Language and Linguistics Compass 2/2 (2008): 332–351, 10.1111/j.1749-818x.2008.00051.x

On the History and Future of Sociolinguistic Data

Tyler Kendall*
Duke University/North Carolina State University

Abstract

Recordings of natural speech play a central role in the diverse subdisciplines of linguistics. The reliance on speech recordings is especially profound in sociolinguistics, where scholars have developed a range of techniques for eliciting and analyzing natural speech. However, sociolinguists have rarely focused explicitly on the storage, management, and preservation of their data – the interfaces to their data – and this lack of focus has had consequences for the advancement of the field. In this essay, I briefly review the history of data-management practices within sociolinguistics, insofar as these practices have been discussed in the literature. I then propose new ways to consider and approach natural speech recordings as data for sociolinguistic analysis and provide examples from the North Carolina Sociolinguistic Archive and Analysis Project, a Web-based digitization and preservation project, to highlight the analytical as well as theoretical benefits of more rigorous considerations of 'data' within sociolinguistics.

Introduction

Linguists' data come in many forms. From grammaticality judgments to reaction times to acoustic measurements, linguists build their theories and understanding of language through a wide variety of data types. Although the notion of what constitutes primary data can differ from research project to research project and from researcher to researcher, within sociolinguistics – the study of language in its social context – 'data' often involve some sort of empirical language recordings, such as recordings of naturally occurring speech. This reliance on naturalistic spoken data is so profound that a large proportion of energy spent developing sociolinguistic practices has focused on the refinement of the 'sociolinguistic interview' as a method for the acquisition of naturalistic, conversational speech (cf. Labov 1984).

The point of this essay, however, is not on data acquisition in sociolinguistic research, nor on the subsequent analysis of those data; there are vast numbers of publications on sociolinguistic analysis elsewhere (textbooks

alone include, for example, Wolfram and Fasold 1974; Milroy 1987; Milroy and Gordon 2003; Tagliamonte 2006). Instead, this essay focuses on two topics not widely discussed within the sociolinguistic literature: what we do with the recordings themselves once we obtain them (and further when the original research project is over) and how we approach these recordings as 'data' in the first place.

Through the work of 'doing' sociolinguistics, we amass a huge quantity of potential data. In the cabinets of the North Carolina Language and Life Project's linguistics lab, for example, we have a growing collection of over 1,500 sociolinguistic interview recordings. These are mostly on cassette tape, but some are in other formats, like CD-ROM or digital videotape. Most of our recent recordings were 'born digital' on solid-state recorders and are housed directly on a number of networked computer hard disks. The bulk of the collection – recordings whose original research project has passed – rest quietly, mostly forgotten until recently, when, not unlike a number of research groups and individual scholars elsewhere, we began the effort of digitizing and organizing our collection of recordings. This initiative, however, titled the North Carolina Sociolinguistic Archive and Analysis Project (NC SLAAP),[1] seeks to do more than just to preserve the recordings. The project centers on the development of Web-based data management and analysis software and, through it, we are reconsidering the role of the sociolinguistic archive, our interfaces with it, and the nature of speech data itself.

This essay begins by briefly reviewing discussions of data-management (e.g., organization, storage, accessibility, and preservation) practices within the sociolinguistic literature. It highlights advances that have been made in other linguistic disciplines, such as documentary linguistics (cf. Himmelmann 1998) and corpus linguistics, and considers how those advances can be used to improve our understanding of and interfaces to sociolinguistic data. It then discusses the *conceptual* and *analytical* benefits of improved, explicit data-management strategies, providing examples from our work with the NC SLAAP.[2]

A Short History of (the Treatment of) Data Within Sociolinguistics

There are two main areas within the sociolinguistic literature in which sociolinguistic methodology is typically discussed: research reports, which explicate the methods used by specific projects, and textbooks (e.g., Wolfram and Fasold 1974; Milroy 1987; Milroy and Gordon 2003; Tagliamonte 2006), which provide overviews and syntheses of general practices. Research reports in the early days of sociolinguistic variation studies (e.g., Labov 1966; Wolfram 1969; Sankoff and Sankoff 1973; Trudgill 1974) in particular published thorough accounts of their methods, ranging from explications of their sampling techniques – how and why they chose the informants they did – to discussions of their interviewing techniques and even of

Language and Linguistics Compass 2/2 (2008): 332–351, 10.1111/j.1749-818x.2008.00051.x

training their fieldworkers (cf. Shuy et al. 1968). These methodological reports were an important and necessary step in establishing sociolinguistics as a credible and quantitatively oriented social science and served to aid future scholars by explicitly sharing 'the knowledge of the problems [the researchers] faced and the solutions [they] tried' (Sankoff and Sankoff 1973: 12).

However, throughout these literatures, authors' discussions of methods almost always jump from data acquisition to data analysis and/or to demographic and theoretical issues pertaining to analyzing language in relation to social attributes of speakers (as in Wolfram and Fasold 1974; Milroy 1987; Milroy and Gordon 2003). There are numerous robust discussions of issues like how to choose informants, how to elicit and obtain 'good' speech,[3] and how to analyze that speech. However, almost across the board these have neglected to discuss issues in how the speech goes from a 'recording' to 'data' and further how those 'data' are organized and accessed.

Of course, there are exceptions.[4] Poplack (1989) provides an excellent discussion of the Ottawa-Hull French Project's data archive and methodology, a project with a goal to improve methodologies inherent in working with large sets of data for sociolinguistic analysis. She writes in response to the fact that

> one area in which development has been sporadic at best is in the construction of major sociolinguistic data bases. The trade-off between sociological representativeness and ethnographic thoroughness has resulted in insufficient data from a large sample of speakers, or masses of data of questionable generalizability from a few speakers. Efforts to increase quantity or authenticity of recordings are still marked by losses in the quality of the data obtained. And even as a data base reaches respectable size, its accessibility is concurrently hampered by the uneconomical effort needed to search it systematically in studies of individual variables (Poplack 1989: 413).

Her paper provides the most thorough discussion in the literature of many of the steps from determining a sample population, to interviewing and recording that population, to organizing the resulting collection, and to developing a computer-based corpus of the recordings.

Tagliamonte's recent textbook, *Analysing Sociolinguistic Variation* (2006), may be an indication that data-management strategies are becoming more explicit within sociolinguistics. Her text has an entire chapter, 'Data, data, and more data', that reviews a wide range of data-management tasks, from labeling and organizing interviews into a coherent corpus to transcribing the data and working with computerized transcripts and recordings. While I hope this is an indication of a growing awareness of the importance of explicit documentation of data-management methods, it is still the case that sociolinguistics as a field operates with individualistic, and underexamined, data-management strategies. This has resulted in a state where the relationship between 'speech' and 'data' is often poorly defined.

