University of Pennsylvania Working Papers in Linguistics

Volume 13 2007
Issue 2 *Selected Papers from NWAV 35*

Article 2

10-1-2007

Enhancing sociolinguistic Data Collections: The North Carolina Sociolinguistic Archive and Analysis Project

Tyler Kendall
Duke University

Enhancing Sociolinguistic Data Collections: The North Carolina Sociolinguistic Archive and Analysis Project[*]

Tyler Kendall[**]

1 Introduction

Recordings of natural speech play a central role in the diverse subdisciplines of linguistics. The reliance on naturalistic data is especially profound in sociolinguistics, and, as a result, sociolinguists have developed and deployed a range of techniques for acquiring such data, such as the *sociolinguistic interview* (cf. Labov 1984). However, with few exceptions (e.g. Poplack 1989), sociolinguists have not focused a great deal on the storage and preservation of their data or on ensuring future access to it. As a consequence, sociolinguists are often not particularly good at preserving and managing their often large collections of data. Furthermore, a (potentially unnecessarily) large portion of the sociolinguistic enterprise is spent on data collection and (re-)analysis since existing data collections are frequently not well-organized or accessible for future work.

This paper introduces the North Carolina Sociolinguistic Archive and Analysis Project (NC SLAAP),[1] an exploration of new approaches to storing, managing, and interacting with natural speech data. The project centers on the creation of an online archive and analytic toolset for the sociolinguistic data collection of the North Carolina Language and Life Project. The primary goals behind NC SLAAP are twofold; at the practical level, it seeks to provide researchers with better access to and interfaces for their data, and at the theoretical level, NC SLAAP seeks to question and rethink current linguistic and sociolinguistic conceptions of the nature of speech data, its representations, and the sorts of questions that can be asked of it.

[*] This paper was first presented at NWAV 35 as a poster under the title "The North Carolina Sociolinguistic Archive and Analysis Project: Empowering the Sociolinguistic Archive."

[**] I would like to gratefully acknowledge funding and support from the North Carolina Language and Life Project, The North Carolina State University Libraries, the William C. Friday Endowment at North Carolina State University, and the Duke University Graduate School. I would also like to thank the many people who have helped to make NC SLAAP possible, in particular Charlotte Vaughn for her extensive help on this paper and its earlier poster-incarnation.

[1] The NC SLAAP website is located at *http://ncslaap.lib.ncsu.edu/*.

2 The Archive

2.1 The North Carolina Language and Life Project

The North Carolina Language and Life Project (NCLLP) is a sociolinguistic research initiative at North Carolina State University with one of the largest audio collections of sociolinguistic data on Southern American English in the world. 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 continues to grow with the addition of upwards of one hundred new interviews a year.[2]

The NCLLP's large and growing collection of interviews is an important resource for linguists in general and for other scholars interested in the American South.[3] As a part of the NC SLAAP initiative, all of these sociolinguistic interviews are being digitized.

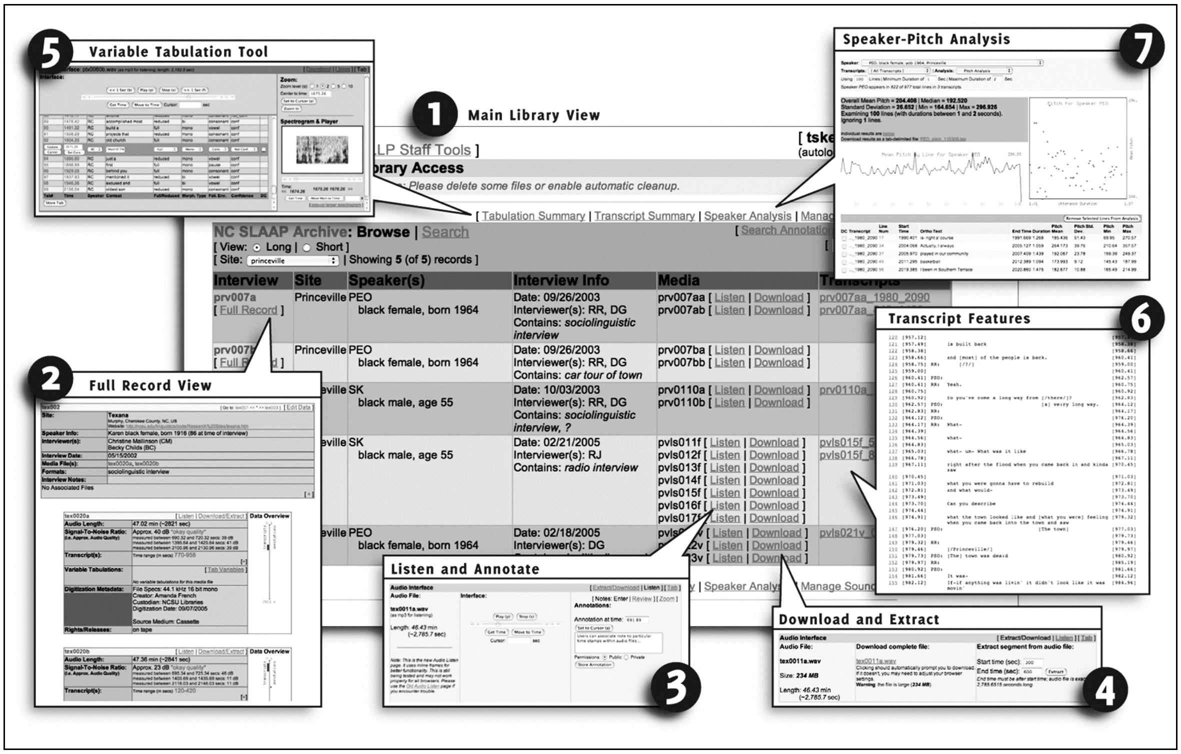
2.2 Features of the Archive and Software

