12 Data Preservation and Access

Tyler Kendall

Sociolinguistic research creates a huge amount of data of various kinds, from recordings to derived transcripts and spreadsheets of coded variables and other measurements and materials such as demographic information about speakers and ethnographic notes. Traditionally, we sociolinguists have tended not to be very explicit about what we do with these data over the course of and beyond our research projects. Recently, however, issues of data sharing, management, and preservation have become important and common topics of discussion among sociolinguists. Recent articles (e.g., Kendall, 2008; 2011; Kretzschmar et al., 2006), edited volumes (e.g., Beal, Corrigan, & Moisl, 2007a; 2007b; Kendall & Van Herk, 2011), and conference presentations and workshops (e.g., Buchstaller, Corrigan, Mearns, & Moisl, 2011; Coleman, Hall-Lew, & Temple, 2011) have addressed issues of managing data, the compilation of sociolinguistic data into "corpora," and data preservation and access.

Data generated through sociolinguistic fieldwork and other forms of sociolinguistic data collection, such as experimentation and corpus aggregation, are valuable and can be of use for a range of investigations unforeseen during the original project for which they are created. Sociolinguistic recordings in particular can provide a wealth of data of interest not only for future sociolinguistic purposes but also for other linguistic studies, oral history research, and public outreach. For sociolinguists, the existence and availability of older recordings has enabled real-time studies of language change to an impressive time-depth – as in projects like the Origins of New Zealand English (ONZE: Gordon, Maclagan, & Hay, 2007) and LANCHART (LANguage CHange in Real Time: Gregersen, 2009). ONZE, for instance, traces the English language in New Zealand back to its first English-speaking settlers, thanks to the availability of recordings made by non-linguists in the 1940s (Gordon et al., 2007). These kinds of projects become possible only if recordings and the information about them (e.g., who the speakers are) are preserved and kept accessible. The success of recent real-time projects may seem to indicate that this is a trivial issue by suggesting that large amounts of data are available to those who look for them, but in fact, upon closer inspection, the majority of speech recordings, sociolinguistic and otherwise, appear to get lost over time. The issue is not necessarily their untimely destruction or lack of preservation but rather their lack of accessibility and/or discoverability.

In this chapter, I review many of the issues involved in preserving and maintaining access to sociolinguistic data. Many of these concerns are intimately tied to topics explored elsewhere, such as research ethics, particularly confidentiality and anonymity (see Trechter, Chapter 3; Besnier, Vignette 3a; Mann, Vignette 3b; Ehrlich, Vignette 3c; and Sadler, Vignette 3d), technical challenges in data collection (see De Decker & Nycz, Chapter 7, and Hall-Lew & Plichta, Vignette 7a), methods for transcription and annotation (see D'Arcy, Vignette 11c, and Vigouroux, Vignette 11d); discussion of making sociolinguistic data accessible to the public is also provided by Kretzschmar (Vignette 12a). In this chapter, I focus on data storage and management, only touching on issues of ethics and rights management. Much of my discussion draws heavily from the language documentation and description literature (e.g., Austin, 2006; Bird & Simons, 2003), where the details of data preservation and access have been considered to great depth.

Legacy Materials

In discussing the preservation of sociolinguistic recordings, it seems fitting to begin with the fact that many legacy materials have proven invaluable to sociolinguistic study. Some of these were created for non-sociolinguistic purposes but have been crucial in developing larger sociolinguistic pictures of varieties and phenomena. The "Mobile Unit" recordings created by Radio New Zealand in the 1940s are a central part of the larger ONZE corpus and have greatly extended the time-depth of the ONZE project's examinations into the origins of New Zealand English (Gordon et al., 2007). As a second example, in North America the existence of recordings made with ex-slaves in the early 20th century has enabled deeper insights into the origins and early forms of African American English (Bailey, Maynor, & Cukor-Avila, 1991).

As time passes, "legacy" materials have come to include recordings created directly by sociolinguists, such as the "S1" studies in the LANCHART project's collection of Danish materials. These "S1" studies are sets of interviews collected from six sites in Denmark between 1973 and the 1990s, with most of the recordings made in the mid-1980s (Gregersen, 2009). The availability of these recordings motivated the creation of the LANCHART project, which aggregated these older data and then resampled many of the same participants in "S2" studies in the early 2000s. By having access to the original recordings – but also complete descriptions of the design of the "S1" studies and the actual data from and information about the original informants – LANCHART has been able to build on the original studies to conduct an unprecedented panel survey (cf. Bailey, 2002) for investigating language change in real-time (Gregersen & Barner-Rasmussen, 2011). These kinds of projects represent important directions for the study of language variation and change, and they become possible only through the availability of legacy recordings and earlier primary sociolinguistic research materials.