How Does Speech Constitute Data?

One of the issues here is that term 'data' means a lot of different things to different sociolinguists (and more broadly, of course, linguists). That is, throughout many of the diverse approaches subsumed under the rubric of sociolinguistics the term 'data' is used to refer loosely to some sort of captured real-world speech event,[5] but the 'data' on which analyses are based differs greatly depending on the theoretical perspective of the researcher and the substantive questions being asked. Variationist sociolinguists, for example, work with a very different sort of 'data' than do, say, discourse analysts or linguistic anthropologists.

To take a specific example, let us consider the quotative system in English. Since Butters (1982) mentioned the use of *be like* as a preface to a quotation (e.g. 'And he*'s like* "No way!" '), a number of studies (e.g. Romaine and Lange 1991; Ferrara and Bell 1995; Tagliamonte and Hudson 1999; Buchstaller 2006) have investigated the increasing use of *be like* as a quotation introduction. While many studies of quotative *be like* (such as those listed above) are conducted in similar ways – roughly by counting instances of *be like* and other quotative markers (such as *say* or *go*) and comparing the relative occurrences of each with other linguistic features of the discourse and social attributes of the speakers – a study of the feature could just as well be undertaken in widely different ways. Different approaches could depend not only on the researcher's theoretical background, but also on the format and type of 'data' used for the analysis or even on the original data sample or instance that inspired the particular analysis in the first place. So, a researcher who becomes interested in studying *be like* from seeing its use in transcripts may conceptualize a study very differently – most likely focusing on morphosyntactic, discourse, or corpus-based approaches – than one who hears instances of *be like* in audio recordings, where phonological and phonetic aspects of the occurrences – such as variation in intonation or stop-release, or the impersonation of others' 'voices' – may be more striking.

In other words, two analysts working within the same theoretical framework, even potentially on the same set of substantive issues, may have very different conceptions of what constitutes the data for their analyses. They may also have different views on just what exactly their substantive issues are depending on the nature of the data they have at hand. The different approaches in turn can influence the researchers' findings and whether or not those findings can (or should) even be compared to one another. Different ways of conceptualizing speech as data have different outcomes, whether or not those outcomes are intentional.

To further exemplify this problem, let us turn briefly to two primary notions of 'data' used within sociolinguistic disciplines: variable tabulations and orthographic transcriptions.

VARIABLE TABULATION IN PRACTICE AND THEORY

Tagliamonte tells us 'the advantage of variation analysis is working with real data, often from representative samples of communities, and from scrutiny of hundreds and perhaps thousands of instances of the linguistic variable' (2006: 74–5). To isolate those 'hundreds and perhaps thousands' of variable realizations involves a great deal of work on the part of the analysts as they filter the complexities of speech down to analyzable formats. Variable tabulating – the extraction and coding of different realizations of the same linguistic variable – is the central methodology within quantitative variationist sociolinguistics and the primary means for undertaking this filtration process (cf. Labov 1966; Wolfram 1993, 2006).

While variationists may agree in principle on how they do their work, there are no agreed upon standards for how they move from a real-world speech event to a set of quantified data for a particular linguistic feature. Wolfram (1993) gives some guidelines for determining and evaluating linguistic variables, but also points out a number of problems with the heuristic, like the difficulty (and under-reporting within the literature) of how to determine what range of variation is subsumed by a single 'linguistic variable' (see also Wolfram 2006), as well as problems of inter-analyst agreement and even intra-analyst reliability.

Blake (1997) and Rickford, Ball, Blake, Jackson and Martin (1999) provide excellent (and exceedingly rare) discussions of some of these issues. Blake (1997) investigates the forms not counted by different groups of researchers in the study of copula deletion in African-American English, while Rickford et al. (1999) examine how frequencies of copula contraction and deletion have been computed by different researchers and how the differences in methodology affect the results of the analysis. Rickford et al. (1999), in particular, demonstrate that different theoretical assumptions and views about data impact the quantitative outcome of a study with consequences of greater theoretical and descriptive import. The significance of these papers is punctuated by the fact that there are countless variable tabulation issues that go undiscussed in the literature.

Finally, it must be noted that this sort of approach on its own may have larger problems, as Macaulay notes in a discussion of the sociolinguistic interview,

> Interviews as a whole have not been used as data. It is somewhat paradoxical that most of the speech collected in sociolinguistic surveys remains unanalyzed. Most investigators have followed Labov's lead in concentrating on a few variables. In this approach a certain number of tokens are extracted from the interview and coded. The analysis then deals with these tokens and the remainder of the interview is ignored. [. . .] The concentration on such variables has, however, an influence on the kind of questions that are asked (1991: 5).

In other words, while often a necessary procedure, the focus on a handful of quantified linguistic features divorced from their original context potentially ignores a great deal of the 'data'. It also may limit the questions that can be asked (and the answers that can be obtained). Thinking further about quotative *be like*, might morphosyntactic or text-based corpus analyses be missing important aspects of the variation?

ORTHOGRAPHIC TRANSCRIPTION IN PRACTICE AND THEORY

Both within and outside linguistics, the orthographic transcript is 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; yet, the act of transcription, especially by beginning transcribers, is often undertaken as a purely methodological activity, as if it were theory neutral. Each decision that is made while transcribing influences and constrains the resulting possible readings and analyses (Ochs 1979; Mishler 1991; Bucholtz 2000; 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 in which the transcript may lead analysts.

Despite the widely declared stance that 'transcriptions are not substitutes for the original recordings but additional tools which can be used to help analyze and understand these recordings' (Liddicoat 2007: 13, on transcripts in Conversation Analysis; see also Tagliamonte 2007), transcripts all too often become the sole means of exemplifying and sharing access to speech recordings. It is also possible that these non-aural representations do end up as the 'data' used for analysis in lieu of the actual recordings, despite the frequent warnings to the contrary.[6]

A major part of the problem behind the use of transcripts for language research is that the text of a transcript is always an incomplete and interpreted record of the original interaction (Edwards 2001). As we will discuss below, transcripts can be improved in ways that can mitigate some of these central problems.

Data vs. Metadata

But, what are variable tabulations and what are transcripts? Are they really 'data'? In a sense, they often are, but in a better sense they should perhaps be considered 'metadata', data about the core data.[7] They are abstractions, representations of the real-world speech events that actually constitute the core data for sociolinguistic research. This statement on its own may not be particularly striking, but it has important consequences. Metadata – such as variable tabulations and transcripts – are inseparable from their

source data and should not be analyzed as things in their own right, divorced from the source recording or real-world interaction.