With the goal to preserve it and make it more accessible to researchers, the digitization of the NCLLP collection is clearly a beneficial and important task in its own right. However, NC SLAAP makes this archive even more useful and accessible to its users by providing new tools and interfaces for interacting with and analyzing the corpus. A collage of screenshots is presented in Figure 1 showing a number of the software's features.

Basic features include: (1) & (2), a browsable and searchable interface to the archive collection, (3) 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 (4) an audio extraction feature that enables users to download excerpts of audio files without having to download or locally store the large files.

[2] More information about the NCLLP is available on the project's website at *http://www.ncsu.edu/linguistics/ncllp/*.

[3] Of course, due to the nature of data derived from human subjects, most of the data in the archive are not publicly available. Built into the NC SLAAP software are strong controls over who can access the resources. Nonetheless, outside scholars with appropriate research interests can request, and gain, access to the archive. While these decisions are ultimately up to the principal investigator(s) of a given research site or project, NC SLAAP's interface to the archive makes this whole process simpler.


1 Main Library View
2 Full Record View
3 Listen and Annotate
4 Download and Extract
5 Variable Tabulation Tool
6 Transcript Features
7 Speaker-Pitch Analysis
LP Staff Tools]
rary Access
Please delete some files or enable automatic cleanup.
[Tabulation Summary | Transcript Summary | Speaker Analysis | Manag
NC SLAAP Archive: Browse | Search
[View: ● Long | ○ Short]
[Site: princeville] | Showing 5 (of 5) records]

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_1980_2090 prv007aa_
prv007b [Full Record]	Princeville	PEO black female, born 1964	Date: 09/26/2003 Interviewer(s): RR, DG Contains: *car tour of town*	prv007ba [Listen | Download] prv007bb [Listen | Download]	
	eville	SK black male, age 55	Date: 10/03/2003 Interviewer(s): RR, DG Contains: *sociolinguistic interview, ?*	prv0110a [Listen | Download] prv0110b [Listen | Download]	prv0110a_
	eville	SK black male, age 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 | Download] pvls016f [Listen | Download] pvls017f [Listen | Download]	pvls015f_5 pvls015f_8
	eville	PEO black female, born 1964	Date: 02/18/2005 Interviewer(s): DG	pvls021v [Listen | Download] pvls022v [Listen | Download] pvls023v [Listen | Download]	pvls021v_

Figure 1: Assorted screenshots from NC SLAAP, discussed in the text

Analytic features include: (5) tools that aid in the extraction and tabulation of linguistic variables (a close-up is provided in Figure 4, below), phonetic analysis features, and (6) sophisticated transcript options. Transcript data are linked to the audio files and transcripts can be viewed in a number of formats at the same time as one listens to the associated audio (see also Figure 2, below). A version of Praat, the open-source phonetics software,[4] is integrated into the NC SLAAP software to allow for the instantaneous retrieval of phonetic data (such as pitch or intensity readings) as well as the generation of spectrograms in-line with the transcript text (see Figure 3, below). Finally, corpus-like tools are in development that will allow for large-scale linguistic analysis across interviews, speakers, and research sites, such as a pitch analysis feature (7), and a pause analysis feature (used and described, for example, in Kendall 2007).

