2 Data in the Study of Variation and Change

TYLER KENDALL

From its modern beginnings with the work of William Labov (1963, 1966), the sociolinguistic study of language variation and change has centered on the empirical investigation of naturalistic talk, primarily gathered in the field (see Feagin, this volume) and elicited through the sociolinguistic interview (Labov 1966, 1972). Other data types and methodologies have also been found to be useful for investigating language variation and change – such as the use of written records (Schneider, this volume), general public corpora (Bauer 2002), attitudinal data (Preston, this volume), and, increasingly, experimental laboratory-based work (e.g. Campbell-Kibler 2007) – but data obtained through conversational, sociolinguistic interviews remain the bread and butter of sociolinguistic research. Yet, what do we actually mean when we talk about "data" in the study of language variation and change? How does speech became data? Every act of analysis involves interpretation and abstraction and the process of going from actual, naturalistic talk in and of the community to the sort of quantitative data that sheds light on the nuanced "orderly heterogeneity" of language (Weinreich, Labov, and Herzog 1968) is often under-discussed in the literature, although with further consideration it is quite complex.

This chapter considers the nature of spoken language data and how it is treated in variationist research. Following somewhat from Kendall (2008),[1] I focus on the meta-theoretical questions of just what variationist data are and how modern computer-based techniques can enhance sociolinguists' connection to and use of their data. I also illustrate how the ways we conceptualize our data and interact with them impacts our analyses and our understanding of the very task of studying language variation and change.

The Handbook of Language Variation and Change, Second Edition.
Edited by J.K. Chambers and Natalie Schilling.
© 2013 John Wiley & Sons, Inc. Published 2013 by John Wiley & Sons, Inc.

1 A Brief History of Data in Sociolinguistics

Since its inception, sociolinguistics has been driven by an interest in natural, authentic language data and, over the history of the discipline, scholars have recorded a huge amount of speech. Some of the earliest modern projects alone, such as Wolfram's (1969) work on African American English in Detroit, collected many hundreds of hours of audio over the course of their fieldwork. The treatment of these recordings – from the more meta-theoretical question of how they have been conceptualized as the data upon which sociolinguistic descriptions and theories are made, to the more methodological issues of how they have been managed and preserved – has changed over the years. I here consider how sociolinguists have treated and described these data, their actual speech recordings, during the past 50 years.

Research reports in the early days of sociolinguistic variation studies (e.g. Labov 1966; Wolfram 1969; Sankoff and Sankoff 1973; Trudgill 1974) tended to publish 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 strategies and even of training their fieldworkers. 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).

Many of these early reports included detailed information about their recordings. For instance, Shuy, Wolfram, and Riley (1968) committed an entire volume to their field methods for the Detroit Dialect Study (Wolfram 1969). Through it, they provide a more thorough account of the treatment of that project's data and recordings than most other projects have, explaining and commenting upon a range of issues from determining a sampling system to designing the format of the files used for coding their data. Similarly, Sankoff and Sankoff (1973) provide a thorough overview of the field and laboratory methods for their sociolinguistic study of Montreal 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 printouts (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)

Much of the work following this first generation of sociolinguistic research did not address its data to such a detailed degree. It may be that the level of description provided by Sankoff and Sankoff became viewed as unnecessarily detailed. Yet, the earliest projects had a world of options surrounding them about what to analyze and even what to consider as their data for analysis. They had to be explicit about each step of their work as they abstracted from real-world speech events to filtered-down quantitative variable data. Detailed accounts of their data and research steps were needed. As these studies found success by yielding robust findings and indicating just how systematic language variation actually is, it became less important to dwell on many of the technical details. Just as fieldwork moved from rigorous, large-scale random sampling techniques (Labov 1966; Wolfram 1969) to network studies and "friend of a friend" sampling techniques (cf. Milroy and Gordon 2003) as it became clear that the systematicity of language variation was discoverable from smaller scale (and more locally sensitive) studies (cf. Eckert 2005), the fine attention to reporting each step of one's analysis process also became less important. In a sense, the data of variationist analysis often jumps from the actual recordings of conversational sociolinguistic interviews to spreadsheets of variable instances.

It is important also to appreciate that recordings during the first decades of modern sociolinguistics were expensive. Before the advent of lightweight tape recorders and now ultra-lightweight solid-state digital recorders, recording equipment was large, cumbersome, and costly. As recording technologies became more accessible via inexpensive and ubiquitous equipment and storage media, sociolinguists' discussions of their methods could focus less and less on the practicalities of recording and the details of the actual, physical recordings. A review of many popular sociolinguistic textbooks shows that their discussions of methods often skip 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, and how to analyze the resulting sociolinguistic variable data. However, almost across the board these discussions neglect issues in how the speech recordings should be organized, stored, preserved, and so on.

Of course, there are some exceptions in the literature. Poplack (1989) discussed the Ottawa-Hull French Project's data archive and methodology in detail, a project with a goal to improve methodologies inherent in working with large sets of data for sociolinguistic analysis. Her methodologically focused paper responded to the fact that:

> One area in which development has been sporadic at best is in the construction of major sociolinguistic databases. 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 database reaches respectable size, its accessibility is concurrently hampered by the uneconomical effort needed to search it systematically in studies of individual variables. (1989: 413)

Poplack's paper provides a thorough treatment 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.

Much recent work in sociolinguistics has returned to dealing explicitly and thoroughly with its data. Tagliamonte's (2006) textbook, *Analysing Sociolinguistic Variation*, 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. Schilling's (2013) book *Sociolinguistic Fieldwork* also discusses a range of important data management and preservation questions. Meanwhile, many funding agencies, such as the National Science Foundation in the US, have recently instituted policies about the management, preservation, and dissemination of data collected under funded research.[2] Likely these kinds of policies will make the explicit treatment of data a larger part of sociolinguistic research endeavors in the coming years. My own work (e.g. Kendall 2008, 2011) has attempted to explore the consequences of our relationships with our data, how the decisions we make – for example, when organizing our data, when transcribing, and so on – impact the kinds of questions that we can ask and the answers that we obtain. In Section 3 I consider this point more thoroughly, but first we consider the status of "corpora" in work on language variation and change.