Clearly, abstracting from the original speech event is an integral part of analyzing that speech event. Despite our best attempts, we can never analyze the 'fleeting events of an interaction' (in the words of Edwards 2001: 321) without using some sort of representation (such as a transcript or a spreadsheet of variable tabulations) as a proxy for the data. The important point is that analysts must always keep in mind their core data – not even the recording itself, but rather the never-fully-reconstructible real-world interaction that underlies the recording.

Figure 1 provides a schematic illustrating the ways in which speech data are traditionally abstracted in their various representations. Each level in

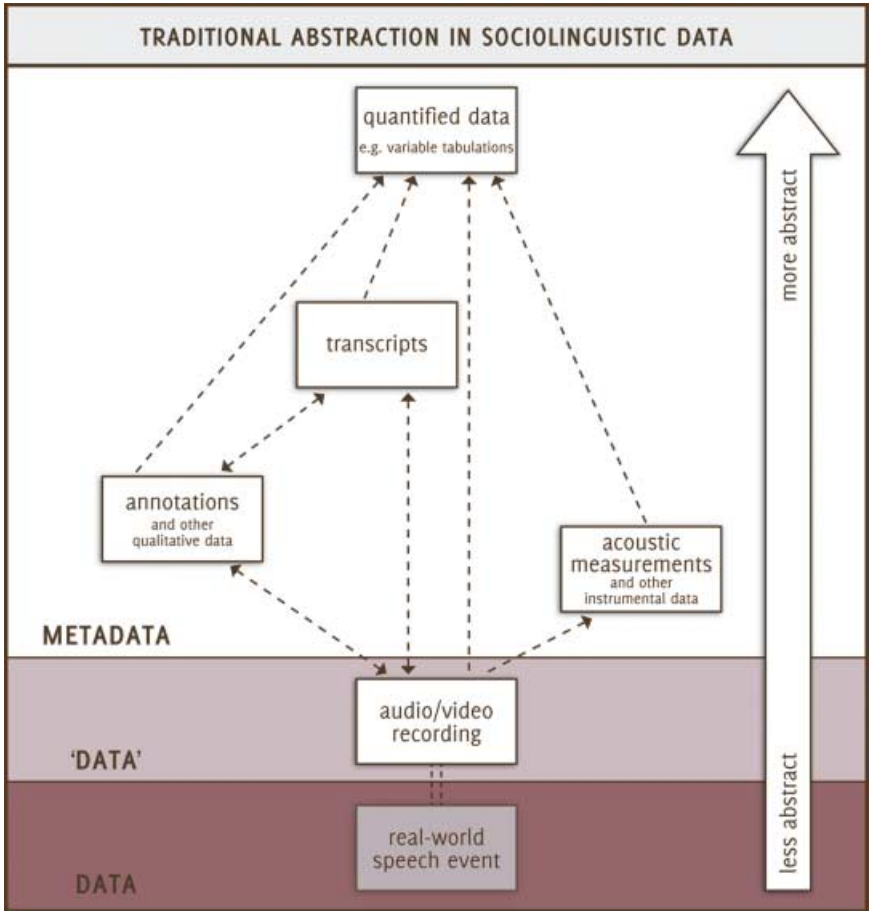

Fig. 1. A schematic indicating the levels of data and metadata for spoken language recordings as used in sociolinguistic research.

the schematic should be considered a layer of metadata over the real-world speech event, the actual 'data' of a particular dataset. The real-world event is shaded and connected to the recording without arrows to represent the fact that it is evanescent, over as soon as it is begun, and rarely (if ever) available to the researcher for deep analysis.

In Figure 1, even the connection between the speech event and the recording is represented with a dashed line, because the very act of recording is, in itself, an abstraction and therefore does not have a solid connection to the event, the actual 'data'. In the schematic, the recording is described as 'DATA' (in quotes), because it is often used in place of the original interaction, as if they were one and the same. A video recording captures more of the real-world interaction than an audio recording, but even then the video recording may be missing important action (e.g. that occurred off camera). It is also still an abstraction in the sense that it is always ultimately divorced from the real-world context(s) in which the original speech event occurred. Ethnographic inquiry can help to strengthen the connection between a recording and its original context, but even then there is only so much an analyst can know about the full scope of an interaction.

Beyond the recording, each further representation in Figure 1 is a further abstraction, and as such each is illustrated with a dashed connector line, sometimes connecting to the recording, but sometimes also to other layers of metadata, multiply removed from the original speech event. Arrowheads indicate the directions to which some layers of metadata contribute to the accuracy and understanding of other layers. Distance is also used to indicate the degree of abstraction from the core data. Through this depiction, we see that many of the quantified data types that are most useful for analysis are often the most abstract.[8]

This is not to argue that these abstractions are 'bad'. They are necessary mediations between a spoken interaction with its full complexity and the categories of interest to the analyst (cf. Edwards 2001). Variable tabulations and other quantified metadata are often highly useful at least in part by virtue of their degree of abstraction. Instead, this is to argue that analysts should keep in mind the distance a particular set of metadata may have from the speech event itself, to be mindful about that distance in each step of the analysis, and to be explicit about it in publication and presentation.

Treating Speech as Data in the Twenty-first Century

In 1973, Sankoff and Sankoff proposed that, while 'detailed descriptive studies of complex urban communities involve a good deal of painstaking and sometimes tedious work[,] much of this can, however, be reduced by maximal use of automated, computerized data processing techniques' (1973: 47). Yet, many approaches to sociolinguistic data management have

not moved toward computerized techniques, or, rather, those that have, have remained individualistic, ad hoc, and implicit. Sankoff and Sankoff's (1973) work may have been ahead of its time, but we are now well into an era in which the technological advancements of the growing digital age enable vastly improved computerized analysis and storage of speech recordings.

This is not a novel observation. Disciplines outside of sociolinguistics are increasingly focusing on spoken language data-management practices through computer-based methods. Associations like the Open Language Archives Community (OLAC; cf. Bird and Simons 2003) and E-MELD (cf. Aristar-Dry 2002) are working toward the development of inter-operable and standards-based archives for language documentation. Bird and Simons (2003), for example, provide a comprehensive enumeration of 'best practice' guidelines at the same time as they highlight often overlooked potential problems that surround building comprehensive language archives. These efforts appear to be finding success and ought to be informing sociolinguistic practice.[9]

Meanwhile, corpus linguistics has grown dramatically in recent years as both a subfield in its own right and as a set of methods used throughout the linguistic disciplines, with tools and corpora becoming more developed, available, and useful for a broad range of pursuits. More projects across all empirical linguistic disciplines are being explicitly designed around the generation of coherent data collections and more publications are appearing on data collections. Even academic libraries are becoming involved in the development and publication of spoken language collections (Kendall and French 2006; Cooper 2007). Groups such as the Linguistic Data Consortium (LDC) at the University of Pennsylvania (cf. Cieri and Liberman 2006), the TalkBank project (cf. MacWhinney 2007), and the Oxford Text Archive have developed and made available a wide range of language corpora and corpus analysis tools.

