion) as it appears in several large digital speech corpora. The publication of the SLX Corpus of Classic Sociolinguistic Interviews further reflects DASL's goals. It not only serves as an example of a digital speech corpus developed specifically to support sociolinguistic research, but also provides a stable benchmark for training in sociolinguistic data collection, digitization, segmentation, transcription, analysis and publication.

2 The SLX Corpus

2.1 Overview

The SLX Corpus of Classic Sociolinguistic Interviews comprises 8 sociolinguistic interviews with a total of 9 speakers, conducted in the 1960s and 70s. All of the interviews are conducted by William Labov or by one of his students. Labov notes that these interviews are not classic in the sense that they form part of a systematic sociolinguistic study of the speech community. Only the Rose B. interview is, drawn from the NYC study. And from the point of view of methodology -- what makes the other interviews good -- it is defective. It still bears the hallmarks of the dialectological format from which it was drawn. The others are all exploratory interviews without any basis in systematic sampling or design.

What makes these interviews classic is that they represent classic solutions to the problems of achieving cross-cultural contact, reducing the effect of the Observer's Paradox and approximating the vernacular of every-day life. Most importantly, they are interviews with extraordinarily gifted, memorable and fluent speakers. They contain a very rich store of narratives in response to the classic set of sociolinguistic interview questions. Several proved useful for the acoustic analysis of Labov, Yaeger & Steiner 1972 (Henry G., Kathy D., Jerry T.); and others provided materials for Labov's work on narrative analysis (Henry G., Bobbie A., Adolphus G., Joe D.).

These particular interviews have also been targeted for inclusion in this corpus because of the sound quality. All interviews were recorded on a Nagra III or IVS with Sennheiser dynamic microphones. Finally, these particular speakers have been selected because publication of the audio data and corresponding transcripts and annotations does not violate any agreement the interviewer made with the speakers regarding data distribution.

The data reflects a broad spectrum of speaking styles, including spontaneous speech, narratives, responses and formal linguistic tasks. The interviews touch on a multitude of topics, and corpus users should note that the language of the interviews represents the uncensored opinions of the speakers, reflecting their every day concerns and personal histories.

Taken as a whole, the speakers exemplify a wide variety of regional and social dialects. Demographic information for each main speaker in the corpus is displayed in the table below.

Speaker	Age	Interview Date	Interviewer	Speech Community	Ethnicity	Occupation	Education	Labov Notes
AdolphusH	81	1971	WL	Near Hillsboro, North Carolina	African American	Farmer	Very little	Adolphus H. is an older farmer living outside of Hillsboro, NC. The interview gives an excellent record of black speech of the period, which reflects local phonetics and phonology but includes features of African American vernacular as well.
BobbieA	22	1973	WL	Ayr, Scotland	Scottish/Italian	Saw doctor	Some technical college	Bobbie A. is a young man from the small town of Ayr in Scotland. The interview contains many effective narratives close to the Scots English vernacular.
HenryG	60	1970	WL	Dekalb Co., Georgia (E. Atlanta)	European American	Railroad foreman	High school graduate	Henry G. is an eloquent exponent of the Southern Shift. He is a retired railroad foreman who shows strong style shifting within the Southern Shift range, with many examples of the characteristic syntax of Southern States speech.
JerryT	19	1969	WL	Near Leakey, Texas	European American	Gas station attendent	Some high school	Jerry T. represents the most advanced form of the Texas variety of Southern States speech. This was recorded late at night at a filling station where there was very little background noise. It shows the most extreme form of schwa contraction of any weak word.
JoeD (interiewed with Eddie M)	21	1971	WL	Liverpool, England	English	Docker	Some high school	The interview with Joe D. and Eddie M. took place in a public park, and there is some noise (from birds and wind). But the quality of the interaction is very high, and the interview is a fairly accurate view of the Scouse dialect. Joe D. is a born ethnographer.
EddieM (interviewed with JoeD)	19	1971	WL	Liverpool, England	English	Docker	Some high school	
KathyD	15	1970	unknown	Rochester, New York	European American	Student	In eleventh grade	Kathy D. is an extreme example of the raising of short-a in the Northern Cities Shift. The interview includes a fluent and intimate account of her dating history.
LouiseA	53	1977	WL	Knoxville, Tennessee	European American	Mother	Unknown	Lousie A is a working class woman from Knoxville who speaks fluently and rapidly on a wide range of subjects. This is probably the best record of the vernacular of the Inland South, and a remarkably long interview with large volumes of speech.
RoseB	43	1963	WL	New York City, New York (Lower East Side)	Italian American	Seamstress in factory	Sixth grade	Rose B is an Italian-American working class woman who is an eloquent speaker and story teller. This is a paradigmatic example of the NYC vernacular.

