10 Working with and Preserving Existing Data

Gerard Van Herk

Part III of this volume, "Working with and Preserving Existing Data," explores the issues and challenges associated with adapting existing data to the needs of sociolinguists. Data treatment, in other words.

It is perhaps useful to consider why sociolinguists, especially variationists, might be more willing and able than other researchers to work with existing data. Variationists' traditional methods of data collection and analysis actually predispose us to a two-step process: first working to collect as naturalistic data as possible, often through the sociolinguistic interview, followed by a close reading of the resulting materials to decide what linguistic variables might best lend themselves to analysis and discussion (see, for example, Wolfram, 1993). This means that much of our data collection is blind to eventual purpose. From there, it is one small step to using data that were not collected for sociolinguistic reasons at all. There are exceptions to this, obviously, since the earliest days of the field: word lists, read passages, Labov's department store study and its Rapid and Anonymous Surveys (Labov, 1966).

Usually, though, sociolinguistic interviews are seen as the gold standard, in large part because they are intended to draw respondents' attention away from the recording process, to access their vernacular. What is it about recordings generally that encourages interviewees to *avoid* vernacular speech? The microphone and recorder? The act of being recorded? Traits of the interviewer (linguist, academic, stranger)? The interviewee's knowledge of the goals of the researcher? Techniques like the danger-of-death question and linguistic modules are designed to overcome such problems, but, to some extent, data from other sources avoids them by not introducing them in the first place.

Once we decide that all the data world is our research stage, certain questions arise:

1. What are "data"? The world is full of linguistic material these days, thanks largely to the internet, and it is extremely easy to access (and, in some ways, easy to do specific types of analysis … an issue perhaps beyond the point of this volume). At what point does material turn from "a bunch of words and stuff" into something we can analyze?
2. What are data "for"? Are they a pool to dip into for multiple studies? Something to share? How do data need treating in order to be shareable?

3. What are "natural" data? What can we do with scripted data? What are our expectations about particular genres?
4. What are the advantages and disadvantages of particular types of existing data? Are there specific caveats for specific data? How do we justify the use of a particular data source? Should we have to?
5. What do existing data give us that we can't get, or get more easily, from sociolinguistic interview data (or similar "in-house" data collection methods)? Different kids of naturalness? Interactions? What do we lose by using such data?

The chapters and vignettes that appear in this section address these questions in researcher-friendly formats.

In Chapter 11, "Written Data Sources," Edgar W. Schneider considers several questions that a researcher might ask before choosing to work with written data. At their most basic, these are writing-specific versions of the kinds of questions we all ask ourselves about our data: Why use this data? How do we find the best instances from all available data? What techniques are most appropriate to the data type? How do we deal with the possible shortcomings of the data? Schneider's description of the rigor required in selection and treatment of written data implicitly argues for the quality of analysis that is possible through careful methodological choices, while his examples of written data sources show us the rich linguistic material that is available for consideration. Vignette 11a, "Accessing the Vernacular in Written Documents," by France Martineau, takes us through the steps involved in choosing her written data sources and the kinds of linguistic information that she found, and argues for a reconsideration of our field's focus on the oral. For both authors, a research focus on historical processes actually requires the use of written data, as recordings rarely give us access to speakers born before about 1880 – see, for example, work on the ex-slave recordings (Bailey, Maynor, & Cukor-Avila, 1991), Quebec folklore recordings (Poplack & St-Amand, 2007), or New Zealand radio field recordings (Gordon, Hay, & Maclagan, 2007). At that early point, some of the problems of representativeness and validity raised by Schneider and Martineau are also relevant to recordings.

A different barrier to access and analysis of data is presented in Vignette 11b, by Philipp Sebastian Angermeyer, "Adapting Existing Data Sources: Language and the Law." Here, the power asymmetries inherent in the legal system may distort the language produced, or its representation, while also raising ethical questions about the use of data. Angermeyer addresses questions of data selection and treatment similar to those raised by Schneider.

Another aspect of turning language into something called Sociolinguistic Data is addressed in the next two vignettes, as Alexandra D'Arcy and Cécile B. Vigouroux discuss the benefits and perils of transcribing data. D'Arcy's "Advances in Sociolinguistic Transcription Methods" (Vignette 11c) stresses the degree to which research goals drive the decisions made during the transcription process, including the decision to transcribe in the first place. If the focus is on variation in linguistic forms, rather than the nature of an interaction, a transcription protocol is needed to ensure accurate representation of those forms, without

getting lost in a bog of (analytically) unnecessary detail. If, on the other hand, the focus is on how participants' linguistic and non-linguistic performances are related, as in Vigouroux's "Transcribing Video Data" (Vignette 11d), then the researcher needs a method that allows for representation and alignment of different aspects of the performance. Vigouroux reminds us that decisions about how to collect and represent data are themselves part of the analytical plan: by choosing to video-record rather than audio-record an interaction, the researcher is making claims about the importance of visual information, and that information must therefore be included in the resulting transcription.