Many corpus-generation projects are of direct interest to sociolinguists (e.g. Pitt et al. 2005; Kretzschmar et al. 2006; Kendall 2007a, see also Bauer 2004). Beal, Corrigan, and Moisl's recent two volumes (2007b,c), *Creating and Digitizing Language Corpora*, contain papers discussing a range of 'unconventional' corpora (from a corpus linguistic perspective) and, as such, have a great deal to offer to sociolinguists. However, many corpora, tools, and other products have not been of great use for sociolinguistic pursuits (as discussed by Kretzschmar et al. 2006: 173–4). Beal et al. (2007a: 2) are correct to point out that their 'volumes are unique, since public output to date has primarily concentrated on describing and assessing the models and methods which underpin conventional corpora and the annotation standards/analytical tools developed specifically for them.'

The point here is that 'normal' sociolinguistic practice could be greatly improved by embracing the sorts of data-management mentalities held by documentary and corpus linguists. In fact, the data collection work that

sociolinguists undertake as part of 'normal' sociolinguistic research could be viewed no differently than the work that goes into the development of a sociolinguistically relevant corpus (in Beal et al.'s 2007a terms, an 'unconventional' corpus), with the exception, perhaps, of how public the final collection might be.

Poplack's forward to Beal et al. (2007b) points out that 'the projected use of the corpus, as *end-product* or *tool*, is clearly the determining factor' (Poplack 2007: xi, emphasis in original) for how a data collection gets treated by its creators. Many explicit corpus creation projects focus on the construction of an 'end-product', whereas for most sociolinguists the utility of the corpus is its role as a 'tool' for researching a particular set of questions. As such, the two groups see their aims as being different and, consequently, the focus in the corpus linguistics literature on compiling corpora has had little traction for sociolinguists. This, at least in part, accounts for the fact that sociolinguistic data collection is rarely treated explicitly as corpus-generation work. Returning to Poplack's preface, considering sociolinguistic data collection and management as corpus work can be of benefit to sociolinguists because of

> the opportunity it affords to serendipitously discover what one wasn't looking for, to characterize the patterned nature of linguistic heterogeneity, and in particular the hidden, unsuspected or 'irrational' constraints that are simply inaccessible to introspection or casual perusal (Poplack 2007: xii).

This is not to argue that sociolinguistics, and all sociolinguists, should move toward corpus linguistic approaches to their data and analyses.[10] Nonetheless, there is no doubt that sociolinguists can learn a lot from thinking about their data as corpora, both for their projects at hand and for future – possibly not yet conceived of – investigations.

Some Benefits of Improved, Explicit Data-Management Work

Returning to my own work with the NC SLAAP, we have found that explicit data-management work leads to more than a better preserved data collection. By digitizing our collection and incorporating the recordings into a centralized repository, we have in a sense put into dialogue our entire collection.[11] This explicit management work creates a level of organization that is more complete and useful than otherwise. It makes for better analyses by giving researchers easier access to their data and better means to support their analyses and investigate problematic cases. It makes it easier to collaborate on research projects and share data and findings. Additionally, it can also create opportunities to evaluate new research questions (e.g., Kendall 2007b). As two examples of this let us briefly consider the approaches to transcription and variable tabulation undertaken within NC SLAAP (a more general overview of NC SLAAP is available in Kendall 2007a).

Fig. 2. Four presentations available in NC SLAAP of the same transcript data.

RE-EXAMINING TRANSCRIPTION

It is possible to ameliorate some of the problems inherent in representing speech in text, if we think deeply about the long-term purpose of our transcripts and reconsider their implementation along those lines (Kendall 2005, 2007a; MacWhinney 2007). This is exemplified, I believe, by the implementation of transcripts in the NC SLAAP software. 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. Transcripts are built using Praat (Boersma and Weenink 2007) to obtain highly accurate start- and end-times for each utterance.[12] By virtue of not being text-documents, the NC SLAAP software can display transcripts in a variety of formats: a 'vertical format' (as in ❶ in Figure 2), a 'column-based' format (as in ❷ in Fig. 2; cf. Ochs 1979), or a 'paragraph' format (as in ❸ in Figure 2). Alternatively, transcripts can also be displayed in a purely visual format (called 'graphicalization', as in ❹ in Figure 2; cf. Kendall 2007a).

With accurate time-alignment, software can also create links between the transcript data and its source audio file, 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 directly from the transcript. 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, although other views are available. The audio for the line can be listened to through an embedded audio player and numerical data (in Figure 3 on pitch) can be obtained at the click of the mouse.

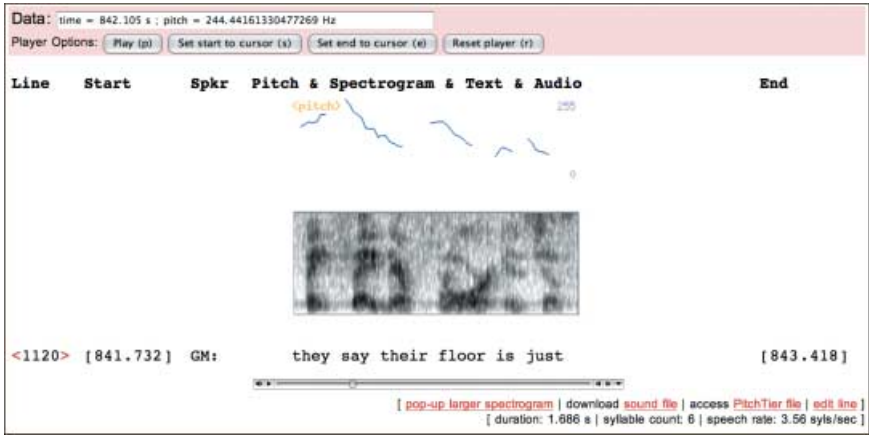

Fig. 3. NC SLAAP screenshot showing a transcript line with phonetic data.

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) is unnecessary and can be reconstructed from the audio itself. At the same time, an approximation of standard orthography (cf. Chafe 1993: 34; Tagliamonte 2007: 211–5)[13] is sufficient for the transcript text, because pronunciation features (e.g. vowel qualities, *r*-vocalization) can be listened for or examined instantly via a spectrogram. The use of standard orthography also allows for easier searching by users. This sort of re-conceptualized model for transcripts makes them more flexible tools for sociolinguistic analysis, useful not only for the projects-at-hand, but for open-ended exploration and future projects.