2 Sociolinguistics, Corpora, and Data Sharing

Several publications consider the ways that standard, publically available corpora can be used to examine language variation and change (Bauer 2002; Baker 2010; Kendall 2011). I do not consider the sociolinguistic analysis of public, or "conventional" (Beal, Corrigan, and Moisl 2007a), corpora in this chapter but rather the fact that sociolinguists are increasingly discussing and thinking about their own data as corpora, a reconceptualization that has potential benefits for the variationist endeavor. Viewing our recording collections and data as corpora – as coherent, self-contained, representative samples of a language variety (see below for a fuller definition) – positions us to be more explicit about just what counts as the data used in a particular project, what is included, what is not, and how we (and others) access them. It also better supports the model of replicable research to which all scientific research should strive.

Corpus linguists primarily view corpus linguistics as a methodology rather than a theoretical stance (cf. McEnery and Wilson 2001; McEnery, Xiao, and Tono 2006; Gries 2009; Kendall 2011) and as such can offer complementary research methods and practices to the investigation of language variation and change. However, corpus linguists differentiate corpora proper from other, less systematically developed collections of language data. Corpora are often defined as involving an explicit focus on:

- *Representativeness* and *balance* – a corpus should accurately represent the full language variety it purports to contain; further, it should be balanced across the proportions of linguistic and social categories that comprise the variety.
- *Machine-readability* – a corpus should be machine-readable (which typically means electronic text).
- A particular (large) *size* – many descriptions of "what makes a corpus a corpus" do not explicitly argue for a size requirement, but in reading the corpus linguistic literature one notes a common focus on size measured in number of words and a growing interest in large corpora.

While corpus linguists pay a great deal of attention to the notions of representativeness and balance, these are often taken to be more of an ideal than a strict requirement. It is, of course, often not possible to represent a language variety precisely in a corpus. Gries writes:

> If I know that dialogs make up 65 percent of the speech of adolescent Californians, approximately 65 percent of my corpus [of adolescent Californian speech] should consist of dialogue recordings. This example already shows that this criterion is more of a theoretical ideal: How would one measure the proportion that dialogs make up of the speech of adolescent Californians? (2009: 8)

Even though corpus linguists accept that these criteria are difficult to meet in actuality, many sociolinguistic recording collections do not attempt to meet the sampling criteria or size to be considered "corpora" by many corpus linguists. (Instead, they might be termed "archives" or "databases" by these scholars.) But, terminology aside, thinking about sociolinguistic data in terms of corpora can benefit both the long- and short-term life of the data. For instance, thinking about sociolinguistic fieldwork as corpus creation (in the terms spelled out in, say, McEnery *et al.*'s 2006 introductory text about corpus-based language study) can lead to better-organized and more manageable data collections.

One lesson in particular to take from the corpus linguistic literature is that it is beneficial to build and organize data collections with the goal that an unfamiliar researcher could make sense of the data without you. It may be the case that you do not plan to share your recordings with anyone else (or that you are unable to, see below), but if you return to your data collection five or 10 years – or even six months – in the future you may find that you approach the data as would a total stranger, for example, no longer remembering how to interpret the file-naming conventions or directory structure.

One of the biggest hurdles to overcome in building sociolinguistic corpora may not actually be technical but rather about who will be able to access the recordings. In some cases, the question of whether or how to share the recordings and other data generated over the course of a research project and who to share with (direct collaborators, other researchers, the general public) is something that each researcher must consider for her- or himself. In other cases, as indicated above, funding agencies or other institutions might enforce a data-sharing or data-accessibility plan, or, as discussed below, human subjects concerns might prevent the sharing of data. Traditionally, sociolinguists have not made a common practice of sharing their data widely and an orientation towards private data versus public corpora is a persistent difference between sociolinguistic and corpus linguistic researchers.

Most sociolinguistic work falls into the category of human subjects research (at least in the US and UK) and is overseen by various kinds of ethics boards. In the US, each university or research institution has an Institutional Review Board (IRB) and each IRB may have slightly different standards for what is permissible – even with the consent of research participants. You should always check with your local IRB before beginning any fieldwork. Regardless of your participants' willingness, it can be the case that your local IRB will not allow you to share the data or that certain steps must be undertaken – such as redaction/anonymization – before the data can be shared (see Childs, Van Herk, and Thorburn 2011).

Corpus linguists deal primarily with text-based data. Documentary linguistics, and the endangered language research community, is another allied discipline, with which sociolinguists have not traditionally collaborated, but which can offer much expertise about the preservation and management of speech recordings. While the nature of their interviews and their analyses – often focusing on qualitative grammatical descriptions – are different than those of sociolinguists, researchers working on endangered languages have a great interest in the preservation of their recordings. They have also put the most focused effort into the development of annotation standards and methodological best practices of any community of researchers who record and analyze natural speech data. Their annotation work (cf. the Open Language Archives Community [OLAC] best-practice recommendations;[3] Simons, Bird, and Spanne 2008) is not always entirely relevant to sociolinguists, but much of the literature on audio preservation (e.g. Bird and Simons 2003) is quite helpful (see Kendall 2013). It is also notable that sociolinguists and endangered language researchers are beginning to collaborate in new and greater ways. For example, Stanford and Preston's (2009) edited volume on variation in indigenous minority languages collects a number of papers that approach lesser-studied languages through a variationist lens. This can lead to further insights into the principles of language variation and change and simultaneously better, more rigorous treatments of the collected data. As sociolinguists move from their traditional foci on languages like English and French to lesser-studied languages, the issues involved in developing metadata, in transcription and other annotation, and in variable coding – the foci of the next section – become all the greater.

