Digital Audio Archives, Computer-Enhanced Transcripts, and New Methods in Sociolinguistic Analysis

Tyler KENDALL
Duke University
North Carolina State University
tsk3@duke.edu

Amanda FRENCH
North Carolina State University
amanda_french@ncsu.edu

Introduction

Traditional methods in sociolinguistic analysis have often relied on the repeated close listening of a set of audio recordings counting the number of times particular linguistic variants occur in lieu of other variants (a classic sociolinguistic example is the tabulating of words using final *–in'* for final *–ing*; cf. Fischer 1958, Trudgill 1974, etc.). These tabulations are normally recorded into a spreadsheet using a program such as Microsoft Excel, or even just into a hard-copy tabulation sheet. The results are then presented as summaries in publications or conference papers as the "data" used for description, explanation, and theory building. Some approaches in linguistics, such as discourse analysis, rely heavily on the development of transcripts of the audio recordings and often the focus of analysis is on the transcript itself and not the original recording or interview event. However, scholars following a wide variety of sociolinguistic approaches have repeatedly highlighted the confounds that arise from these treatments of "pseudo-data" (i.e., analysts' representations of the data) as data. Linguists such as Blake (1997) and Wolfram (e.g., 1993) have discussed problems relating to the tabulation and treatment of linguistic variables and raised the issue that individual scholars' methods are often not comparable. In discussing transcription theory, Edwards has repeatedly pointed out that "transcripts are not unbiased representations of the data" (Edwards 2001: 321). In general, the understanding that linguistic data is more elusive than traditional "hard science" data is widespread but not acted upon. In this paper, we present a project underway at North Carolina State University to argue that computer-enhanced approaches can propel sociolinguistic methodology into a new, more rigorous era.

The North Carolina Sociolinguistic Archive and Analysis Project

The North Carolina Language and Life Project (NCLLP) is a sociolinguistic research initiative at North Carolina State University (NCSU) with one of the largest audio collections of sociolinguistic data on American English in the world. It consists of approximately 1,500 interviews 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, Natalie Schilling-Estes, Erik Thomas, and numerous other scholars. The NCLLP has partnered with the NCSU Libraries on an initiative titled the North Carolina Sociolinguistic Archive and Analysis Project (NC SLAAP). NC SLAAP has two core goals: (1) to preserve the NCLLP's recordings through digitization; and (2) to enable and explore new computer-enhanced techniques for interacting with the collection and for conducting sociolinguistic analysis.

NCSU Libraries has as one of its chief goals the long-term preservation of the recordings made by the NCLLP, and it regards digitization as an appropriate means of preservation. Academic libraries may still be less expert than some commercial organizations when it comes to digitizing and storing audio, but they may be even less equipped to maintain analog audio collections properly (cf. Brylawski 2002, Smith, Allen, and Allen 2004). Archivists and librarians also sometimes point out that digitization and storage of audio may not be worth the expense and difficulty if the sole goal is preservation (cf. Puglia 2003). However, when scholarly digital projects can contribute significantly to the advancement of a discipline, as in the case of NC SLAAP, surely significant investments are called for.

The NC SLAAP project has from the beginning planned to integrate sociolinguistic analysis tools into the archive. This has been achieved to a large degree by integrating

the open source phonetic software application Praat (http://www.praat.org) into the web server software. In brief overview, the NC SLAAP system is an Apache web server currently housed on a Macintosh G5 computer running Mac OS 10.4. Data are stored in a MySQL database and application pages are written in PHP. The web server communicates with third-party open source applications to do most of its "heavy" processing. Most importantly, the web server communicates with Praat to generate real-time phonetic data (such as the pitch data and the spectrogram illustrated in Figure 1).

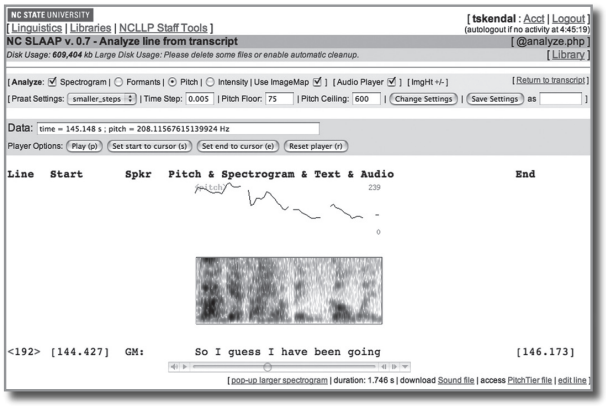

Figure 1: Transcript Line Analysis Example

While certain feature sets are still under development, NC SLAAP, even in its current state, provides a range of tools that greatly enhance the usability of the audio data. These features include an audio player with an annotation tool that allows users to associate notes with particular timestamps, an audio extraction feature that allows users to download and analyze particular segments of audio files, sophisticated transcript display options (as partly illustrated in Figure 1, above), and extensive search and query tools. Importantly, the NC SLAAP software helps to address concerns about the treatment of "pseudo-data" as data, because it enables scholars to better access, check, and re-check their (and their colleagues') variable tabulations, analyses, and conclusions. In short, the NC SLAAP software is an attempt to move us one step – hopefully, a large step – closer to the "real" data.

