Data Collection in Sociolinguistics

METHODS AND APPLICATIONS

EDITED BY

Christine Mallinson

Becky Childs

Gerard Van Herk

Data Collection in Sociolinguistics

Data Collection in Sociolinguistics: Methods and Applications is an accessible, contemporary guide to data collection, a central pursuit of sociolinguistic research. Contributions by veteran as well as up-and-coming sociolinguists cover methods of data collection, whether generating new data or when working with existing data, and tackle important questions of ethics, data sharing, data preservation, and community outreach. Other cutting-edge topics, featured in main chapters and in shorter, reader-friendly vignettes, include the use of public documents, virtual world research, collecting online data, working with video data, and cross-cultural issues in data collection. A comprehensive volume, *Data Collection in Sociolinguistics* is set to become an indispensable guide for students and scholars interested in both the broad outlines and the finer details of sociolinguistic data collection.

Christine Mallinson is Associate Professor of Language, Literacy, and Culture and Affiliate Associate Professor of Gender and Women's Studies at the University of Maryland-Baltimore County.

Becky Childs is Associate Professor of English at Coastal Carolina University.

Gerard Van Herk is Associate Professor and Canada Research Chair in Regional Language and Oral Text at Memorial University of Newfoundland.

Data Collection in Sociolinguistics

Methods and Applications

Edited by Christine Mallinson,
Becky Childs, and Gerard Van Herk

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2013
by Routledge
711 Third Avenue, New York, NY 10017

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging in Publication Data
Data Collection in Sociolinguistics : Methods and Applications/ Edited by Christine Mallinson, Becky Childs, and Gerard Van Herk ; University of Maryland, Baltimore County, Coastal Carolina University, and Memorial University of Newfoundland.
pages cm
1. Sociolinguistics–Research–Methodology. 2. Qualitative research.
I. Mallinson, Christine. II. Childs, Becky. III. Van Herk, Gerard.
P40.3.D38 2013
306.44072′3–dc23

2012042040

ISBN: 978-0-415-89856-0 (hbk)
ISBN: 978-0-415-89857-7 (pbk)
ISBN: 978-0-203-13606-5 (ebk)

Typeset in Minion
by Wearset Ltd, Boldon, Tyne and Wear

Contents

PART II
Generating New Data

PART IV
Sharing Data and Findings 251

Illustrations

Figures

Tables

Foreword

Observing the Observers

J. K. Chambers

The ideal data for studying the social uses of language, as all the authors in this book agree, are speech produced in natural circumstances, unmonitored and carefree. Getting access to that kind of speech is challenging, for obvious reasons. For starters, in order to make the speech accessible for study it must be elicited, and eliciting speech usually introduces “foreign” elements into the speech act, including the presence of an outsider (the investigator), recording devices (microphones in plain view), controlled ambience (full attention and relative quiet), and strange tasks interspersed with conversation (reading a text or word list, identifying pictures and other means to guarantee comparable material from all subjects). The act of observing speech alters its nature.

Sociolinguists seek to observe speech as people use it when they are not being observed. That is the “observer’s paradox,” and it has been a central preoccupation of sociolinguistic methodology from the beginning (Labov, 1972, p. 61). It is, naturally, a central preoccupation of this book; it is cited explicitly in the chapters and vignettes by Kara Becker, Niko Besnier, Charles Boberg, Becky Childs, Cynthia G. Clopper, Paul De Decker and Jennifer Nycz, and Sara Trechter. It is implicit almost everywhere.

In the four eventful decades since William Labov gave a name to the observer’s paradox, sociolinguists have come up with several ways of neutralizing it. Indeed, one of the rewarding sub-themes of this book is discovering how the experienced fieldworkers who contributed the chapters and vignettes got around it.