3 Re-examining Transcription

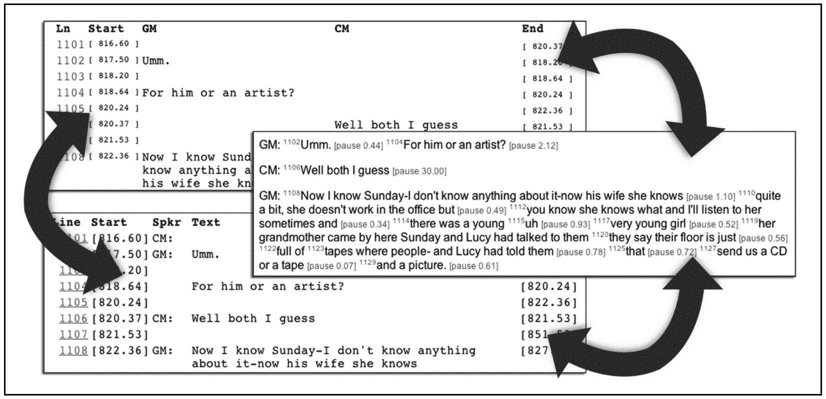
The transcript is, without doubt, the primary representation used to present speech in a non-aural format. Within language research, it is often the chief mediating apparatus between theory and data. As such, NC SLAAP seeks to make a large contribution to our thinking about and use of transcripts.

Researchers from a wide array of linguistic disciplines and across the social sciences rely on transcripts for the analysis and presentation of their data, yet despite some important interventions (e.g. Ochs 1979, Edwards and Lampert 1993, Edwards 2001) most transcripts remain text-based documents, varying in their conventions from researcher to researcher, and limited in their utility to the project at hand. While we know, as Jane Edwards wrote, that “transcripts are invaluable [since] they provide a distillation of the fleeting events of an interaction, frozen in time, freed from extraneous detail, and expressed in categories of interest to the researcher” (Edwards 2001:321), we also know that the form of and information in a given transcript will influence our interpretations of the data (Ochs 1979, Edwards 2001). Decisions as seemingly straightforward as how to lay out the text, to those more nuanced—like how much non-verbal information to include and how to encode minutiae such as pause length and utterance overlap—have far-reaching effects on the utility of a transcript and the directions that the transcript may lead the analysts.

NC SLAAP adopts the hypothesis that speech data can be treated and stored as data, just as one might treat and store other types of data (such as financial or customer information, to use business comparisons). Along

[4] Information about Praat is available at *http://www.praat.org/*.

those lines, NC SLAAP seeks to apply standard data management and presentation methodologies to the treatment and representation of natural speech data. One major premise therein is the separation of content and format. Separating the data from its formatting provides a huge amount of flexibility in terms of the presentation of the information. One direct result of this is that transcripts can be presented in any number of formats. For example, Figure 2 displays three different views of the same transcript data as currently available in the NC SLAAP software.



```
Ln   Start    GM                                  CM                  End
1101 [ 816.60 ]                                                       [ 820.37
1102 [ 817.50 ] Umm.                                                  [ 818.2
1103 [ 818.20 ]                                                       [ 818.64 ]
1104 [ 818.64 ] For him or an artist?                                 [ 820.24 ]
1105 [ 820.24 ]                                                       [ 822.36 ]
     [ 820.37 ]                           Well both I guess           [ 821.53 ]
     [ 821.53 ]
108  [ 822.36 ] Now I know Sund
                know anything a
                his wife she kn
```

GM: ^{1102}Umm. [pause 0.44] ^{1104}For him or an artist? [pause 2.12]

CM: ^{1106}Well both I guess [pause 30.00]

