NCSLAAP

NORTH CAROLINA SOCIOLINGUISTIC ARCHIVE AND ANALYSIS PROJECT

➡http://ncslaap.lib.ncsu.edu

tyler kendall | tsk3@duke.edu | duke university/nc state university

nwav35 poster | the ohio state university | november 10, 2006

about the project ■

The North Carolina Sociolinguistic Archive and Analysis Project (NCSLAAP) is a research and preservation initiative being conducted jointly between the North Carolina Language and Life Project (NCLLP) and the North Carolina State University Libraries.

ncslaap's core goals:

- ■ TO PRESERVE THE NCLLP'S RECORDINGS THROUGH DIGITIZATION
- ■ TO ENABLE AND EXPLORE NEW COMPUTER-ENHANCED TECHNIQUES FOR SOCIOLINGUISTIC ANALYSIS

about the ncllp■

The NCLLP is a sociolinguistic research initiative at North Carolina State University with one of the world's largest collections of sociolinguistic interviews focusing on Southern American English. It consists of approximately 1,500 interviews conducted from the late 1960s up to the present, most on analog cassette tape, but some in formats ranging from reel-to-reel tape to digital video. The collection features the interviews of Walt Wolfram, Erik Thomas, Natalie Schilling-Estes, Kirk Hazen, and numerous other scholars. For more information about the NCLLP see http://www.ncsu.edu/linguistics/.

archives and digitization ■

Sociolinguists have developed a range of excellent techniques for acquiring naturalistic speech data through the recording of *sociolinguistic interviews*. However, with few exceptions (e.g. Poplack 1989), they have not focused a great deal on the storage and preservation of their data and future access to them. NCSLAAP seeks to preserve and make more accessible the large collection of sociolinguistic interviews of the NCLLP.

The digitization of this archive is major task in its own right with important dividends. The centralization of the digitization process ensures a consistent method as opposed to an *ad hoc* approach to digitization that scholars are often forced to follow, digitizing particular tapes (or parts of tapes) as needed for specific projects. NCSLAAP's web interface means that analysts can have instant access to their data from anywhere.

re-examining transcription■

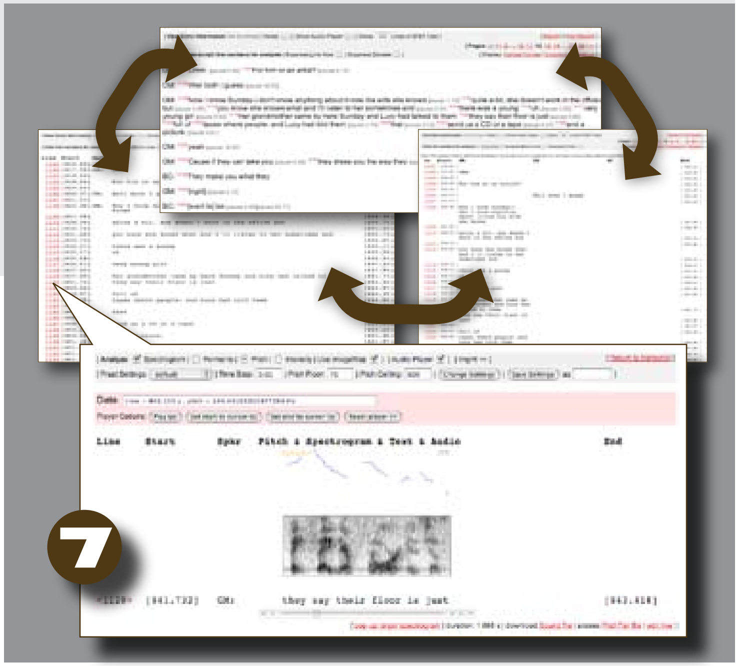
Unlike the *textual accuracy* that many transcript theorists aim for (cf. Du Bois et al. 1993), NCSLAAP transcripts target *temporal accuracy* with the assumption that everything else can be constructed from the audio file, either automatically by the software, or manually by examining the audio for the given time range. With the start- and end-times for each utterance captured in the database and a linkage maintained with the audio much of the other information that is often tagged or coded (e.g., latching, overlap, pause length, etc.) is unnecessary. Additionally, storing the transcript data separate from its formatting allows for a wide variety of different presentations, all viewable at the click of a mouse, as illustrated in ❼.