Primary Speakers in the SLX Corpus

2.2 Processing: Digitization, Segmentation and Transcription

To create the current corpus, interview sessions were digitized from the original open reel tapes onto DAT/disk at 16bit, 44KHz sampling. The monaural signal was passed through 2 channels at levels differing by 20% to capture the best digital copy in a single pass. Technicians monitored the recording process, and adjusted for sustained changes in speech levels. The digital files show no significant clipping in the digital domain.

In the next stage, the audio files were processed to produce a single transcript file for each speaker. The target speaker was distinguished from the interviewer, other speakers, silence and noise. A time-aligned transcript was then created for each speaker. Audio segmentation proceeded in two passes. The first identified basic utterance boundaries. Sentence or phrase boundaries, significant pauses and breath groups were identified and marked in the second pass.

Once individual utterance-length segments had been created for each speaker, transcription began. The transcripts are written in standard orthography with normal punctuation and standard spelling. No attempt has been made to correct speakers' pronunciation or grammar, and problematic or incomprehensible sections are transcribed using special conventions. Unintelligible speech or doubtful best guesses, for example, are enclosed in ((double parentheses)). Non-standard lexical items are marked with special symbols. All transcripts have undergone at least two further passes: the first to catch errors and revisit unintelligible sections, and the second to allow a native speaker of British English to check for errors based on unfamiliarity with cultural items, slang or pronunciation.

The corpus documentation includes the complete set of annotation guidelines used in the segmentation and transcription of the SLX Corpus.

2.3 The Sociolinguistic Variable Survey

After the time-aligned transcripts had been created and checked for quality, researchers reviewed the data to identify sociolinguistic variables of interest. The interview subjects do not belong to a coherent speech community in space or time, except the speech community consisting of all English speakers. Thus certain phonological and grammatical variables in this survey were selected because they exist in all English dialects studied to date. They include among others the realization of -ing, t/d deletion and negative concord. Other variables in the survey are associated with certain of the dialects under consideration: habitual 'be' in African American English, for example, or the frication of syllable initial stops in Liverpool speech. Since some of these dialect features are categorically absent from non-speakers of the dialect, not all speakers were coded for all variables; nor were speakers coded exhaustively for any variable. In all, over 90 phonological, phonetic and prosodic and over 60 grammatical and lexical variables have been identified in the corpus. Tokens for each speaker are further annotated for speaking style. Researchers developed a style annotation guide consisting of seven basic categories ranging from casual speech to formal linguistic tasks. Each category is broken down into several sub-types. A complete description of the sociolinguistic variables identified in the survey and the full style annotation guide are included in the SLX Corpus documentation.

It is important to note in general that the survey is experimental, non-systematic and principally descriptive. It is not intended to give an exhaustive account of the variation that exists within the corpus. Instead, its purpose is to provide snapshot of intra- and inter-speaker variation in the corpus, and to highlight a broad range of sociolinguistic variables that are attested in the data. It further gives us an idea of how the DASL Project could eventually allow sociolinguists to examine a general variable like t/d deletion across many speech communities.

2.4 Value of the SLX Corpus

The SLX corpus is an example of a multi-community sociolinguistic corpus. Containing classic interview material in the Labovian tradition, it is a valuable teaching tool for linguists. The recordings demonstrate successful interviewing techniques, the sound quality is high, and the digitization, segmentation and transcription of the data represent best practice in these areas. The variable survey highlights over 150 sociolinguistic variables attested in the corpus and suggests avenues for further research. Most importantly, the SLX corpus represents an important step in creating infrastructure, tools and data to support a collaborative digital methodology for the quantitative study of linguistic variation in society.

3 Using the Corpus

3.1 SLX Corpus Tools

This publication includes a number of annotation tools that were developed to support browsing and exploration of the SLX Corpus, as well as those used in the creation of the corpus. The current release includes a pre-compiled package of these tools for Windows users. We plan to create a pre-compiled package for i386 Linux users in the future.