GM: ^{1108}Now I know Sunday-I don't know anything about it-now his wife she knows [pause 1.10] ^{1110}quite a bit, she doesn't work in the office but [pause 0.49] ^{1112}you know she knows what and I'll listen to her sometimes and [pause 0.34] ^{1114}there was a young ^{1115}uh [pause 0.93] ^{1117}very young girl [pause 0.52] ^{1119}her grandmother came by here Sunday and Lucy had talked to them ^{1120}they say their floor is just [pause 0.56] ^{1122}full of ^{1123}tapes where people- and Lucy had told them [pause 0.78] ^{1125}that [pause 0.72] ^{1127}send us a CD or a tape [pause 0.07] ^{1129}and a picture. [pause 0.61]

```
ine  Start    Spkr Text                                        
 01  [816.60] CM:
     7.50]    GM:  Umm.
110  .20]
1104 818.64]       For him or an artist?                      [820.24]
1105 [820.24]                                                 [822.36]
1106 [820.37] CM:  Well both I guess                          [821.53]
1107 [821.53]                                                 [851
1108 [822.36] GM:  Now I know Sunday-I don't know anything    [827
                   about it-now his wife she knows
```

Figure 2: Three presentations of the same transcript data

Transcript data in NC SLAAP are stored in database tables. Each transcript is a table in the database, and each line is an entry in the database table representing an utterance by a speaker.[5] Transcripts for NC SLAAP are built using Praat to obtain highly accurate start- and end-times for each utterance. Unlike the textual accuracy that many transcript theorists aim for,[6] NC SLAAP transcripts target temporal accuracy with the belief that everything

[5] The determination of exactly what should constitute a transcript line (and, more broadly, how we define *utterance*) is not a straightforward question. For NC SLAAP, a line is based simply on an utterance as defined as a single *phonetic unit*, an unbroken stretch of speech (silence-speech-silence). Other scholars (e.g. Chafe 1993) focus on *intonation units* as the principal spoken unit. While it is outside the scope of this paper to pursue this further, Figueroa (1994) provides a valuable discussion about some major definitions and treatments of utterance.

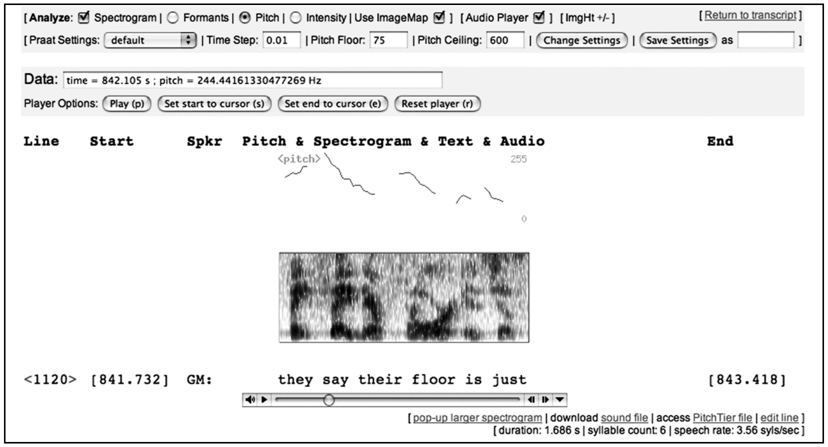
[6] See, for example, Du Bois, Schuetze-Coburn, Cumming, and Paolino (1993) for a comprehensive outline of potential transcription conventions and features that a textually accurate transcript may seek to encompass.

else can be (re-)constructed from the audio file, either automatically by software, or manually by examining the audio for the given time range.

Speaker Reference	Utterance Start Time	Utterance Textual Representation	Utterance End Time

Table 1: Core data elements for a data-based transcript

In a data-based transcript model, the only data required are those represented in Table 1 (Kendall 2005). This very simple data model is actually quite powerful. Software, like NC SLAAP, can then create links between the transcript data and the audio file from which the transcript is based, and phonetic software (such as Praat in the case of NC SLAAP) can be integrated with the transcript to allow for real-time phonetic analysis. In other words, 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 and can be reconstructed from the audio itself. At the same time, an approximation of standard orthography (following Chafe 1993) is sufficient for the transcript text because pronunciation features (e.g. vowel qualities, *r*-vocalization, etc.) can be listened for or examined instantly via a spectrogram. The use of standard orthography also allows for easier searching by users.