Core data elements for a data-based transcript

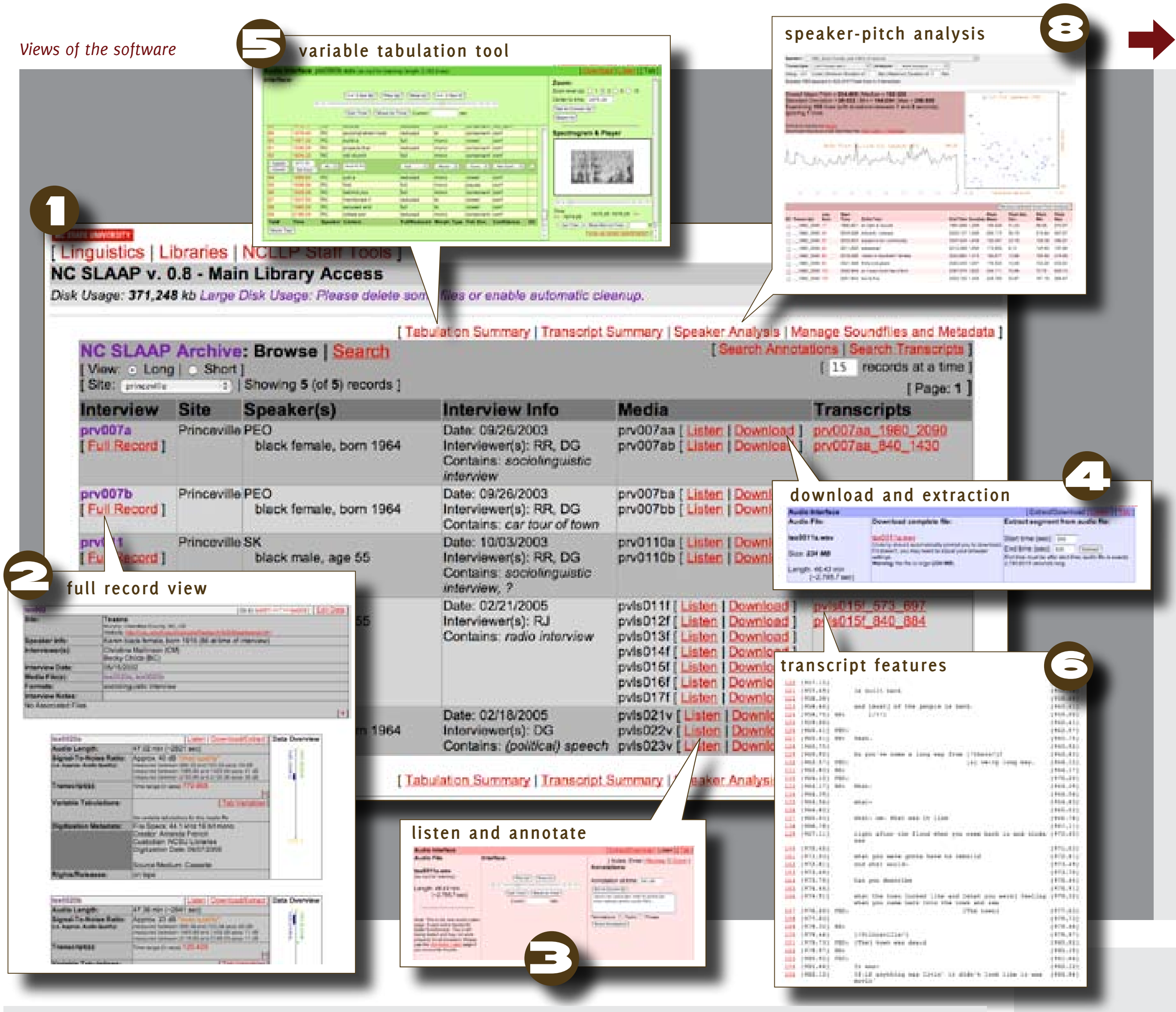
SPEAKER | UTTERANCE START TIME | UTTERANCE TEXTUAL REPRESENTATION | UTTERANCE END TIME

In a data-based transcript model, the only data strictly required for an utterance are the speaker, a textual representation, and the start- and end-times. This very simple data model is actually quite powerful. Software, like NCSLAAP, can then create links between the transcript data and the source audio file, and can conduct real-time phonetic analysis. In other words, there is no need to code for features such as loudness or pitch because these features can be reconstructed from the audio itself.