3 What You Put In Impacts What You Get Out

Sociolinguistic fieldwork obtains some of the most authentic language data available within any of the branches of linguistics. However, what sociolinguists get out of their data is ultimately a function of how those data are treated in the research steps following fieldwork. In the process of going from recordings of actual speech to some sort of quantitative or qualitative data for analysis, numerous decisions must be made and each of these can have ramifications on the rest of the research project and on future – and possibly unexpected – uses of the data. How and where do you record and describe information about the speakers in your recordings and other metadata about the recording? ("Metadata" are data about the data – e.g. what are the properties of the recordings?) Further, do you transcribe the recordings? If so, do you transcribe phonetically or orthographically? What spelling conventions do you adopt and what kinds of variable features do you include? Do you focus on temporal accuracy (i.e. finely calibrated time-alignment) and/or orthographic accuracy? Many more questions are important as well. To review many of these issues further, we will consider a specific research example.

Say that we are interested in examining change in the quotative system – the use of *say*, *go*, *be like*, and so on, to introduce reported speech or thought – in a community of English speakers (cf. Romaine and Lange 1991; Ferrara and Bell 1995; Tagliamonte and D'Arcy 2004, 2007; Buchstaller 2006, 2011; Buchstaller and D'Arcy 2009). We would begin our study by conducting sociolinguistic interviews with a range of age and social groups in the community.[4] We might think that we now have our data and are ready to analyze it, but in fact a number of steps are intermediate before we can analyze the data. Even ignoring the crucial questions about how we define the variable context itself and undertake the variable coding (cf. Tagliamonte (2006) on variable analysis in general and Buchstaller (2011) on quotatives specifically), a number of questions remain about just what the data are. What is the scope over which one can define the variable in the first place? Is it the actual audio recordings themselves or a transcribed representation of the recordings? Put differently, do we extract the data directly from the audio? Or do we transcribe first and then extract variables from the transcripts? Or do we choose a middle route coding variables from the audio but using the transcripts as a guide to the audio? Finally, what specific information do we include with each variable code?

Transcription – the creation of a textual representation of the audio – is standard practice in some research traditions and by some research groups, but is rare or deemed unnecessary by others. This is often a function of the kind of features the researchers are interested in. Scholars working on phonological or phonetic variants may decide that transcription is too costly and time-consuming without high enough returns. Many scholars who work extensively on morphosyntactic or discourse features – like our quotative example – do make common practice of transcribing. In some cases, our preconception of whether we are examining a phonological or a morphosyntactic variable influences how we go about the variable coding, but then – and yes, this is circular – how the data are coded influences how we must analyze and interpret them.

To digress momentarily from our quotative example, variable (ing), the alternation between productions like *talking* and *talkin'*, is a widely studied variable in English (e.g. Fischer 1958; Labov 1966; Trudgill 1974; Tagliamonte 2004; Campbell-Kibler 2007; Hazen 2008). One could consider it to be a phonological alternation, a morphological alternation, a lexical alternation, or even a combination of these phenomena. The underlying conception of what the data are – for example, orthographic transcriptions versus phonetic transcriptions versus actual audio recordings – will ultimately play a role in how the variable is encountered and then studied by its researchers. If I work from orthographic transcripts (with the variable realizations of (ing) coded either phonetically or in the orthography – for example, using the spelling *talkin'* for the production [tɔkɪn]), I can conduct a rigorous analysis of many potential factors behind (ing) realization, even including some phonological factors (determining, for example, preceding and following environments from the transcribed text), but I could not examine the role that discourse-level stress patterns and other prosody plays and I may miss different phonetic realizations of (ing) such as variation in the vowel separate from the nasal segment.

Returning to the quotative example, while many studies of quotatives are conducted in similar ways – roughly by counting instances of the quotative markers (such as *say*, *go*, *be like*) and comparing the relative occurrences of each with other linguistic features of the discourse and social attributes of the speakers – a study could be undertaken in other ways.[5] This could depend not only on one's theoretical background and interest, but also on the format and type of data used for the analysis. 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 – from one who hears instances of *be like* in audio recordings, where phonological and phonetic aspects of the occurrences – such as variation in stop-release, intonation, or voice quality – may be more striking.

Further, let's say our interviews were conducted with pairs of participants who were friends or siblings or spouses. We might later become interested in the question of whether quotative use is sensitive to entrainment, that is, to the use by a speaker's interlocutor(s). Depending on how the original variable data were coded, this is either a very easy question to pursue or a very difficult one. If each coded variable contains information about when it was spoken, who it was spoken to, who else was present, and so on, we can readily compare usage across interlocutors and examine the data as properties of interactional events among multiple speakers rather than individual instances of language use divorced from their discourse context. If not, this would be impossible or would involve redoing the entire data coding.

The process of going from a recorded speech event to quantitative, variationist data, involves three types of derived (and interpretive) information. Each of these is a kind of data. These are (i) information about the speakers, the interactional context, and the recording itself – the *metadata*, (ii) intermediate *annotations*, such as transcripts and analysts' notes, and (iii) *extracted variables* and instrumental *measurements*. I comment on each of these in turn.

3.1 Metadata

As explained above, the term metadata means data about data. Typical kinds of metadata – such as the Dublin Core[6] standard used by many libraries – include information about when a particular resource (like an audio recording) was created, who created it, and with what, its format and technical specifications, and so forth. Yet, it is also valuable to think about speaker- and interaction-level information as metadata – not only when an interview was recorded, but who was present, who was miked, where it took place, and what was going on in the background. In addition, we ideally want to note various sociolinguistically relevant details for each speaker in a given recording (including interviewers as well as interviewees; see Schilling 2013, Chapter 6), including sex, ethnicity, age, year of birth, and so on.

One of the long-standing difficulties in developing metadata for sociolinguistic-like purposes is that different settings, communities, studies, and language/social groups have different properties. Ideally, metadata are objective "facts" about the data and are comparable across individual "units" of data. Many of the most interesting social categories to sociolinguists, however, are not facts about speakers that can be compared across projects but are locally relevant and interpretive (cf. Eckert 2005; see, for instance, Cheshire 2002 and Queen, this volume, for a relevant discussion of the difference between sex and gender). Even though we may not be able to come up with a set of metadata elements that will work for all studies and all sociolinguistic recording collections, we can still go a long way towards enumerating the technical, social, and interactional information about our data, and doing so – and storing it in easily retrievable ways – will benefit both short-term and long-term use of the recordings.