The features of the NC SLAAP software have potentially tremendous implications for a wide range of linguistic approaches. We focus on only one such feature here: the implications relating to transcription theory.

Transcription Method and Theory

Improvements to the traditional text transcript are extremely important because the transcript is often the chief mediating apparatus between theory and data in language research. Language researchers have long been concerned with the best method and format for transcribing natural speech data (cf. Edwards 2001). Researchers frequently incorporate a number of different transcription conventions depending on their specific research aims. Discourse analysts (e.g., Ochs 1979) traditionally focus most heavily on transcription as theory and practice, but researchers studying language contact phenomena (as in Auer 1998) also have their own transcription conventions for analyzing and presenting their data. At the other end of the spectrum are variationists and dialectologists, who also use transcripts, even if often only for presentation and illustration.

Despite the importance of the transcript for most areas of linguistics, little work has been done to enhance the usability and flexibility of our transcripts. Yet the way a researcher builds a transcript has drastic effects on what can be learned from it (Edwards 2001). Concerns begin with the most basic decision about a transcript: how to lay out the text. Further decisions must be made throughout the transcript-building process, such as decisions about how much non-verbal information to include and how to encode minutiae such as pause-length and utterance overlap. Furthermore, the creation of a transcript is a time- and energy-intensive task, and researchers commonly discover that they must rework their transcripts in mid-project to clarify aspects of the discourse or speech sample.

The NC SLAAP software seeks to improve the linguistic transcript by moving it closer to the actual speech that it ideally represents (Kendall 2005). In the NC SLAAP system, transcript text is treated as annotations on the audio data: transcripts are broken down into utterance-units that are stored in the database and directly tied to the audio file through timestamping of utterance start and end times. Transcript information can be viewed in formats mimicking those of traditional paper transcripts, but can also be displayed in a variety of dynamic ways – from the column-based format discussed by Ochs (1979) to a finer-level focus on an individual utterance complete with phonetic information (as shown in Figure 1, above).

Conclusion

NC SLAAP is a test case for new ways of approaching linguistic analysis, using computers to maintain a strong tie between the core audio data and the analysts' representations of it. In many senses the project is still in a "proof of concept" stage. However, we feel that it has made large steps towards new and more rigorous methods for sociolinguistic analysis and data management. In addition, it can serve as a model for academic libraries as a project that incorporates digital preservation with significant scholarly advancement.

References

Auer, P. (ed.) (1998). *Code-Switching in Conversation*, London: Routledge.

Blake, R. (1997). Defining the Envelope of Linguistic Variation: The Case of "Don't Count" Forms in the Copula Analysis of AAVE. *Language Variation and Change* 9: 57-79.

Brylawski, S. (2002). Preservation of Digitally Recorded Sound, in *Building a National Strategy for Preservation: Issues in Digital Media Archiving*, Council on Library and Information Resources Publication 106.http://www.clir.org/pubs/abstract/pub106abst.html

Edwards, J. (2001). The Transcription of Discourse, in *Handbook of Discourse Analysis*, eds. Deborah Tannen, Deborah Schiffrin, and Heidi Hamilton: 321-348, Oxford and Malden, Ma: Blackwell.

Fischer, J. (1958). Social Influences on the Choice of a Linguistic Variant. *Word* 14: 47-56.

Kendall, T. (2005). Advancing the Utility of the Transcript: A Computer-Enhanced Methodology, paper presented at the Twelfth International Conference on Methods in Dialectology: Moncton, New Brunswick, Canada. August 2005.

Puglia, S. (2003). Overview: Analog vs. Digital for Preservation Reformatting, paper presented at the 18th Annual Preservation Conference, March 27, 2003, at University of Maryland College Park. http://www.archives.gov/preservation/conferences/papers-2003/puglia.html

Smith, A., D. Allen, and K. Allen (2004). *Survey of the State of Audio Collections in Academic Libraries*. Council on Library and Information Resources Publication 128. http://www.clir.org/pubs/abstract/pub128abst.html

Ochs, E. (1979). Transcription as Theory, in *Developmental Pragmatics*, eds. Elinor Ochs and Bambi Schieffelin: 43-72, New York: Academic Press.

Trudgill, P. (1974). *The Social Differentiation of English in Norwich*, Cambridge: CUP.

Wolfram, W. (1993). Identifying and Interpreting Variables. *American Dialect Research*. Ed. Dennis R. Preston. Amsterdam: John Benjamins Publishing Company. 193-221.