VARIABLE TABULATION IN A DATA-BASED MODEL

Following the general model of time-aligned transcription discussed above, we can consider some ways to make variable tabulation, that is, coding and extraction, practices more powerful. In the NC SLAAP variable extraction and coding feature, variable tabulations are time-stamped entries linked to the core audio. As with 'normal' tabbing practice, each tabulation consists of a set of enumerated fields. To extract an occurrence of a variable, an analyst simply clicks a button to retrieve the timestamp from the audio player, enters a short string of text describing the context of the occurrence, and then selects the appropriate attribute from a pop-up menu for each field. For illustration, Figure 4 shows a part of the variable tabulation screen in NC SLAAP for the variable 'syllable-coda consonant cluster reduction' (cf. Guy 1980; Wolfram et al. 2000).

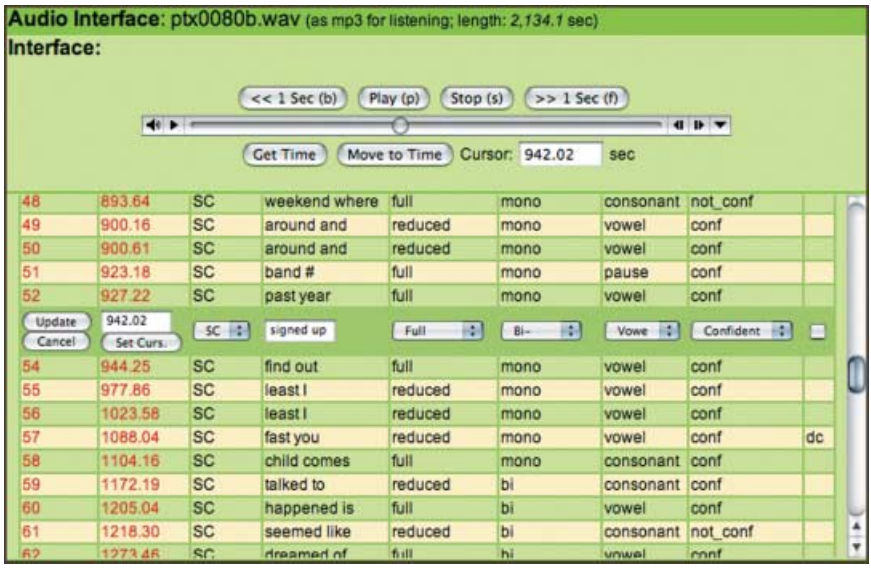

Fig. 4. NC SLAAP screenshot of a tabulation form and audio player.

Through the linkage with the source recording, analysts are able to review their own tabulations by returning to the appropriate moment in the audio 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 with each tab, which provides a helpful mechanism for the review of putative or less confident tabulations. Since tabulations are time-aligned to the source recording, the system maintains a connection between these quantified data and the greater context of the interaction; they remain situated in terms of the larger discourse, enabling more holistic analyses of the data (such as examining the role of topic, stance, or interlocutor effects on variable productions over the course of the speech event).

In addition to the benefits of coding transparency, improved accuracy, and better-situated quantitative data, this method also provides simple logistical benefits. Through the Web-based interface, analysts can tabulate their data from any Internet-connected computer and can leave their work and return to it without losing their place in the audio. Through the timestamps, variable tabulations can be connected to or generated from the transcripts. That is, users can search the transcripts (for features like quotative *be like*), listen to the relevant utterance, and then extract and code the relevant variable all from the same screen. NC SLAAP also allows users to view tabulation summary results directly from the website as well as to download the tabulation sheet in formats suitable for opening in Microsoft Excel or other spreadsheet applications. In summary, the

procedural enhancements provided by the NC SLAAP implementation of variable tabulation enable general methodological and theoretical advances to this foundational component of quantitative sociolinguistics.

Concluding Thoughts

Explicit introspection about what sociolinguists 'do' with their data (and how they do it) can lead to enhanced speech data collections, which in turn are more useful for study. As I have illustrated through the brief discussion of two features of NC SLAAP's software, sociolinguists can move their data collections from 'tapes in a cabinet' to interactive and powerful tools for linguistic analysis. The steady accumulation of metadata – researcher's notes, transcripts, variable tabulations, and so forth – in a system like NC SLAAP enhances the collection overall and makes it richer and more usable over time instead of less usable (as the original analysts move on, notes are misplaced, the audio tapes deteriorate, etc.).

Through a system like NC SLAAP, where all types of metadata are linked to their source audio file and time stamped to the appropriate moment in that audio, we can revise the schematic first depicted in Figure 1. Figure 5 presents this improved schematic, as is conceptualized within NC SLAAP, alongside the original illustration for comparison.[14] In this version, metadata types are closer together, and are connected in all cases with bi-directional arrows that depict the stronger connection between the levels of representation. Thus, metadata are interconnected, each making the others richer, more accurate, and more fully understood. There are also fewer levels of abstraction, because of the iterative connection between the base-data of the recording and each level of metadata.

Of course, we cannot remove all abstraction from the representation of linguistic information (hence, the not-completely-solid lines); it is not possible to do so entirely. However, levels of metadata can be moved closer to their source recording by maintaining strong linkages between the two. A system such as NC SLAAP also provides a level of organization and some basic shared methodologies for all of its users. Thus, it makes more explicit the data-management and analytic practices that operate on the data within the archive.

In closing, the goal of this essay has been to highlight some of the ways in which sociolinguists can reconsider their interfaces to their data collections and the subsequent benefits both for their current work and for future work with those collections. Two features of NC SLAAP's software – its implementations of variable tabulation and transcription – have been presented here, because they instantiate an approach to improved sociolinguistic data management, and, ultimately, demonstrate that regardless of the specific model adopted, it is time for sociolinguists to be more explicit and rigorous about their data-management practices and the ways in which they conceptualize speech as data.