Digitizing Analog Recordings and Preserving Digital Recordings

Until recently, much audio and video recording was done on analog devices, such as cassette tape, reel-to-reel tape, and even cylinder- and disc-based phonograph devices. The past couple of decades have seen massive initiatives in the digitization – the transfer from analog to digital format – of these legacy materials. (I do not detail the process of digitization, which in sum involves playing the analog recording using an appropriate device and sending its output directly to a digital recorder, such as computer-based recording software. Numerous websites and documents provide detailed descriptions of the digitization process and how one can achieve the best-quality results. For a good linguistically oriented presentation, see Bartek Plichta's website, http://bartus.org/akustyk/adc.html.) These initiatives have come from a diverse range of groups beyond linguists and other academics, including government agencies (e.g., through various initiatives and grant opportunities by federal and local funders) and libraries (e.g., the University of North Carolina Library's Documenting the American South project, http://docsouth.unc.edu/sohp/).

Solid-state digital recorders have become the recording device of choice for many sociolinguistic fieldworkers (see De Decker & Nycz, Chapter 7, and Hall-Lew & Plichta, Vignette 7a), and most sociolinguistic recordings are now completely digital. Thus, it is increasingly the case that most of the available audio recordings of speech, new and old, are available in digital versions. Thanks to the internet, digital files can easily be duplicated and even potentially shared with, accessed, and discovered by new users.

At first glance, digital recordings, and their ability to be duplicated cheaply and easily, may seem to "solve" the problem of ensuring that materials stay preserved and accessible over the passage of time, but the long-term preservation of these digital resources actually involves a host of problems. Bird and Simons (2003) provide a thorough consideration of what they term the portability problem, arguing that the problem of data preservation is a part of a larger issue in data management (i.e., portability):

> If digital language documentation and description should transcend time, they should also be reusable in other respects: across different software and hardware platforms, across different scholarly communities (e.g. field linguistics, language pedagogy, language technology), and across different purposes (e.g. research, teaching, development).
>
> (p. 558)

They highlight seven problem areas for data portability – content, format, discovery, access, citation, preservation, and rights – and propose a series of recommended best practices, which they arrive at through the articulation of value statements about their research community's needs. Readers are referred to Bird and Simons' paper and follow-up papers (such as Boas's 2006 discussion of implementing their recommendations) for detailed discussions. For sake of

space, I do not review Bird and Simons' seven problem areas completely but instead consider four broader points that come out of their proposals:

1. *Digital formats and digital media have relatively short lifetimes and thus require active curation to survive the passage of time.* Bird and Simons report that

   > [m]uch digital language documentation and description becomes inaccessible within a decade of its creation. Linguists who have been quick to embrace new technologies, create digital materials, and publish them on the web soon find themselves in technological quicksand. Funded documentation projects are usually tied to software versions, file formats, and system configurations having a lifespan of three to five years. Once this infrastructure is no longer tended, the language documentation is quickly mired in obsolete technology.
   >
   > (p. 557)

   In other words, ensuring that materials survive the passage of time is a larger project – and takes a larger commitment – than simply posting copies of files to a website or backed up hard drive. (I return to this point in the next section.)
2. *The use of open – i.e., non-proprietary and transparent – formats and the adherence to standards and best practices increase the usability of resources and their likelihood of long-term preservation.* Thus, recordings should be stored in common formats (like WAV) and should not, for example, be compressed using proprietary software or locked using proprietary password protection. Transcripts and language metadata should be stored in open, standards-based text formats, such as XML (www.w3.org/XML/; cf. Austin, 2006, pp. 101–107), and should adopt standard mark-up conventions, like those described by the Text Encoding Initiative (TEI, www.tei-c.org). They should not be stored in proprietary file types such as Microsoft Word documents, which are not readily usable without the proprietary software. In general, preservationists advise the avoidance of data formats that are linked only to specific software programs. Not only does this mean that a specific program is needed to read the file (such as Microsoft Word), but it also means that users of the data must rely on future versions of that software not changing their data structure or maintaining backward compatibility. Software versions (and the long-term survival of specific software programs) have proven to be extremely volatile, which is a bigger issue for the preservation of digital files than is often assumed. At the same time, many current linguistic analysis and annotation tools, such as Praat (Boersma & Weenink, 2011, www.fon.hum.uva.nl/praat/) and ELAN (Sloetjes & Wittenburg, 2008, www.lat-mpi.eu/tools/elan/), store their files in formats that are readable and parseable from outside the specific applications. This means that, for instance, Praat's TextGrid files can be read and imported by other software (such as ELAN) and, importantly, can be