[Analyze: ☑ Spectrogram | ○ Formants | ● Pitch | ○ Intensity | Use ImageMap ☑] [Audio Player ☑] [ImgHt +/-]

[Praat Settings: default | Time Step: 0.01 | Pitch Floor: 75 | Pitch Ceiling: 600 | Change Settings | Save Settings as]

Data: time = 842.105 s ; pitch = 244.44161330477269 Hz

```
Line    Start      Spkr  Pitch & Spectrogram & Text & Audio        End
<1120>  [841.732]  GM:       they say their floor is just          [843.418]
```

[duration: 1.686 s | syllable count: 6 | speech rate: 3.56 syls/sec]

Figure 3: Screenshot showing a transcript line with phonetic data

Figure 3 shows a screenshot from the NC SLAAP software demonstrating an in-depth view of one transcript line. This example shows a pitch plot as well as a spectrogram, though other views are available. Note also that the audio for the line can be listened to through an embedded audio player and that numerical data (in Figure 3 on pitch) can be obtained at the click of the mouse. Additionally, multiple transcript lines can be displayed in this detailed format on the same page, allowing for the comparison between utterances or individual word-tokens.

4 Towards Holism in Quantitative Sociolinguistics

Another major benefit of the NC SLAAP approach to the treatment of natural speech data is that quantitative and qualitative analyses can be better integrated with one another. Since linkages are maintained between the quantified data and the speech events that the data are extracted from, analysts can better situate their quantitative data and analyses in terms of the larger discourse. Meanwhile, discourse-level work, typically focused on more qualitative questions, can more easily integrate quantitative measures. This section seeks to illustrate some of these benefits by highlighting just two of the features of the NC SLAAP software.

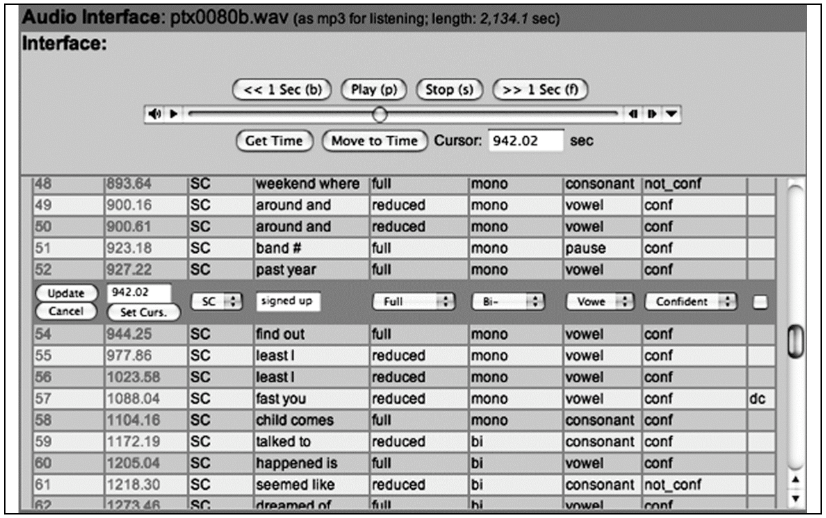
4.1 Variable Tabulation in a Data-Based Model

Variable tabulating—the counting and comparing of different realizations of the same linguistic variable—is a methodological centerpiece of quantitative variationist sociolinguistics (cf. Labov 1966, Wolfram 1993, 2006). Yet, behind variable tabulation practices, summaries, and analyses, there exist theoretical questions (e.g. which forms should be counted or not counted as significant, or meaningful, language variation? cf. Blake 1997) and methodological questions (such as those involving inter-analyst agreement; e.g., do multiple analysts obtain the same counts from the same source recording?[7]). These sorts of concerns remain, with few exceptions (such as Blake 1997), under-examined and under-reported.

NC SLAAP's variable tabulation tool helps to counter some of these problems by making tabulation practices more transparent and individual

[7] Interestingly, despite a relatively long history of explications of sociolinguistic methodology and the use of the linguistic variable construct (e.g. Labov 1966, Wolfram and Fasold 1974, Wolfram 1993, Milroy and Gordon 2003, etc.), inter-analyst agreement has been little discussed. It is often mentioned in passing in discussions of methodology (e.g. Wolfram 1993:215–216, Milroy and Gordon 2003:151), but it is telling that there is not an archetypal (or even good) citation for a thorough account of the *inter-analyst agreement* problem.

tabulation data more accessible for easy review. Following the focus on temporal accuracy behind transcription implementation (as discussed above), tabulations in NC SLAAP are time-stamped entries comprised of enumerated fields linked to the core audio. Analysts are able to review their own tabulations at the click of the mouse and colleagues can easily share and review each other's work. Furthermore, coding analysts are prompted to mark their level of confidence for each tab, which provides a helpful mechanism for the review of putative or less confident tabulations. For illustration, Figure 4 shows a part of the variable tabulation screen in NC SLAAP for the variable *syllable-coda consonant cluster reduction*.[8]