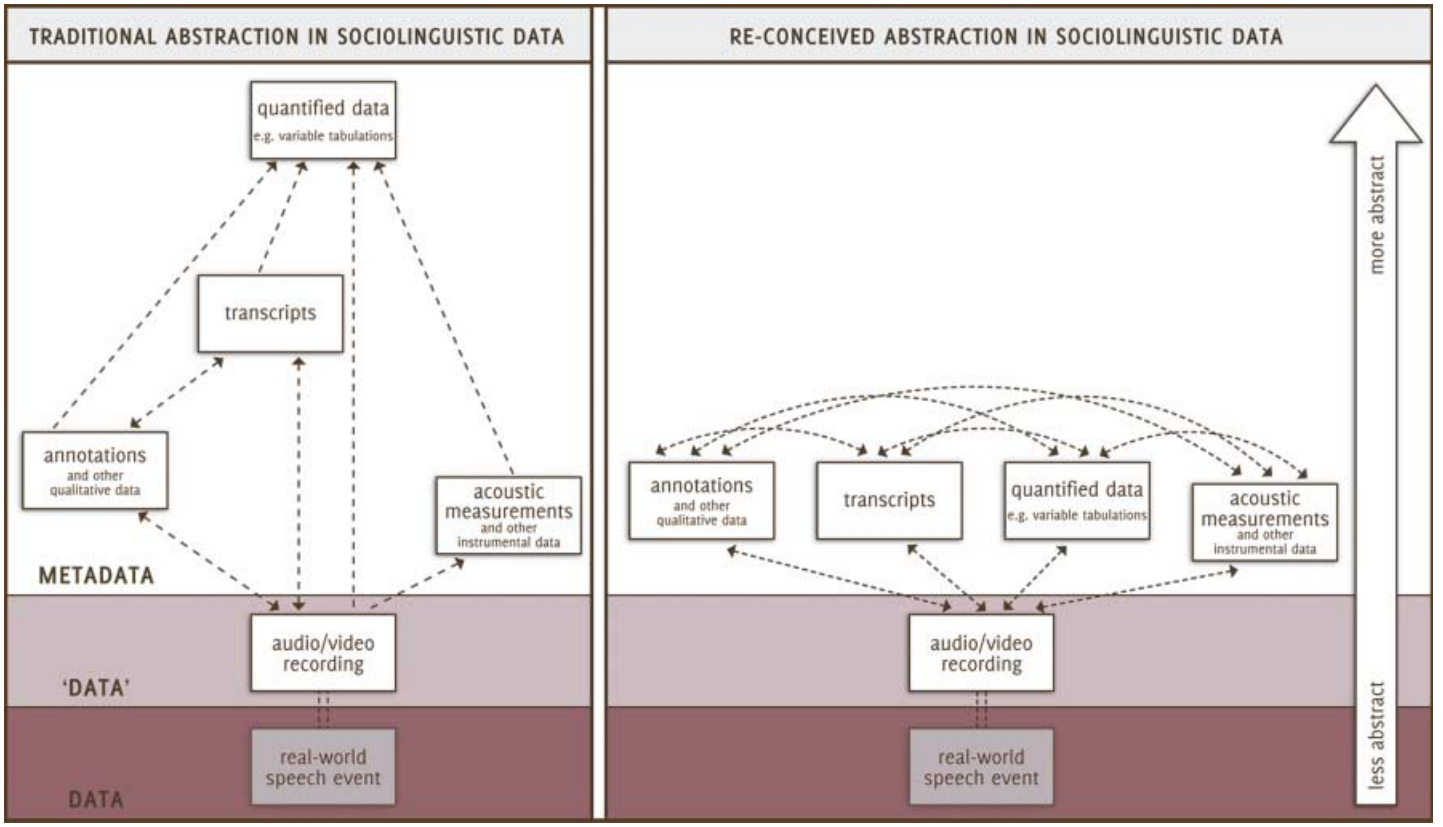

Fig. 5. The schematic from Figure 1 alongside a revised schematic showing the levels of data and metadata as conceptualized within NC SLAAP.

Language and Linguistics Compass 2/2 (2008): 332–351, 10.1111/j.1749-818x.2008.00051.x

Acknowledgements

The impetus for this paper came from discussions with Natalie Schilling-Estes about two intersecting projects: my work over the past few years on NC SLAAP; and, a panel – *Re-examining Language Data in the American South* – given by myself, Todd Cooper, Christine Mallinson, and Walt Wolfram at the 74th Southeastern Conference on Linguistics (SECOL 74) in April 2007. NC SLAAP was begun with support from the North Carolina State University Libraries and the William C. Friday Endowment at North Carolina State University. The Duke University Graduate School has also contributed funding. Much of the data stored in NC SLAAP were obtained through support by the National Science Foundation (including grant number BCS-0542139). I am deeply grateful to all of these funding sources for making the project possible. The figures in this paper were made with the help of Charlotte Vaughn. Christine Mallinson, Walt Wolfram, Erik Thomas, Charlotte Vaughn, Lisa Selkirk, and Gerard Van Herk have all been influential in my work on this project, and consequentially in this paper. I thank them all. Finally, I also thank the anonymous reviewer whose comments have greatly improved this paper.

Short Biography

Tyler Kendall is a PhD Candidate in the joint program in English Linguistics at Duke University and North Carolina State University. His dissertation work focuses on sociolinguistic methodology and notions of 'data' within sociolinguistics, using the North Carolina Sociolinguistic Archive and Analysis Project as a testing ground for new, computer-driven approaches to sociolinguistic analysis. His primary research interests center on understanding variation in spoken language, discourse, and interaction (and what that tells us about linguistic theory) through quantitative discourse analysis, variationist sociolinguistics, and sociophonetics. He has led NC SLAAP since the project's inception in early 2005.

Notes

⋆ Correspondence address: Tyler Kendall, Department of English, Box 90017, Duke University, Durham, NC 27708, USA. Email: tsk3@duke.edu.

[1] More information about NC SLAAP is available on the project's website, at http://ncslaap.lib.ncsu.edu/. Also, see Kendall (2007a) for a more complete discussion of the project than is given here. Information about the North Carolina Language and Life Project – a sociolinguistic research and outreach initiative at North Carolina State University – is also available online, at http://www.ncsu.edu/linguistics/ncllp/.

[2] While this paper touches on the *technical* benefits of following 'best practices' in the management of linguistic data, it does not seek to provide a thorough discussion along those lines. See Bird and Simons (2003) and the OLAC website (http://www.language-archives.org/) for important discussions of 'best practice' guidelines for language resources. Plichta and

Kornbluh (2002) and the National Initiative for a Networked Cultural Heritage's guide (NINCH 2003) are good resources for specific information about digitizing audio recordings.

[3] The discussions about recording 'good' speech, however, center on methods to elicit and record naturalistic, or casual, speech. None of the major overviews of sociolinguistic field methodology give more than cursory mentions of the technical aspects of recording interviews – which kind of recording device to use, what kind and how many microphones to use, and where to place them, etc. This, I believe, remains a major methodological problem for sociolinguistics, but is, unfortunately, outside the scope of this paper.