accessed and parsed by customized scripts (or even plain text viewers) should this become necessary in the future. (Note, for example, that the webpage http://ncslaap.lib.ncsu.edu/tools/praat_to_text.php will convert certain types of Praat TextGrid files to tab-delimited plain text versions.) So, while it is conceivable that Praat might become unavailable someday or that its designers might change the format of its data structure (so that older files are no longer readable by newer software versions), it will be possible to salvage and use the data contained in the TextGrid files, provided that documentation is kept about the file structure. Praat has been used here as an example, but the same issues apply to XML-based data formats: unless the underlying structure of the data files is documented, maintained, and made available to users of the data, even the use of open formats can be problematic, as future users may not be able to interpret the information in the files.

3. *Preservation alone is insufficient without a corresponding plan to allow for the access and discovery of the data by potential users*. Some language data cannot be shared beyond the original research group – many sociolinguistic projects have constraints on the sharing of data based on human subjects-related or other agreements related to ethics and/or confidentiality – but to preserve data forever is ultimately a waste of effort, storage space, and money if those data cannot be accessed or discovered by anyone, ever. Researchers should think about the short-term, medium-term, and long-term life of their data. The short term can be thought of as the immediate future, the course of the actual research project, and one's individual interest in those data as "active" research data. The medium term may encompass one's complete research career and/or the lifetimes of the informants in the recordings. The long term is the unforeseeable future: what use can future scholars gain from the data as a part of the historical record of a language variety or a community?
4. *What rights exist for the sharing of data in the short, medium, and long term?* Questions of ethics, ownership of data and copyright, and sensitivities to the content of and participants in research are complex and are the subject of many current discussions in sociolinguistics and other disciplines. For sake of space and because they are addressed elsewhere (e.g., Trechter, Chapter 3; Besnier, Vignette 3a; Mann, Vignette 3b; Ehrlich, Vignette 3c; Sadler, Vignette 3d; Ngaha, Chapter 16; Charity Hudley, Chapter 17; and Starks, Vignette 17c; see also Childs, Van Herk, & Thorburn, 2011; Milroy & Gordon, 2003, pp. 79–87), I do not fully review these issues here. Instead, I sum up these discussions by recommending that researchers give full consideration to the questions of (short-, medium-, and long-term) access rights in the earliest stages of their research – before beginning fieldwork. Many of the future limitations on the use of preserved data can be alleviated by negotiating up front with the relevant human subjects authority (such as one's institutional ethics board) and by giving research participants a wide range of explicit options for how their recordings and derived data can be used in the future. (Austin, 2006, p. 101, further recommends that researchers assign future rights about data into their wills to ensure that procedures are in place to manage access to data after researchers die.)

As Bird and Simons (2003) make clear, within linguistics, members of the endangered language research community have pioneered the biggest efforts in data preservation and the development of best practices and standards for linguistic data management. For those researchers, the preservation of documentary evidence is crucial, as the languages themselves are endangered. Organizations such as the Open Language Archives Community (OLAC, www.language-archives.org), the Electronic Metastructure for Endangered Languages Data project (E-MELD, http://emeld.org), and the Hans Rausing Endangered Languages Project (HRELP, www.hrelp.org) have led web-based initiatives, practical tutorials, and workshops and in general have provided leadership and organization. The literature by those researchers, especially by those working on language documentation, is well developed and is quite relevant for sociolinguists. (In addition to Bird & Simons, 2003, I recommend Austin, 2006, for its extended discussion of a range of data-processing and archiving issues.)