3.2 *Annotations*

Metadata tell us about the properties of a recording and the participants of the recording. They do not tell us about the actual linguistic (or interactional) content. Annotations provide a representation of and information about what is in the recordings. The most common type of annotation in linguistic research is transcription but other kinds of annotation, like part-of-speech tagging and syntactic parsing, are also common in some areas of work, especially in corpus and computational linguistics (again, see resources like McEnery *et al.* 2006).[7] Importantly, every kind of annotation – for example, an orthographic transcription or the part-of-speech tags for an utterance – is always a representation of the actual recording and the outcome of an interpretive process.

Let us consider transcription further. Both within and outside linguistics, the orthographic transcript is the primary representation used to present speech in a non-aural format. The act of transcription, especially by beginning transcribers, is often undertaken as a purely methodological activity, as if it were theory-neutral. Yet, each decision that is made while transcribing influences and constrains the resulting possible readings and analyses (Ochs 1979; Mishler 1991; Bucholtz 2000; Edwards 2001).[8] Decisions as seemingly straightforward as how to lay out the text and what spelling conventions to use, to those more nuanced – like how much nonverbal 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. It is important to remember that the text of a transcript is always an incomplete and interpreted record of the original interaction (Edwards 2001).

Thinking about other kinds of annotations, like part-of-speech tags, brings the interpretive nature of annotation more to the front. Is the word *waiting* in the phrase *the waiting man sat down* a verb, a gerund, an adjective, or . . . ? Ultimately, the answer is dependent on the analyst's theoretical framework and the tag-set being used. Deciding with what grammatical category to code the word is a form of analysis.

3.3 Extracted variables and measurements

At a certain point we come to what we often think of most squarely as "data" – the specific measurements and pieces of information that are submitted to our quantitative analyses. The principal data analyzed in variationist research are sociolinguistic variables (see Bayley, this volume, for a discussion of the quantitative paradigm) and instrumental measurements (see Thomas, this volume) extracted from speech recordings. But, while variationists may agree in principle on what these data are, there are no agreed upon processes for how to move from a collection of recordings 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 Tagliamonte 2006; Wolfram 2006), as well as problems of inter-analyst agreement and potentially intra-analyst reliability.

Blake (1997) and Rickford *et al.* (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.*, 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 coding issues that go undiscussed in the literature. This is not to argue that variation analysis is impossibly difficult, or that some researchers do it "wrong," but rather a reminder that the process of moving from recordings to quantitative data involves an array of decisions that must be made, and it is important to be thorough and explicit about each decision.

If we step back for a moment we see that all of the issues discussed here – the various layers of social, interactional, and recording-based metadata, the transcripts we generate, the extracted and coded variables – are in actuality not all that different from one another. Each reflects the output of a kind of analysis – some more subjective and interpretive, some more objective, some "harder" and some "easier," and so on – but each is just a different kind of information that we might want to use as we conduct our analyses. Each tells us something different about the recordings. And none is theory-neutral. If we are explicit about each of our data processing maneuvers and if we can develop better ways to organize all of these types of data, we can simultaneously make our analyses easier, by decreasing the amount of manual data entry and re-entry work that we do, and better, by decreasing the room for errors to creep into our workflow and by making our work more transparent and replicable.

4 Processing and Storing Speech Data

Recent years have seen a rising interest in the development of models and tools for the storage and management of speech data. Projects outlined in the two volumes on "unconventional" corpora edited by Beal, Corrigan, and Moisl (2007b, 2007c), like the Origins of New Zealand English (ONZE; Gordon, Maclagan, and Hay 2007) and the Newcastle Electronic Corpus of Tyneside English (NECTE; Allen *et al*. 2007), as well as projects described elsewhere, such as the Bergen Corpus of London Teenage Language (COLT; Stenström, Andersen, and Hasund 2002) and LANCHART (LAnguage CHAnge in Real Time; Gregersen 2009) are actively bridging the gap between corpus linguistic projects and sociolinguistic field-based research and developing innovative, new approaches to the treatment of (socio)linguistic data. Yet, best practices are still being developed and there is much room for advancements to be made. In a 2006 paper, Kretzschmar *et al*. argued for the need for broad collaboration among sociolinguists in the creation of sociolinguistic corpora and best practices for sociolinguistic corpus development. As the above-mentioned projects indicate, many sociolinguists are indeed thinking along these lines now, although to date the work continues to progress primarily via individual initiatives.

My own work, first (and ongoing) on the Sociolinguistic Archive and Analysis Project (SLAAP;[9] Kendall 2007, 2008) and more recently on the Online Speech/Corpora Archive and Analysis Resource (OSCAAR;[10] Kendall 2010), has sought to explore the possibilities of specialized, web-based spoken language archives. These projects provide *speech data management systems* in which disparate collections of recordings are stored, managed, and accessed. Poplack (2007), in a short foreword to the Beal *et al*. volumes, explains that the architecture of a sociolinguistic corpus is dependent on whether it is viewed by its creators as a *tool* or as an *end-product*. Unlike many variety- or community-oriented corpus-building projects, like ONZE, NECTE, COLT and LANCHART, SLAAP and OSCAAR are about the creation of tools rather than specific corpus end-products. The projects are outlined in detail elsewhere (Kendall 2007, 2008, 2010; see also Newman 2008), with screenshots and discussions of aspects like the transcript model and conventions, and for sake of space I do not review them in depth here. Rather, I focus on two of their main features in order to describe how computer-enhanced methods for the storage and management of sociolinguistic data can help our research practices.

4.1 Centralized archives and web accessibility

One benefit of projects like SLAAP and OSCAAR is that they store all of a research group's recordings, metadata, transcripts and annotations, variable data, and other derived materials (including articles, conference handouts, PowerPoint files, etc.) in a single, centralized location. This allows the complete data archive to exist separately from researchers' individual computers and ensures that all of a project's data remain accessible to all of their users. Even for recording collections that have only one owner or user (say, recordings collected for a student's Masters project for which there are no rights to share or reuse the data later), storing the recordings in a central archive still has benefits, such as ensuring that the data survive unforeseeable accidents, like spilled coffee on a laptop.