Transcript views


Line
Start
Spkr
Pitch & Spectrogram & Text & Audio
End
<1128>
[841.732]
GM:
they say their floor is just
[843.418]
7

Views of the software



[Linguistics | Libraries | NCLLP Staff Tools]

NC SLAAP v. 0.8 - Main Library Access

Disk Usage: 371,248 kb Large Disk Usage: Please delete son files or enable automatic cleanup.

[Tabulation Summary | Transcript Summary | Speaker Analysis | Manage Soundfiles and Metadata]

NC SLAAP Archive: Browse | Search [Search Annotations | Search Transcripts]

[View: ○ Long | ○ Short] [15 records at a time]

[Site: princeville] | Showing 5 (of 5) records] [Page: 1]

Interview	Site	Speaker(s)	Interview Info	Media	Transcripts
prv007a [Full Record]	Princeville	PEO black female, born 1964	Date: 09/26/2003 Interviewer(s): RR, DG Contains: *sociolinguistic interview*	prv007aa [Listen | Download] prv007ab [Listen | Download]	prv007aa_1960_2090 prv007aa_840_1430
prv007b [Full Record]	Princeville	PEO black female, born 1964	Date: 09/26/2003 Interviewer(s): RR, DG Contains: *car tour of town*	prv007ba [Listen | Downl prv007bb [Listen | Downl	
prv 11 [Fu Record]	Princeville	SK black male, age 55	Date: 10/03/2003 Interviewer(s): RR, DG Contains: *sociolinguistic interview, ?*	prv0110a [Listen | Downl prv0110b [Listen | Downl	
		55	Date: 02/21/2005 Interviewer(s): RJ Contains: *radio interview*	pvls011f [Listen | Download] pvls012f [Listen | Download] pvls013f [Listen | Download] pvls014f [Listen | Download pvls015f [Listen | Downlo pvls016f [Listen | Downlo pvls017f [Listen | Downlo	pvls015f_573_697 pvls015f_840_884
		n 1964	Date: 02/18/2005 Interviewer(s): DG Contains: *(political) speech*	pvls021v [Listen | Downlo pvls022v [Listen | Downlo pvls023v [Listen | Downlo	

[Tabulation Summary | Transcript Summary | eaker Analys

➡ http://ncslaap.lib.ncsu.edu

software features ▪

While one of the major benefits of the project for scholars in general is the creation of a large online archive of sociolinguistic interviews, NCSLAAP also provides new tools and interfaces for interacting with and analyzing the corpus.

Basic features include: ❶ & ❷ a browsable and searchable interface to the archive collection, ❸ an audio player with an annotation tool that allows users to associate searchable notes to specific times within the audio files (and to listen to those particular passages at the click of the mouse), and ❹ an audio extraction feature that enables users to download excerpts of audio files without having to download or locally store the large files.

Analytic features include: ❺ tools that aid in the extraction and tabulation of linguistic variables, phonetic analysis features, and ❻ sophisticated transcript options. Transcript data are linked to the audio files and transcripts can viewed in a number of formats at the same time as one listens to the associated audio. A version of Praat, the open-source phonetic analysis software, is integrated with the software to allow for instantaneous retrieval of phonetic data (such as pitch or intensity readings) as well as ❼ the generation of spectrograms in line with the transcript text.

Finally, corpus-like tools are in development that allow for large-scale linguistic analysis across interviews, speakers, and research sites, such as a pitch analysis feature, as shown in ❽ .

future directions ▪

At present, both the archive and software are under development. New features – such as support for multilingual transcripts and new corpus-like analysis tools – are scheduled for development. It is hoped that the over the course of the next year or so the entire collection of the NCLLP's interviews will be digitized and included in the archive and much of the software features will be completed.

Meanwhile, the eventual goal is to make the NCSLAAP software available to the greater sociolinguistic community – either via a more widely accessible web server or through the distribution of the source-code – so that other researchers can make use of the software to store and interact with their own archives.

selected references ▪

▪ Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino (1993), Outline of Discourse Transcription. In Edwards, Jane and Martin Lampert, eds., *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale, NJ: Lawrence Erlbaum: 45-89.

▪ Poplack, Shana (1989), The Care and Handling of a Mega-Corpus: The Ottawa-Hull French Project. In Fasold, Ralph and Deborah Schiffrin, eds., *Language Change and Variation*, Amsterdam: Benjamins: 411-451.