[4] Many of the early, foundational sociolinguistic projects have entire publications dedicated to their field and laboratory methods, and a handful of these do at least touch on aspects of their data management. For example, Labov (1984: 52) ends with a short section mentioning the Philadelphia Language Change and Variation Project's archive of recordings and describing its size and who can access the archive. Shuy, Wolfram, and Riley (1968) provide a more thorough account of the treatment of the Detroit Dialect Study's data and recordings than most other projects give; they give an extensive accounting of their field methods, explaining and commenting upon everything from determining a sampling system to computer coding their data and from hiring interviewers and an administrative assistant to training them in field interviewing and phonetic transcription. Similarly, Sankoff and Sankoff (1973) provide a thorough overview of the field and laboratory methods for their sociolinguistic study of Montréal French, including a discussion of their computerized transcription system and an overview of their complete database, which they enumerate in detail:

> (i) 120 reels of taped interviews (2 copies);
> (ii) 64 boxes, most of them full, of computer cards containing transcriptions, about 100,000 cards in all;
> (iii) computer print-outs (in several copies) in readable format;
> (iv) in addition, we are presently storing corrected transcriptions on a master computer tape. Thus, to date, 40 interviews, over 20 boxes of cards, are now stored on a single reel of tape at the Centre de Calcul (Sankoff and Sankoff 1973: 42).

At first glance, it may seem that the sort of detail provided by Sankoff and Sankoff is superfluous. One could ask: what benefit do readers gain by their listing the detailed contents of their linguistic closet? I believe, however, that even the simple list of the data provided by Sankoff and Sankoff (1973) is greatly useful. On one level, it provides specific and straightforward information about the data collection (how it is organized and where one can find it). On another level, it indicates the ways in which the analysts interacted with their data.

[5] Of course, there are a number of sociolinguistic pursuits that do not focus on speech recordings as their primary data (e.g. perceptual dialectology; cf. Preston 1989). As is clear to the reader, those sorts of data are not the focus of the present essay.

[6] Some of the time this is validated and valid. For example, Poplack (1989: 436–7) was 'confident that the Ottawa-Hull corpus may be used to study morphosyntactic and lexical phenomena without the necessity of recourse to the audio tapes.' Others (e.g. Rickford and Théberge-Rafal 1999; Buchstaller 2006) have also shown that many kinds of morphosyntactic and lexical studies can be carried out successfully on transcripts. However, what is concerning here is that since no standard publication practices exist for explicating analytic methods, it is quite possible that any number of studies do base their analyses on transcripts instead of the source recordings in ways that are problematic.

[7] The term 'metadata' is used here more generally than it is sometimes used elsewhere. Bird and Simons (2003: 573), for example, use the term for information specifically about 'a physical object or a digital resource,' but I choose to use the term more broadly for any sort of data about data.

[8] This conception does privilege instrumental quantitative data over impressionistic quantitative data since instrumental data, such as vowel formant measurements, are measured directly from the recording and, therefore, can be considered to be closer to the recording. A number of factors still contribute to the abstractness of acoustic measurements, such as the facts that analysts determine where in the speech to measure based on interpreted cues and choose the settings to use for the measurements based on impressionistic assessments of the speech.

[9] Cf. Boas (2006) for a discussion and assessment of one implementation of Bird and Simons' (2003) 'best practice' recommendations.
[10] Conversely, it perhaps is to argue that corpus linguists should look to the data collection work of sociolinguists – especially those whose fieldwork is ethnographically informed – for thorough sources of data for their corpora.
[11] For example, our descriptive metadata, that is, the information stored about each interview, speaker, and research project (again cf. Bird and Simons 2003 for an excellent discussion of this kind of metadata), along with our transcripts and researcher notes are all searchable both within and across projects. Older 'data' and metadata are just as easily retrieved as new materials.
[12] See also Barbiers, Cornips, and Kunst (2007) on using Praat for time-aligned transcription. MacWhinney (2007) discusses some other tools that can be used for this purpose, such as Transcriber, and CLAN. Kendall (2005), MacWhinney (2007), and others (e.g. Edwards 2001) provide general discussions on the benefits of time-aligned transcripts.
[13] Also, see Preston (2000) on the importance of the choice of orthography in transcription.
[14] Of course, NC SLAAP is not the only system or approach to lessen the distances between data and metadata. For example, the 'annotation graph' framework of Bird and Liberman (2001) also achieves this 're-conceived' data model.

Works Cited

Aristar-Dry, Helen. 2002. E-MELD: overview and update. Paper presented at the International Workshop on Resources and Tools in Field Linguistics (LREC 2002), 26–27 May, Las Palmas, Canary Islands, Spain.

Barbiers, Sjef, Leonie Cornips, and Jan Pieter Kunst. 2007. The syntactic atlas of the Dutch dialects (SAND): A corpus of elicited speech and text as an online dynamic atlas. Creating and digitizing language corpora. Volume 1: synchronic databases, ed. by Joan Beal, Karen Corriga and Hermann Moisl, 54–90. New York, NY/Basingstoke, UK: Palgrave-Macmillan.

Bauer, Laurie. 2004. Inferring variation and change from public corpora. The handbook of language variation and change, ed. by J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes, 97–114. Malden, MA/Oxford, UK: Blackwell.

Beal, Joan, Karen Corrigan, and Hermann Moisl. 2007a. Taming digital voices and texts: models and methods for handling unconventional synchronic corpora. Creating and digitizing language corpora. Volume 1: synchronic databases, ed. by Joan Beal, Karen Corrigan and Hermann Moisl, 1–16. New York, NY/Basingstoke, UK: Palgrave-Macmillan.

—— (eds.) 2007b. Creating and digitizing language corpora. Volume 1: synchronic databases. New York, NY/Basingstoke, UK: Palgrave-Macmillan.

—— (eds.) 2007c. Creating and digitizing language corpora. Volume 2: diachronic databases. New York, NY/Basingstoke, UK: Palgrave-Macmillan.

Bird, Steven, and Mark Liberman. 2001. A formal framework of linguistic annotation. Speech Communication 33.23–60.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. Language 79.557–82.

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.

Boas, Hans C. 2006. From the field to the web: implementing best-practice recommendations in documentary linguistics. Language Resources and Evaluation 40.153–74.

Boersma, Paul, and David Weenink. 2007. Praat: doing phonetics by computer [computer program]. <http://www.praat.org/>.

Bucholtz, Mary. 2000. The politics of transcription. Journal of Pragmatics 32.1439–65.

Buchstaller, Isabelle. 2006. Diagnostics of age-graded linguistic behaviour: the case of the quotative system. Journal of Sociolinguistics 10.3–30.

Butters, Ronald. 1982. Editor's note [on 'be + like']. American Speech 57.149.

Chafe, Wallace. 1993. Prosodic and functional units of language. Talking data: transcription and coding in discourse research, ed. by Jane Edwards and Martin Lampert, 33–43. Hillsdate, NJ: Lawrence Erlbaum.

Cieri, Christopher, and Mark Liberman. 2006. More data and tools for more languages and research areas: a progress report on LDC activities. Paper presented at the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

Cooper, Todd. 2007. Documenting the American South. Paper Presented at the Southeastern Conference on Linguistics (SECOL) 74, Natchitoches, LA.