In the case of SLAAP, the archive is accessible from anywhere in the world, but is housed on a designated computer in the North Carolina State University Libraries server room. SLAAP is custom-built (see, for example, Kendall 2007) but other software is available from university information technology divisions and from private companies that can help in the management of audio data. My own university, for instance, makes available a secure file sharing system for faculty and research use that makes it easy to store (and back up) files and share them among a research group, and my university library provides a web-based publication system for faculty and student use which supports the long-term distribution and preservation of audio and video materials, in addition to more traditional file types, like PhD dissertations.[11] To some degree, even making the move from storing research data on a personal laptop or desktop computer to a centralized, networked drive has preservational and organizational benefits. However, many systems for the generic storage and sharing of files have pitfalls that must be kept in mind as well. For instance, unless a rigorous file access permission system is set up on a networked hard drive, it is easy for a user to accidentally rename, move, alter, or – worst of all – irrevocably delete files on the drive. Bird and Simons (2003) review a number of potential problems that arise from linguists' move from analog to digital files; the fact that digital files are so much "easier" to store and manage perhaps creates as many potential problems for their long-term preservation as it solves (see also Kendall 2013; Schilling 2013, Chapter 6).

SLAAP and OSCAAR, as specially designed speech data management systems, are built to address some of these basic issues. They provide highly structured access control systems so that different users of the systems have access to different sets of recordings and have access to different sets of features for those recordings. One user, for instance, may be registered as the "owner" or "manager" of a recording collection and can make changes to the collection (and, for example, can add new transcripts to the collection), while other user accounts have "read-only" access to those recordings.

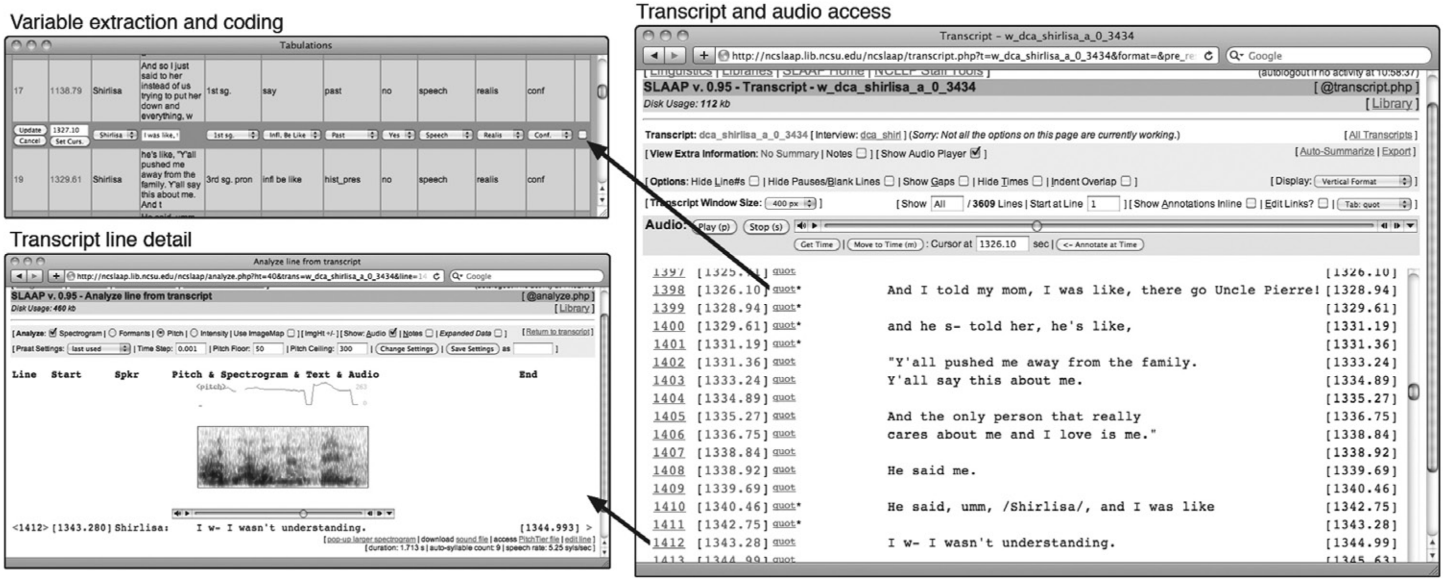
Centralizing all of one’s collected recordings makes it easier to reuse and reexamine older data. Returning to our quotative analysis example, for instance, we could imagine the benefits of having easy access to the files from earlier, related projects. We might, perhaps, want to test whether an unexpected pattern that arises in our own data exists in our colleague’s data from an earlier project. If the data are structured in similar ways and stored in the same place we can quickly test this idea. If not, it could take days of hunting around for the right files and more days to reformat them. Or, we might never find the necessary files.

4.2 Linked layers of data

SLAAP attempts to resolve some of the issues illustrated by our hypothetical quotative analysis by maintaining links between levels of annotation in the data. Each transcript line is accurately time-aligned and linked to its source audio and each extracted variable code is also stored with a time-stamp. Thus, through the web-based software, researchers can listen to the source audio at a click of the mouse and can jump between the different kinds of annotation. If we are interested in quotatives, we can code the instances of the variable directly from the transcript as we might in a “traditional” analysis, but also simultaneously listen to the audio to assess prosodic or voice quality factors that we would not be able to get from the transcript alone. Figure 2.1 displays some sample screens from SLAAP. The screen on the right is a transcript from an interview with a young African-American woman. The image to the upper-left is a (small area of a) variable coding window, where a researcher can extract and code an instance of a quotative. The image to the lower-left shows a close-up of a single transcript line, with a spectrogram and pitch information extracted and displayed. Both views of transcript information include audio players.

SLAAP and OSCAAR, the projects I have been involved in, are only two possible approaches to speech data management and they represent explorations and attempts at “proofs-of-concept” more than definitive solutions for sociolinguistic data management. Other systems are being developed – like LaBB-CAT[12] (formerly called ONZE Miner; Fromont and Hay 2008) – and a number of research groups are building sophisticated data management and dissemination systems (e.g. some of the projects described in Beal *et al*. 2007b, 2007c).