In Chapter 12, "Data Preservation and Access," Tyler Kendall addresses a basic question that often remains unanswered (or answered through its avoidance): once we have a bunch of data, what do we do with those data? In particular, how do we make sure that sociolinguistically useful data remain available and known to other researchers? The "forward compatibility" of existing recordings and transcriptions can be compromised by the format in which material is stored, while shareability can be limited by confidentiality and other ethical requirements, as well as by a lack of awareness that materials even exist. Kendall's suggestions echo what careful readers may be seeing as a recurring theme in this book: think about issues of data collection at the beginning of your research project.

Vignette 12a, "Making Sociolinguistic Data Accessible," by William A. Kretzschmar, Jr., takes us through parts of that thought process, with a special focus on considering the needs of all the potential audiences for your data, everything from the original interviewee to future generations of researchers who may have technical or research requirements that we have not even thought of yet. Kretzschmar concludes his vignette with a call to us to "give it all away," a theme picked up by Mark Davies in Vignette 12b, "Establishing Corpora from Existing Data Sources." As an example, Davies offers the Corpus of Contemporary American English, which he created in less than a year by using existing materials. Here, the challenges are more like those described by Schneider and Martineau: how does a researcher decide which of the available materials are sociolinguistically good? Joan C. Beal and Karen P. Corrigan's Vignette 12c, "Working with 'Unconventional' Existing Data Sources," uses their work on the Newcastle Electronic Corpus of Tyneside English to illustrate how potentially competing needs (such as ethics, searchability, preservation, and accessibility) can be addressed.

Chapter 13, "Working with Performed Language: Movies, Television, and Music," picks up on an idea central to Davies' vignette: how "natural" are scripted media data, and what are they good for? Robin Queen uses recent media discussions of "vocal fry" (creaky voice) to exemplify the sociolinguistic issues that can be considered in performed language data. She then discusses how performed data requires (or permits) particular theoretical approaches (linguistic ideology, styles, indexicalities, enregisterment) and methods of organizing data (representativeness, preservation, selection, transcription, copyright).

In Vignette 13a, "Working with Scripted Data: A Focus on African American English," Tracey L. Weldon gives us an example of how such research can work. As Weldon points out, research on African American English is well known for

an obsession with tapping the vernacular, but by considering filmic representations of the variety, we can address questions of authenticity, audience, and negotiation of dialogue. This issue of negotiation and change (especially between original scripts and released materials) is discussed in greater detail by Michael Adams in Vignette 13b, "Working with Scripted Data: Variations among Scripts, Texts, and Performances." A central idea is that there are multiple versions of the "text" of a performance, and each can take on a life of its own.

Finally, Chapter 14, "Online Data Collection," by Jannis Androutsopoulos, addresses some of the concerns specific to online data (both the harvesting of data from existing sources and the creation of new data, through interaction with language users). The chapter includes a detailed breakdown of the characteristics of online language (text) and social organization (place) that may require consideration. Online language is plentiful, written, and organized into multiple modes and genres, while online social organization involves new contexts and, often, limited information on social characteristics of participants. Some distinctions already familiar to sociolinguists – language focused vs. speaker focused, ethnographic vs. non-ethnographic, macro vs. micro – can be adapted to discussions of online data, while other distinctions may be more immediately relevant to online data, especially those relating to mode and genre. Our ability to "eavesdrop" online may encourage a greater focus on the interactional aspects of language use, as well as introducing new wrinkles to the problems of ethical use and anonymity.

Many of the ideas discussed in this section's chapters and vignettes are the things that sociolinguists discuss over beverages at the margins of conferences and workshops and get-togethers, the things that do not always make it into the "final cut" of academic papers. The authors, through their discussions and reminiscences, remind us of the tight links in our field between the daily decisions in data collection and the theoretical questions we try to address.

References

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

Gordon, E., Hay, J., & Maclagan, M. (2007). The ONZE corpus. In J. C. Beal, K. P. Corrigan, & H. L. Moisl, *Creating and digitizing language corpora*, Vol. 2, *Diachronic Databases*. New York: Palgrave Macmillan.

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

Poplack, S., & St-Amand, A. (2007). A real-time window on 19th-century vernacular French: The Récits du français québécois d'autrefois. *Language in Society, 36*(5), 707–734.

Wolfram, W. (1993). Ethical considerations in language awareness programs. *Issues in Applied Linguistics, 4*(2), 225–255.