Edwards, Jane. 2001. The transcription of discourse. Handbook of discourse analysis, ed. by Deborah Tannen, Deborah Schiffrin and Heidi Hamilton, 321–48. Malden, MA/Oxford, UK: Blackwell.

Ferrara, Kathleen, and Barbara Bell. 1995. Sociolinguistic variation and discourse function of constructed dialogue introducers: the case of *Be* + *Like*. American Speech 70.265–90.

Guy, Gregory. 1980. Variation in the group and the individual: the case of final stop deletion. Locating language in time and space, ed. by William Labov, 1–36. New York, NY: Academic Press.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. Linguistics 36.161–95.

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.

——. 2007a. The North Carolina sociolinguistic archive and analysis project: empowering the sociolinguistic archive. Penn Working Papers in Linguistics 13.2.15–26.

——. 2007b. On the status of pause in sociolinguistics. Paper presented at the Linguistic Society of America 2007 Annual Meeting. Anaheim, CA.

Kendall, Tyler, and Amanda French. 2006. Digital audio archives, computer-enhanced transcripts, and new methods in sociolinguistic analysis. Paper presented at Digital Humanities (ALLC/ACH), Paris, France.

Kretzschmar, William, Jr, Jean Anderson, Joan Beal, Karen Corrigan, Lisa Lena Opas-Hänninen, and Bartlomiej Plichta. 2006. Collaboration on corpora for regional and social analysis. Journal of English Linguistics 34.172–205.

Labov, William. 1966. The social stratification of English in New York City. Washington, DC: Center for Applied Linguistics.

——. 1984. Field methods of the project on linguistic change and variation. Language in use: readings in sociolinguistics, ed. by John Baugh and Joel Sherzer, 28–53. Englewood Cliffs, NJ: Prentice-Hall.

Liddicoat, Anthony. 2007. An introduction to conversation analysis. London/New York, NY: Continuum.

Macaulay, Ronald. 1991. Locating dialect in discourse: the language of honest men and Bonnie Lasses in Ayr. New York, NY/Oxford, UK: Oxford University Press.

MacWhinney, Brian. 2007. The talkbank project. Creating and digitizing language corpora. Volume 1: synchronic databases, ed. by Joan Beal, Karen Corrigan and Hermann Moisl, 163–80. New York, NY/Basingstoke, UK: Palgrave-Macmillan.

Milroy, Lesley. 1987. Observing and analyzing natural language. Malden, MA/Oxford, UK: Blackwell.

Milroy, Lesley, and Matthew Gordon. 2003. Sociolinguistics: methods and interpretation. Malden, MA/Oxford, UK: Blackwell.

Mishler, Elliot. 1991. Representing discourse: the rhetoric of transcription. Journal of Narrative and Life History 1.255–80.

NINCH. 2003. The NINCH guide to good practice in the digital representation and management of cultural heritage materials, Version 1.1. [online resource]. <http://www.ninch.org/guide.pdf>.

Ochs, Elinor. 1979. Transcription as theory. Developmental pragmatics, ed. by Elinor Ochs and Bambi Schieffelin, 43–72. New York, NY: Academic Press.

Pitt, Mark, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication 45.89–95.

Plichta, Bartek and Mark Kornbluh. 2002. Digitizing speech recordings for archival purposes [online resource]. <http://www.historicalvoices.org/papers/audio_digitization.pdf>.

Poplack, Shana. 1989. The care and handling of a mega-corpus: the Ottawa-Hull French project.

Language change and variation, ed. by Ralph Fasold and Deborah Schiffrin, 411–51. Amsterdam, The Netherlands: John Benjamins.

——. 2007. Foreward. Creating and digitizing language corpora. Volume 1: synchronic databases, ed. by Joan Beal, Karen Corrigan and Hermann Moisl, ix–xiii. New York, NY/Basingstoke, UK: Palgrave-Macmillan.

Preston, Dennis. 1989. Perceptual dialectology: nonlinguists' views of areal linguistics. Dordrecht, The Netherlands: Foris Publications.

——. 2000. Mowr and mowr bayud spellin': confessions of a sociolinguist. Journal of Sociolinguistics 4.614–21.

Rickford, John, Arnetha Ball, Renée Blake, Raina Jackson, and Nomi Martin. 1999. Rappin' on the copula coffin: theoretical and methodological issues in the analysis of copula variation in African American Vernacular English. African American Vernacular English, ed. by John Rickford, 61–89. Malden, MA/Oxford, UK: Blackwell.

Rickford, John, and Christine Théberge-Rafal. 1999. Preterite *Had* + verb *–ed* in the narratives of African American preadolescents. African American Vernacular English, ed. by John Rickford, 34–60. Malden, MA/Oxford, UK: Blackwell.

Romaine, Suzanne, and Deborah Lange. 1991. The use of *Like* as a marker of reported speech and thought: a case of grammaticalization in progress. American Speech 66.227–79.

Sankoff, David, and Gillian Sankoff. 1973. Sample survey methods and computer-assisted analysis in the study of grammatical variation. Canadian languages in their social context, ed. by Regna Darnell, 7–64. Edmonton, Cananda: Linguistic Research.

Shuy, Roger, Walt Wolfram, and William Riley. 1968. Field techniques in an urban language study. Washington, DC: Center for Applied Linguistics.

Tagliamonte, Sali. 2006. Analysing sociolinguistic variation. Cambridge, UK: Cambridge University Press.

——. 2007. Representing real language: consistency, trade-offs and thinking ahead! Creating and digitizing language corpora. Volume 1: synchronic databases, ed. by Joan Beal, Karen Corrigan and Hermann Moisl, 205–40. New York, NY/Basingstoke, UK: Palgrave-Macmillan.

Tagliamonte, Sali, and Rachel Hudson. 1999. *Be Like* et al. beyond America: the quotative system in British and Canadian Youth. Journal of Sociolinguistics 3.147–72.

Trudgill, Peter. 1974. The social differentiation of English in Norwich. Cambridge, UK: Cambridge University Press.

Wolfram, Walt. 1969. A sociolinguistic description of Detroit Negro speech. Washington, DC: Center for Applied Linguistics.

——. 1993. Identifying and interpreting variables. American dialect research, ed. by Dennis Preston, 193–221. Amsterdam, The Netherlands: John Benjamins.

——. 2006. Variation in language: overview. Encyclopedia of languages and linguistics, 2nd edn, ed. by Keith Brown, 333–40. Oxford, UK: Elsevier.

Wolfram, Walt and Ralph Fasold. 1974. The study of social dialects in American English. Englewood Cliffs, NJ: Prentice-Hall.

Wolfram, Walt, Rebecca Childs, and Benjamin Torbert. 2000. Tracing English dialect history through consonant cluster reduction: comparative evidence from isolated dialects. Journal of Southern Linguistics 24.17–40.