While I have thus far focused on data management systems for sociolinguistic data, it is also the case that general technological advances in natural language processing, acoustic signal processing, and wider computational methods are changing the ways that we examine our data. Easy-to-use, free software, like Praat[13] and ELAN,[14] make analysis and annotation processes, which once required expensive specialized equipment, possible for near-beginners with only a personal computer. Methods like the forced alignment of plain orthographic text to audio speech recordings (Yuan and Liberman 2008)[15] and the automatic extraction of vowel formant information (Evanini, Isard, and Liberman 2009; Labov, Rosenfelder, and Fruehwald 2013) point to new possibilities for the large-scale analysis of language variation. There is no doubt that the coming years will see an explosion of



Variable extraction and coding

Tabulations

17	1138.79	Shirlisa	And so I just said to her instead of us trying to put her down and everything, w	1st sg.	say	past	no	speech	realis	conf	
Update Cancel	1327.10 Set Curs.	Shirlisa	I was like, '	1st sg.	Infl. Be Like	Past	Yes	Speech	Realis	Conf.	□
19	1329.61	Shirlisa	he's like, "Y'all pushed me away from the family. Y'all say this about me. And t	3rd sg. pron	infl be like	hist_pres	no	speech	realis	conf	

Transcript line detail

Analyze line from transcript

http://ncslaap.lib.ncsu.edu/ncslaap/analyze.php?ht=40&trans=w_dca_shirlisa_a_0_3434&line=14

SLAAP v. 0.95 - Analyze line from transcript [@analyze.php]
Disk Usage: 460 kb [Library]

[Analyze: ☑ Spectrogram | ○ Formants | ● Pitch | ○ Intensity | Use ImageMap □] [ImgHt +/-] [Show: Audio ☑ | Notes □ | Expanded Data □] [Return to transcript]
[Praat Settings: last used | Time Step: 0.001 | Pitch Floor: 50 | Pitch Ceiling: 300 | Change Settings | Save Settings as]

```
Line   Start      Spkr        Pitch & Spectrogram & Text & Audio              End

<1412> [1343.280] Shirlisa:    I w- I wasn't understanding.               [1344.993] >
```

[pop-up larger spectrogram | download sound file | access PitchTier file | edit line]
[duration: 1.713 s | auto-syllable count: 9 | speech rate: 5.25 syls/sec]

Transcript and audio access

Transcript – w_dca_shirlisa_a_0_3434

http://ncslaap.lib.ncsu.edu/ncslaap/transcript.php?t=w_dca_shirlisa_a_0_3434&format=&pre_re

SLAAP v. 0.95 - Transcript - w_dca_shirlisa_a_0_3434 [@transcript.php]
Disk Usage: 112 kb [Library]

Transcript: dca_shirlisa_a_0_3434 [Interview: dca_shirl] (Sorry: Not all the options on this page are currently working.) [All Transcripts]
[View Extra Information: No Summary | Notes □] [Show Audio Player ☑] [Auto-Summarize | Export]
[Options: Hide Line#s □ | Hide Pauses/Blank Lines □ | Show Gaps □ | Hide Times □ | Indent Overlap □] [Display: Vertical Format]
[Transcript Window Size: 400 px] [Show All / 3609 Lines | Start at Line 1] [Show Annotations Inline □ | Edit Links? □ | Tab: quot]
Audio: Play (p) Stop (s)
Get Time | Move to Time (m) : Cursor at 1326.10 sec | <- Annotate at Time

```
1398  [1326.10] quot*     And I told my mom, I was like, there go Uncle Pierre! [1328.94]
1399  [1328.94] quot*                                                           [1329.61]
1400  [1329.61] quot*     and he s- told her, he's like,                        [1331.19]
1401  [1331.19] quot*                                                           [1331.36]
1402  [1331.36] quot      "Y'all pushed me away from the family.                [1333.24]
1403  [1333.24] quot      Y'all say this about me.                              [1334.89]
1404  [1334.89] quot                                                            [1335.27]
1405  [1335.27] quot      And the only person that really                       [1336.75]
1406  [1336.75] quot      cares about me and I love is me."                     [1338.84]
1407  [1338.84] quot                                                            [1338.92]
1408  [1338.92] quot      He said me.                                           [1339.69]
1409  [1339.69] quot                                                            [1340.46]
1410  [1340.46] quot*     He said, umm, /Shirlisa/, and I was like              [1342.75]
1411  [1342.75] quot*                                                           [1343.28]
1412  [1343.28] quot      I w- I wasn't understanding.                          [1344.99]
```

Figure 2.1 Screenshots from SLAAP.

computationally sophisticated analysis, annotation, and data dissemination techniques. The challenge for sociolinguists in the years to come will be to continue to lead in the collection, analysis, and interpretation of authentic, richly contextualized language data while not becoming isolated from broader advances in the technical and computational treatment of speech data.

5 Moving Ahead

As I pointed out earlier, sociolinguists have collected a huge amount of data over the past half-century. We have also developed sophisticated theoretical and methodological toolkits for the analysis and understanding of our data. We have not – as a field – developed rigorous, shared models of data and this has hampered some areas of potential progress. Across linguistic disciplines, researchers are increasingly turning to corpora of authentic language in use (cf. Wasow 2009) and sociolinguists more than any other language research community could be contributing to the availability of language data. More importantly for sociolinguists, thinking seriously about the nature of speech data will allow us to get more out of them. Twenty years ago, Macaulay wrote:

> 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. (1991: 5)

Macaulay was writing about the sociolinguistic interview and was commenting specifically about how individual interviews are used as data. Out of context we can take his words even more seriously – sociolinguists have a wealth of rich, authentic data available to us but typically use only a very small percentage of that data. Moving ahead we must develop richer models and methods for dealing with all of this data without losing sight of the ultimate goal, to understand the sources, processes, and outcomes of language variation and change.