Sociolinguists (and most other linguists) have lagged behind the endangered language community in these kinds of centralized initiatives, although some work has recently moved in this direction (Kendall, 2008; Kretzschmar et al., 2006), and, as was noted earlier, many scholars have come together in recent years through special panels at conferences and in workgroups to address these issues (e.g., Buchstaller et al., 2011; Coleman et al., 2011). While the endangered language community has very real needs with respect to the preservation of disappearing resources, much sociolinguistic research on minority dialects also records dying and endangered language varieties (Wolfram, 2002), and the importance of preserving these resources is quite clear. Meanwhile, even for extensively spoken and studied varieties, including many dialects of English, relatively few authentic spoken language data are actually publicly available, and any additions sociolinguists can make to the pool are valuable contributions.

The needs and data management requirements of sociolinguists are somewhat different from those of the language documentation community, however. The recordings and texts produced by language documentarians are generally seen as the end products of fieldwork and research (Austin, 2006, pp. 87–88). As such, the recordings themselves are (or at least should be) created in ways that are designed for the largest possible audience and are sensitive to the cultural norms and wishes of their informants. Sociolinguistic recordings are often just the first step in the generation of "data" (Kendall, 2008). They are often also, in fact, quite personal exchanges and can be private communicative events not intended for sharing (cf. Tagliamonte, 2012, pp. 115–116). Nonetheless, sociolinguists should turn to the initiatives of the documentation community for inspiration and guidance in the preservation and access of sociolinguistic data. Bird and Simons (2003) conclude their paper by saying:

> Today, the community of scholars engaged in language documentation and description is in the midst of transition between the paper-based era and the digital era. We are still working out how to preserve knowledge that is stored in digital form. During this transition period, we observe unparalleled confusion in the management of digital language documentation and description.

> A substantial fraction of the resources being created can only be reused on the same software/hardware platform, within the same scholarly community, for the same purpose, and then only for a period of a few years. However, by adopting a range of best practices, this specter of chaos can be replaced with the promise of easy access to highly portable resources.
>
> (p. 579)

It is unlikely that language documentarians feel that they have fully replaced the "specter of chaos," but their work has made great strides toward better practices. Many sociolinguists likely feel this same sense of chaos about how to preserve and manage sociolinguistic data in the long term. This situation seems to me only natural as the field negotiates the sorts of issues described by Bird and Simons. Sociolinguists have not yet done the same work of articulating shared values across researchers and negotiating shared best practices as language documentarians have done, but recent and ongoing conversations lead in this direction and are an important part of the process.

Managing Sociolinguistic Data

I turn now to some topics in the management of sociolinguistic data, which underlie the process of data preservation and access. This discussion is not meant to suggest particular best practices but rather to be explicit about the processes of one archiving initiative. Greater explicitness about how individual researchers and research groups manage and treat their data can lead to better research and to the eventual development of shared best practices.

Although we often treat them as such, issues of data management and preservation are not just problems to solve. They present opportunities to think deeply about the very nature of our data and how we interact with and conceptualize them (Kendall, 2008). As an example, the Sociolinguistic Archive and Analysis Project (SLAAP, http://ncslaap.lib.ncsu.edu; Kendall, 2007) is a web-based digitization and preservation project housed at North Carolina State University, featuring a growing archive of sociolinguistic audio recordings along with dynamic interfaces to those recordings. As of October 2011, over 2,400 interview recordings are stored in and accessible through SLAAP, amounting to over 1,900 hours of speech. The web-based archive has allowed a number of researchers around the world to access a shared, centralized recording and data archive. Access to the archive is password-protected and controlled at the level of the individual user account and at the level of the individual data collections, so different users have different levels of access to different sets of materials. This setup allows the same archive to house highly restricted collections (accessible to very few researchers) along with widely accessible collections. By aggregating many different collections and storing them within a unified architecture, SLAAP also allows the otherwise diverse materials to be put in communication with one another. Researchers can ask new questions of old data and search across collections for particular phenomena.

Beyond its shared, web-accessible interface, the centerpiece of the SLAAP software is a time-aligned annotation framework that is integrated with analytic

software, including Praat and R (R Development Core Team, 2011; http://R-project.org), allowing for features such as the automatic generation of spectrograms within a web-based audio player, the extraction of phonetic data from within a recording's transcript, multiple and dynamic displays of each transcript, and corpus linguistic analyses across the diverse materials in the archive. SLAAP has proven valuable for a wide range of uses (e.g., Carter, 2009; Dunstan, 2010; Herman, 2009; Kendall, 2009; Kendall, Bresnan, & Van Herk, 2011; Kohn, 2008; Thomas, 2010; 2011). (The SLAAP software and data model, as well as transcription method and conventions, are detailed elsewhere, e.g., Kendall, 2007; 2008; and http://ncslaap.lib.ncsu.edu/userguide/.)