First, the corpus release includes Transcriber, a multi-platform transcription tool developed jointly by LDC and DGA. LDC staff used this tool to create the initial audio segmentation and transcription for the SLX corpus. Transcriber was selected for these tasks a number of reasons. First, the audio segmentation process is facilitated by Transcriber's "real time" mechanism, which

allows annotators to label natural breakpoints in the audio by just hitting <Enter> as they listen to interview. This means that initial segmentation can proceed at close to real time. Second, because Transcriber is freely-available and works across different computing platforms, accepting many different audio and transcript formats as input, it allows the segmentations and transcriptions to be shared with other linguists and even allowing multi-site annotation.

The corpus also contains a version of MultiTrans, recently developed at LDC as an alternative to Transcriber, expanding on many of Transcriber's most useful features. The original segmentation and transcription process for SLX created individual transcript files for each speaker in an interview. MultiTrans allows corpus users to explore these individual speaker files, but also provides a merged view of these same transcripts that is time-aligned with the audio, making browsing the transcripts while listening to the audio very straightforward.

The DASLTrans tool, developed specifically for the SLX Corpus, provides an interactive view of the sociolinguistic variable survey while allowing users to add new annotations or modify existing ones. The tool gives users a spreadsheet-style view of the tokens that have been coded for each speaker, along with all corresponding annotations. Each token is displayed within its surrounding utterance, and each utterance is linked to the corresponding audio signal for easy playback. The variable survey forms are also provided in their original Microsoft Excel format.

To further facilitate exploration of the SLX Corpus, this release also includes the SLX Corpus Browser, a Windows-based interactive assistant that steps users through the corpus, allowing them to view corpus documentation, complete transcripts plus audio, or the sociolinguistic variable survey for any speaker with just a few mouse clicks. The corpus documentation directory contains user manuals for the SLX Corpus tools.

The versions of MultiTrans and DASLTrans included within this publication have been optimized for exploring the SLX Corpus. Future tool releases will integrate the segmentation and transcription features of MultiTrans and the annotation features of DASLTrans into a single tool, allowing users to go from segmentation to transcription to coding of variables within a single system.

Both DASLTrans and MultiTrans were developed using the Annotation Graph Toolkit (AGTK - http://agtk.sf.net), also developed by LDC. Future versions of DASLTrans and MultiTrans will be available from the AGTK website as well as the SLX Corpus website: http://www.ldc.upenn.edu/Projects/DASL/SLX

Updates and additional information about Transcriber can be found at: http://www.etca.fr/CTA/gip/Projets/Transcriber/

3.2 File Naming Conventions

Each interview contains the speech of several speakers; at minimum, this includes the interviewer and the respondent. Because separate transcript files were created for each speaker within the interview, multiple transcript files correspond to each audio file. There is also a "merged" transcript file for each interview that shows the complete transcript for all speakers in one file. Each individual speaker also has a variable survey file containing dozens of examples of sociolinguistic variables attested for the speaker. The files follow a consistent naming convention.

In each case, the first element of the file name is the name of the target speaker(s). This is followed by a number to indicate the tape volume (01, 02 or 03). For the files that correspond to individual speaker transcripts, this is followed by the name of the transcribed speaker in ALL CAPS. Finally, the file extension represents the type of file.

For example, for Tape 01 of Labov's interview with AdolphusH, the corresponding files are:

Tool (file extension)	**All tools (.wav)**	**DASLTrans (.tsv)**	**MS Excel (.xls)**	**MultiTrans (.lcf)**	**Transcriber (.trs)**
File Description	Audio File	Socioling. Variable Survey File	Socioling. Variable Survey File	Transcript Files	Transcript Files
Example	*AdolphusH_01.wav*	*AdolphusH_01.tsv*	*AdolphusH_01.xls*	*AdolphusH_01_MERGED.lcf* *AdolphusH_01_ADOLPHUS.lcf* *AdolphusH_01_LABOV.lcf* *AdolphusH_01_WIFE.lcf* *AdolphusH_01_FRIEND.lcf*	*AdolphusH_01_ADOLPHUS.trs* *AdolphusH_01_LABOV.trs* *AdolphusH_01_WIFE.trs* *AdolphusH_01_FRIEND.trs*

4 SUPPORT

The SLX Corpus was funded in part through a 5-year grant (BCS-998009, KDI, SBE) from the National Science Foundation via TalkBank (http://www.talkbank.org/), an interdisciplinary project to foster research and development in communicative behavior by providing tools and standards for analysis and distribution of language data. Additional funding was provided by Linguistic Data Consortium.