Audio Interface: ptx0080b.wav (as mp3 for listening; length: *2,134.1* sec)

Interface:

Cursor: 942.02 sec

48	893.64	SC	weekend where	full	mono	consonant	not_conf	
49	900.16	SC	around and	reduced	mono	vowel	conf	
50	900.61	SC	around and	reduced	mono	vowel	conf	
51	923.18	SC	band #	full	mono	pause	conf	
52	927.22	SC	past year	full	mono	vowel	conf	
Update Cancel	942.02 Set Curs.	SC	signed up	Full	Bi-	Vowe	Confident	□
54	944.25	SC	find out	full	mono	vowel	conf	
55	977.86	SC	least I	reduced	mono	vowel	conf	
56	1023.58	SC	least I	reduced	mono	vowel	conf	
57	1088.04	SC	fast you	reduced	mono	vowel	conf	dc
58	1104.16	SC	child comes	full	mono	consonant	conf	
59	1172.19	SC	talked to	reduced	bi	consonant	conf	
60	1205.04	SC	happened is	full	bi	vowel	conf	
61	1218.30	SC	seemed like	reduced	bi	consonant	not_conf	
62	1273.46	SC	dreamed of	full	bi	vowel	conf	

Figure 4: Screenshot of a tabulation form and audio player

In addition to the benefits of coding transparency and improved accuracy, this method also provides simple logistical benefits. Through the web-based interface, analysts can tab their data from any Internet-connected computer and can leave their work and return to it without losing their place in the audio. NC SLAAP also allows users to view tabulation summary results directly from the website as well as to download tab-delimited versions of the tabulation sheet suitable for opening in Microsoft Excel or other spreadsheet applications. In sum, the procedural enhancements provided by the NC SLAAP implementation of variable tabulation enable

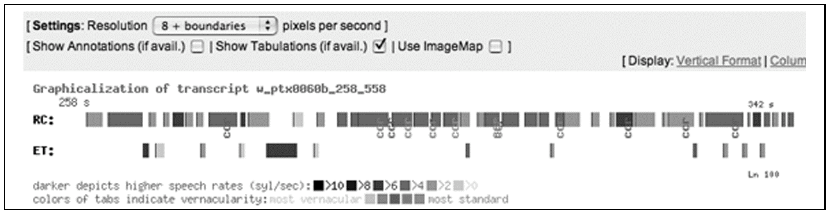
[8] See, for example, Guy (1980) for a thorough discussion of this variable.

general methodological and theoretical advances to this foundational component of quantitative sociolinguistics.

4.2 Visualizing Speech as Data

Since transcript information in NC SLAAP is stored as data separate from its formatting and linkages are maintained between the transcript information, additional data and metadata (such as variable tabulations), and the core audio recordings themselves, the NC SLAAP software is able to perform sophisticated transformations and visualizations on the aggregate data. Figure 5 displays the results of one such visualization within the NC SLAAP software, *graphicalization*. In this presentation, a transcript is displayed in a purely visual format with no text. Shading indicates speech rate,[9] while variable tabulation data are overlaid on the depiction of the transcript so variable constellations are viewable along with a summary representation of the speech event. Importantly, this presentation gives analysts a simple visual overview of the unfolding of the speech event.

Who talks when, and in response to whom? Do the interviewers and interviewees accommodate to one another over the course of the interview (in terms of variable productions, speech rate, gap length, etc.)? This sort of view of the data can help motivate answers to these questions and importantly allows us to see speech data in a more holistic way than traditional transcription or tabulation presentation methods have allowed.



[**Settings**: Resolution 8 + boundaries pixels per second]
[Show Annotations (if avail.) □ | Show Tabulations (if avail.) ☑ | Use ImageMap □]
[Display: Vertical Format | Colum

Graphicalization of transcript w_ptx0060b_258_558

darker depicts higher speech rates (syl/sec): ■>10 ■>8 ■>6 ■>4 ■>2 ■>0
colors of tabs indicate vernacularity: most vernacular ■■■■■ most standard

Figure 5: Screenshot showing an excerpt of a transcript *graphicalization*

[9] Darker shading represents faster speech. Speech rate is determined by an algorithm that counts syllables in the orthographic representation of the speech and then divides that count by the exact duration of the utterance. The algorithm only approximates a syllable count (at about 77% accuracy; tests indicate, however, that less than 2% of the syllable counts are off by more than 1 syllable). While the algorithm could use some improvement (and is quite limited since it is based on orthography—e.g., is "probably" two or three syllables?), it is, I argue, sufficient for its present uses.