While SLAAP, I believe, illustrates a number of benefits of thinking deeply about data management, a persistent issue in the long-term preservation and accessibility of research recordings is the problem of institutionalization, which presents a larger hurdle than the availability of specific tools and methods or any of the technical problems of data preservation. Many sociolinguistic data collections depend on their original collector to maintain them, and many researchers create impressive websites about their work and may even maintain their own data in a web-accessible format. However, these kinds of resources take extensive time (and cost) to maintain. Traditionally, these activities have not been evaluated as a part of researchers' academic "credit" for advancement, so we are often, in fact, disincentivized to spend the extensive and sustained effort required to ensure that our materials are accessible to others and maintained in the long term.

SLAAP has attempted to address the problem of institutionalization by consolidating the data collections of many researchers into one centralized archive and by teaming with the North Carolina State University Library to manage it. This relationship has proven to be an extremely valuable and rewarding partnership. The library provides the infrastructure and expertise to support the archive system's ongoing operation, while the sociolinguists provide the domain- and need-specific knowledge to develop the actual software and user interfaces. SLAAP admittedly does not fully solve the problem of institutionalization. Its long-term maintenance and some aspects of its day-to-day operations depend on one or two administrators who are academic linguists and not full-time archivists. The determination of a long-term management plan and the eventual scope of SLAAP's archive (e.g., what data are relevant additions to the archive, how to manage increased growth) are issues still being thrashed out. Nonetheless, academic libraries can make excellent partners in the preservation and even short-term management of resources.

Recent changes at several levels of academic structure indicate that work on data management and preservation will be an increasing part of researchers' obligations and may become academically "credited" activities. For instance, many funding agencies, such as the National Science Foundation in the United States and the Social Sciences and Humanities Research Council of Canada, have recently instituted policies about the management, preservation, and dissemination of data collected under funded research. It is likely that these kinds of policies will make the explicit treatment of data a larger part of sociolinguistic

research endeavors and will further promote the development of central solutions in the coming years.

SLAAP is one example of a speech data management system of value to sociolinguists and represents just one possible approach to data management and preservation. Further, it represents one collaborative group's attempts to explore possible models for data management rather than to provide a definitive solution for all sociolinguistic data and all sociolinguists. Other systems are being developed, such as LaBB-CAT (formerly called ONZE Miner: Fromont & Hay, 2008; http://onzeminer.sourceforge.net), and a number of research groups are building sophisticated data management and dissemination systems for their own data (e.g., some of the projects described in Beal et al., 2007a; 2007b). However, to my knowledge none of the current projects (including my own) adequately surmounts the issue of true institutional support for the long-term storage of diverse sociolinguistic research materials. Organizations like the Linguistic Data Consortium (www.ldc.upenn.edu) and the Oxford Text Archive (http://ota.ahds.ac.uk) represent perhaps the closest and best options, but as of now even these impressive archive centers are not well suited to the needs of sociolinguists and their data. As Kretzschmar et al. (2006) argue, it seems crucial that the field develop shared models and tools for data preservation and data sharing. This seems to me to go beyond the need for shared (and best) practices in terms of data preservation to the very issues of where to store and how to manage our actual data files. The current model of "everyone for him- or herself" is untenable in the long run. And preservation is clearly an issue for the long run.

Moving Forward

This chapter has reviewed a number of issues in the preservation, management, and larger accessibility of sociolinguistic data. I have provided examples of recent projects that have benefited from preserved data (e.g., ONZE) and those that have developed data management solutions (e.g., SLAAP). I have drawn from and pointed to work from language documentation and description and the endangered language community, where researchers have well articulated their needs for data preservation and begun to develop best practices. I have also made several suggestions about how data preservation and access issues should be thought about from the very earliest stages of research.

I have not presented a specific how-to guide for preserving sociolinguistic data because one does not yet exist. Instead, I urge sociolinguists to collaborate to explore and develop shared best practices for our data. By being explicit about our data management practices and plans and by having discussions (e.g., through publications, workshops, and conversations) about how we interact with and conceptualize our data, we can move forward in the treatment and preservation of our data.