NOTES

[1] Sections of this chapter revisit discussions published in Kendall (2008) and Kendall (2011).

[2] See http://www.nsf.gov/sbe/sbe_data_management_plan.jsp for information about the NSF's Data Management Plan requirements.

[3] http://www.language-archives.org/REC/bpr-20080711.html

[4] We might want to craft interviews in such a way as to elicit particular kinds of talk that provide the right environments for quoted discourse. I leave further consideration of the fieldwork strategies for our hypothetical research project, however, to Feagin, this volume.

5 For instance, contrast a study like Klewitz and Couper-Kuhlen's (1999) analysis of the prosodic marking of reported speech with the variationist studies of quotatives cited earlier.

6 See www.dublincore.org

7 In addition to orthographic and phonetic transcripts (in, for example, the International Phonetic Alphabet), linguists also talk about annotation systems like ToBI as "transcription" systems (cf. Beckman, Hirschberg, and Shattuck-Hufnagel 2005), which is to say that *transcription* as a term is often fairly synonymous with *annotation*. It appears to be mostly convention that separates what is called "transcription" and what is called "annotation".

8 Also see Preston (1982).

9 ncslaap.lib.ncsu.edu

10 oscaar.ling.northwestern.edu

11 See scholarsbank.uoregon.edu

12 onzeminer.sourceforge.net

13 http://www.fon.hum.uva.nl/praat

14 http://www.lat-mpi.eu/tools/elan

15 For instance, see the Penn Phonetics Lab Forced Aligner at http://www.ling.upenn.edu/phonetics/p2fa

REFERENCES

Allen, Will, Beal, Joan, Corrigan, Karen *et al*. (2007) A linguistic "time-capsule": The Newcastle Electronic Corpus of Tyneside English. In Joan Beal, Karen Corrigan, and Hermann Moisl (eds.), *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases*. 16–48. New York/Basingstoke, Hampshire: Palgrave Macmillan.

Baker, Paul (2010) *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Bauer, Laurie (2002) Inferring variation and change from public corpora. In J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. 97–114. Oxford: Blackwell.

Beal, Joan, Corrigan, Karen, and Moisl, Hermann (2007a) Taming digital voices and texts: Models and methods for handling unconventional synchronic corpora. In Joan Beal, Karen Corrigan, and Hermann Moisl (eds.), *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*. 1–16. New York/Basingstoke, Hampshire: Palgrave Macmillan.

Beal, Joan, Corrigan, Karen, and Moisl, Hermann (eds.) (2007b) *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*. New York/Basingstoke, Hampshire: Palgrave Macmillan.

Beal, Joan, Corrigan, Karen, and Moisl, Hermann (eds.) (2007c) *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases*. New York/Basingstoke, Hampshire: Palgrave Macmillan.

Beckman, Mary E., Hirschberg, Julia, and Shattuck-Hufnagel, Stefanie (2005) The original ToBI system and the evolution of the ToBI framework. In Sun-Ah Jun (ed.), *Prosodic Models, and Transcription: Towards Prosodic Typology*. 9–54. Oxford: Oxford University Press.

Bird, Steven and Simons, Gary (2003) Seven dimensions of portability for language documentation and description. *Language* 79(3): 557–582.

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(1): 57–79.

Bucholtz, Mary (2000) The politics of transcription. *Journal of Pragmatics* 32: 1439–1465.

Buchstaller, Isabelle (2006) Diagnostics of age-graded linguistic behaviour: The case of the quotative system. *Journal of Sociolinguistics* 10(1): 3–30.

Buchstaller, Isabelle (2011) Quotations across the generations: A multivariate analysis of speech and thought introducers across 5 decades of Tyneside speech. *Corpus Linguistics and Linguistic Theory* 7(1): 59–92.

Buchstaller, Isabelle and D'Arcy, Alexandra (2009) Localized globalization: A multi-local, multivariate investigation of quotative *be like*. *Journal of Sociolinguistics* 13(3): 291–331.

Campbell-Kibler, Kathryn (2007) Accent, (ING), and the social logic of listener perceptions. *American Speech* 82(1): 32–64.

Cheshire, Jenny (2002) Sex and gender in variationist research. In J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. 423–443. Oxford: Blackwell.

Childs, Becky, Van Herk, Gerard, and Thorburn, Jennifer (2011) Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory* 7(1): 163–180.

Eckert, Penelope (2005) Three waves of variation study: The emergence of meaning in the study of variation. http://www.stanford.edu/~eckert/PDF/ThreeWavesofVariation.pdf (accessed February 5, 2013).

Edwards, Jane (2001) The transcription of discourse. In Deborah Tannen, Deborah Schiffrin, and Heidi Hamilton (eds.), *Handbook of Discourse Analysis*. 321–348. Oxford: Blackwell.

Evanini, Keelan, Isard, Stephen, and Liberman, Mark (2009) Automatic formant extraction for sociolinguistic analysis of large corpora. *Proceedings of INTERSPEECH 2009*. 1655–1658.

Ferrara, Kathleen and Bell, Barbara (1995) Sociolinguistic variation and discourse function of constructed dialogue introducers: The case of *be + like*. *American Speech* 70: 265–290.

Fischer, John L. (1958) Social influence on the choice of a linguistic variant. *Word* 14: 47–56.

Fromont, Robert and Hay, Jennifer (2008) ONZE Miner: The development of a browser-based research tool. *Corpora* 3(2): 173–193.

Gordon, Elizabeth, Maclagan, Margaret, and Hay, Jennifer (2007) The ONZE Corpus. In Joan Beal, Karen Corrigan, and Hermann Moisl (eds.), *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases*. 82–104. New York/Basingstoke, Hampshire: Palgrave Macmillan.

Gregersen, Frans (2009) The data and design of the LANCHART study. *Acta Linguistica Hafniensia* 41: 3–29.

Gries, Stefan Th. (2009) *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York/London: Routledge.

Hazen, Kirk (2008) (ING): A vernacular baseline for English in Appalachia. *American Speech* 83(2): 116–140.

Kendall, Tyler (2007) Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. *Penn Working Papers in Linguistics* 13(2): 15–26.