Of course, this is not the first attempt to examine variable clustering in context. For example, Podesva's (2007) recent presentation at the Linguistic Society of America's 2007 Annual Meeting highlighted both the importance of examining bundles of variables and an innovative way to present those bundles. Meanwhile, a number of earlier projects (e.g. Wolfram 1985) situated variable tabulations within their discourse context. Of particular note about the NC SLAAP *graphicalization* feature is that the software creates this presentation automatically and dynamically. As data and metadata accumulate in the system—that is, as users tabulate variables and add transcripts and notes—the richness of the representation grows.

5 Future Directions

NC SLAAP seeks to provide its users better tools and better data with which to undertake their studies, whether traditional sociolinguistic pursuits or investigations of new avenues for research. As this paper has attempted to show, this sort of interactive archive increases the utility of speech data. Over time, the steady accumulation of metadata—researcher's notes, transcripts, variable tabulations, and so forth—enhances the corpus overall. Instead of data becoming less usable over time (as the original analysts move on, notes are misplaced, the audio tapes deteriorate, etc.), the speech data stored in NC SLAAP become richer and more usable.

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 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, an eventual goal is to make the NC SLAAP software available to the greater sociolinguistic community, either via a more widely accessible web server or through the distribution of the software itself, so that other researchers can make use of these tools to store and interact with their own archives.

References

Blake, Renée. 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.

Chafe, Wallace. 1993. Prosodic and functional units of language. In *Talking Data: Transcription and Coding in Discourse Research*, ed. J. Edwards and M. Lampert, 33–43. Hillsdale, New Jersey, Lawrence Erlbaum.

Du Bois, John, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In *Talking Data: Transcription and Coding in Discourse Research*, ed. J. Edwards and M. Lampert, 45–89. Hillsdale, New Jersey, Lawrence Erlbaum.

Edwards, Jane. 2001. The transcription of discourse. In *The Handbook of Discourse Analysis*, ed. D. Tannen, D. Schiffrin, and H. Hamilton, 321–348. Oxford and Malden, Massachusetts, Blackwell.

Edwards, Jane and Martin Lampert, eds. 1993. *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, New Jersey, Lawrence Erlbaum.

Figueroa, Esther. 1994. *Sociolinguistic Metatheory*. Oxford, Pergamon.

Guy, Gregory. 1980. Variation in the group and individual: The case of final stop deletion. In *Locating Language in Time and Space*, ed. W. Labov, 1–36. New York, Academic Press.

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, DC, Center for Applied Linguistics.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In *Language and Use: Readings in Sociolinguistics*, ed. J. Baugh and J. Sherzer, 28–53. Englewood Cliffs, New Jersey, Prentice-Hall.

Kendall, Tyler. 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.

Kendall, Tyler. 2007. On the status of pause in sociolinguistics. Paper presented at the Linguistic Society of America 2007 Annual Meeting, Anaheim, California, January 2007.

Ochs, Elinor. 1979. Transcription as theory. In *Developmental Pragmatics*, ed. E. Ochs and B. Schieffelin, 43–72. New York, Academic Press.

Podesva, Robert J. 2007. Social meaning in the interaction of variables. Paper presented at the Linguistic Society of America 2007 Annual Meeting, Anaheim, California, January 2007.

Poplack, Shana. 1989. The care and handling of a mega-corpus: The Ottawa-Hull French project. In *Language Change and Variation*, ed. R. Fasold and D. Schiffrin, 411–451. Amsterdam, Benjamins.

Wolfram, Walt. 1985. Variability in tense marking: A case for the obvious. *Language Learning* 35, 2:229–253.

Wolfram, Walt. 1993. Identifying and interpreting variables. In *American Dialect Research,* ed. D. Preston, 193–221. Amsterdam, Benjamins.

Wolfram, Walt. 2006. Variation in language: Overview. In *Encyclopedia of Languages and Linguistics* (2nd Edition), ed. K. Brown, 333–340. Oxford, Elsevier.

Wolfram, Walt and Ralph W. Fasold. 1974. *The Study of Social Dialects in American English*. Englewood Cliffs, New Jersey, Prentice-Hall.

English Department
Duke University
Durham, NC 27708
tsk3@duke.edu