References

Austin, P. K. (2006). Data and language documentation. In J. Gippert, N. Himmelmann, & U. Mosel (Eds.), *Essentials of language documentation* (pp. 87–112). Berlin: Mouton de Gruyter.

Bailey, G. (2002). Real and apparent time. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 312–332). Malden, MA: Blackwell.

Bailey, G., Maynor, N., & Cukor-Avila, P. (Eds.). (1991). *The emergence of Black English: Text and commentary*. Amsterdam: John Benjamins.

Beal, J. C., Corrigan, K. P., & Moisl, H. L. (Eds.). (2007a). *Creating and digitizing language corpora*, Vol. 1: *Synchronic databases*. Basingstoke, UK: Palgrave Macmillan.

Beal, J. C, Corrigan, K. P., & Moisl, H. L. (Eds.). (2007b). *Creating and digitizing language corpora*, Vol. 2: *Diachronic databases*. Basingstoke, UK: Palgrave Macmillan.

Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language, 79*(3), 557–582.

Boas, H. C. (2006). From the field to the web: Implementing best-practice recommendations in documentary linguistics. *Language Resources and Evaluation, 40*(2), 153–174.

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer [Computer software]. Amsterdam: Phonetic Sciences, University of Amsterdam.

Buchstaller, I., Corrigan, K. P., Mearns, A., & Moisl, H. (Organizers). (2011). Dialect and heritage language corpora for the Google generation. Workshop presented at the Methods in Dialectology 14 conference. London, ON.

Carter, P. M. (2009). Speaking subjects: Language, subject formation, and the crisis of identity. (Unpublished doctoral dissertation). Duke University, Durham, NC.

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

Coleman, J., Hall-Lew, L., & Temple, R. (2011). New methods for community sharing of spoken corpora. Paper presented at The UK Language Variation and Change 8 conference. Ormskirk, Lancashire, UK.

Dunstan, S. B. (2010). Identities in transition: The use of AAVE grammatical features by Hispanic adolescents in two North Carolina communities. *American Speech, 85*(2), 185–204.

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

Gordon, E., Maclagan, M., & Hay, J. (2007). The ONZE Corpus. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora*, Vol. 2: *Diachronic databases* (pp. 82–104). Basingstoke, UK: Palgrave Macmillan.

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

Gregersen, F., & Barner-Rasmussen, M. (2011). The logic of comparability: On genres and phonetic variation in a project on language change in real time. *Corpus Linguistics and Linguistic Theory, 7*(1), 7–36.

Herman, D. (2009). *Basic elements of narrative*. Malden, MA: Blackwell.

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

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

Kendall, T. (2009). Speech rate, pause, and linguistic variation: An examination through

the Sociolinguistic Archive and Analysis Project. (Unpublished doctoral dissertation). Duke University, Durham, NC.

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

Kendall, T., Bresnan, J., & Van Herk, G. (2011). The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory, 7*(2), 229–244.

Kendall, T., & Van Herk, G. (Eds.). (2011). Corpus linguistics and sociolinguistic inquiry. Special issue of *Corpus Linguistics and Linguistic Theory, 7*(1): introduction.

Kohn, M. (2008). Latino English in North Carolina: A comparison of emerging communities. (Unpublished master's thesis). North Carolina State University, Raleigh, NC.

Kretzschmar, W. A., Jr., Anderson, J., Beal, J. C., Corrigan, K. P., Opas-Hänninen, L. L., & Plichta, B. (2006). Collaboration on corpora for regional and social analysis. *Journal of English Linguistics, 34*(3), 172–205.

Milroy, L., & Gordon, M. (2003). *Sociolinguistics: Method and interpretation*. Malden, MA: Blackwell.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software]. Vienna: R Foundation for Statistical Computing.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco, May 28–30.

Tagliamonte, S. A. (2012). *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Blackwell.

Thomas, E. R. (2010). A longitudinal analysis of the durability of the Northern/Midland dialect boundary in Ohio. *American Speech, 85*, 375–430.

Thomas, E. R. (2011). *Sociophonetics: An introduction*. New York: Palgrave Macmillan.

Wolfram, W. (2002). Language death and dying. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 764–787). Malden, MA: Blackwell.