Kendall, Tyler (2008) On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2(2): 332–351.

Kendall, Tyler (2010) Developing web interfaces to spoken language data collections. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2) University of Chicago.

Kendall, Tyler (2011) Corpora from a sociolinguistic perspective (Corpora sob uma perspectiva sociolinguística) In Stefan Th. Gries (ed.), Corpus studies: Future directions, special issue of *Revista Brasileira de Linguística Aplicada* 11(2): 361–389.

Kendall, Tyler (2013) Data preservation and access. In Christine Mallinson, Becky Childs, and Gerard Van Herk (eds.), *Data Collection in Sociolinguistics: Methods and Applications*. 195–205. New York/London: Routledge.

Klewitz, Gabriele and Couper-Kuhlen, Elizabeth (1999) Quote-unquote? The role of prosody in the contextualization of reported speech sequences. *Pragmatics* 9(4): 459–485.

Kretzschmar, William Jr., Anderson, Jean, Beal, Joan *et al*. (2006) Collaboration on corpora for regional and social analysis. *Journal of English Linguistics* 34(3): 172–205.

Labov, William (1963) The social motivation of a sound change. *Word* 19: 273–309.

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

Labov, William (1972) *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William, Rosenfelder, Ingrid, and Fruehwald, Josef (2013) One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89(1): 30–65.

Macaulay, Ronald (1991) *Locating Dialect in Discourse: The Language of Honest Men and Bonnie Lasses in Ayr*. New York/Oxford: Oxford University Press.

McEnery, Tony and Wilson, Andrew (2001) *Corpus Linguistics*. Second Edition. Edinburgh: Edinburgh University Press.

McEnery, Tony, Xiao, Richard, and Tono, Yukio (2006) *Corpus-based Language Studies: An Advanced Resource Book*. New York/London: Routledge.

Milroy, Lesley (1987) *Observing and Analyzing Natural Language*. Oxford: Blackwell.

Milroy, Lesley and Gordon, Matthew (2003) *Sociolinguistics: Methods and Interpretation*. Oxford: Blackwell.

Mishler, Elliot (1991) Representing discourse: The rhetoric of transcription. *Journal of Narrative and Life History* 1(4) 255–280.

Newman, John (2008) Spoken corpora: Rationale and application. *Taiwan Journal of Linguistics* 6(2): 27–58.

Ochs, Elinor (1979) Transcription as theory. In Elinor Ochs and Bambi Schieffelin (eds.), *Developmental Pragmatics*. 43–72. New York: Academic Press.

Poplack, Shana (1989) The care and handling of a mega-corpus: The Ottawa-Hull French Project. In Ralph Fasold and Deborah Schiffrin (eds.), *Language Change and Variation*. 411–451. Amsterdam: John Benjamins.

Poplack, Shana (2007) Foreword. In Joan Beal, Karen Corrigan, and Hermann Moisl (eds.), *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*. ix–xiii. New York/Basingstoke, Hampshire: Palgrave Macmillan.

Preston, Dennis (1982) "Ritin" fowklower daun 'rong: Folklorists failures in phonology. *Journal of American Folklore* 95(377): 304–326.

Rickford, John, Ball, Arnetha, Blake, Renée *et al*. (1999) Rappin' on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African American Vernacular English. In John Rickford, *African American Vernacular English*. 61–89. Oxford: Blackwell.

Romaine, Suzanne and Lange, Deborah (1991) The use of *like* as a marker of reported speech and thought: A case of grammaticalization in progress. *American Speech* 66(3): 227–279.

Sankoff, David and Sankoff, Gillian (1973) Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Regna Darnell (ed.), *Canadian Languages in their Social Context*. 7–64. Edmonton, Canada: Linguistic Research.

Schilling, Natalie (2013) *Sociolinguistic Fieldwork*. Cambridge: Cambridge University Press.

Shuy, Roger, Wolfram, Walt, and Riley, William (1968) *Field Techniques in an Urban Language Study*. Washington, DC: Center for Applied Linguistics.

Simons, Gary, Bird, Steven, and Spanne, Joan (eds.) (2008) Best practice recommendations for language resource description. Open Language Archives Community document.

Stanford, James N. and Preston, Dennis R. (eds.) (2009) *Variation in Indigenous Minority Languages*. Amsterdam: John Benjamins.

Stenstöm, Anna-Brita, Andersen, Gisle, and Hasund, Ingrid Kristine (2002) *Trends in Teenage Talk: Corpus Compilation, Analysis, and Findings*. Amsterdam: John Benjamins.

Tagliamonte, Sali (2004) Somethi[ŋ]'s goi[n] on!: Variable (ing) at ground zero. In Britt-Louise Gunnarsson, Lena Bergström, Gerd Eklund *et al*. (eds.), *Language Variation in Europe: Papers from ICLaVE 2*. 390–403. Uppsala, Sweden: Uppsala Universitet.

Tagliamonte, Sali (2006) *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Tagliamonte, Sali and D'Arcy, Alexandra (2004) He's like, she's like: The quotative system in Canadian youth. *Journal of Sociolinguistics* 8(4): 493–514.

Tagliamonte, Sali and D'Arcy, Alexandra (2007) Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change* 19: 199–217.

Trudgill, Peter (1974) *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Wasow, Thomas (2009) Gradient data and gradient grammars. *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society*. 255–271.

Weinreich, Uriel, Labov, William, and Herzog, Marvin (1968) Empirical foundations for a theory of language change. In Winfred P. Lehmann and Yakov Malkiel (eds.), *Directions for Historical Linguistics*. 95–188. Austin, TX: University of Texas.

Wolfram, Walt (1969) *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics.

Wolfram, Walt (1993) Identifying and interpreting variables. In Dennis Preston (ed.), *American Dialect Research*. 193–221. Amsterdam: John Benjamins.

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

Wolfram, Walt and Fasold, Ralph (1974) *The Study of Social Dialects in American English*. Englewood Cliffs, NJ: Prentice-Hall.

Yuan, Jiahong and Liberman, Mark (2008) Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics 2008*: 5687–5690.