One obvious stratagem is diversion. One of the most ingenious examples in my experience was devised by an undergraduate in a course I taught in the 1970s. In those days, the stressed vowel in the word *tomato* had three variants in Toronto: either [ei], the North American variant, or [a], modeled on the British pronunciation, or [æ], a distinctive Canadianism that came into being as a fudge between the other two variants. In order to discover the social correlates of the three variants, my student mounted four pictures on a poster: a cauliflower, a carrot, an apple, and (inevitably) a tomato. He visited department stores frequented by different social classes (following Labov’s famous department-store study described, for instance, by Barbara M. Horvath in this book). He approached shoppers, and, after a friendly introduction, he showed them the poster and asked, “How many of these are vegetables?” If they said “two,” he challenged them: “Why not three?” They inevitably answered, “Because a tomato is a fruit, not a vegetable." And, conversely, if they answered "three," he queried their answer, and was told, "Because a tomato is a vegetable, not a fruit (whatever other people might say)." His subjects had no idea, of course, that he was eliciting their pronunciations; they assumed he was challenging their botanical acumen, in which the classification of the tomato is a well-known point of contention. In a short time, he accumulated hundreds of responses and he was able to show that social class sometimes interacted with age: people under 40 all used the [ei] variant except for a few oddballs from the upper middle class. (Since then, they too have disappeared, and the [ei] variant is nearly unanimous throughout Canada.)

This method has proven practicable for small-scale studies like the *tomato* variable, known as "rapid and anonymous surveys" (discussed by Charles Boberg in Chapter 8 and Gerard Van Herk in Chapter 10). Nevertheless, the basic idea of framing the interview context so that the subject's attention is fixed on something other than the speech act is one of the key devices for blunting the paradox or, put positively, for eliciting unmonitored speech. Several authors in this book make suggestions and provide models toward that end.

Special communities require specialized methods, and they too are covered incisively in this book. Among them are immigrant communities (discussed by Rajend Mesthrie, James A. Walker, and Michol F. Hoffman, among others), closed enclaves ("clans" in James Stanford's vignette), and moribund dialects and endangered languages (discussed, respectively, by Patricia Causey Nichols and D. Victoria Rau). There are also data sources that are far removed from unmonitored natural speech but, with suitable precautions, can yield sociolinguistic insights. Prominent among these are the "public" languages of the mass media (called "performed language" by Robin Queen in Chapter 13 and "scripted data" in the vignettes by Tracey L. Weldon and Michael Adams). Equally public if less prominent are courtroom transcripts (discussed from different perspectives in vignettes by Susan Ehrlich and Philipp Sebastian Angermeyer).

Besides speech, there are also data sources in written materials. As Edgar W. Schneider says in his discussion of written material (in Chapter 11), writing "represents a secondary encoding of speech." The language we write is a kind of abstraction of the language we speak, hemmed in as it is by spelling conventions and stylized formatting. Nevertheless, written records existed for a millennium or more before audio recordings. Comparative linguistics, the most vital branch of language studies until the early 20th century, made monumental advances in genetic classification based almost entirely on classical texts. Those materials and other written documents continue to yield insights, and those insights are all the more astute now that we have deeper understandings of spoken vernaculars. Knowing the dynamics of living languages enriches our understanding of ancient processes on the understanding that linguistic processes were the same in nature and kind hundreds of years ago as the ones we now observe.

One type of written documentation that has long proved useful in supplementing our linguistic knowledge is the written questionnaire, discussed by Charles Boberg (Chapter 8). The obvious limitation of asking people to tell us what they say is that they can only tell us what they *think* they say, which is not always the same thing. Subtle phonetic distinctions often require some training to recognize, and rare syntactic constructions sometimes strike users as strange even though they themselves may use them. Self-administered questionnaires definitely work best on well-defined and easily discernible features. Countering this limitation, as Boberg makes clear, is their efficiency. Written surveys can cover a large territory with great density in a short time.

One of the minor cavils in Boberg's account is illustrated from my Dialect Topography survey, and I cannot resist showing that subsequent information gives it a rather more positive spin than was originally evident. Open-ended questions, Boberg notes, sometimes "elicit an overwhelming variety of minority responses." As a case in point, he cites a question about the schoolyard prank now widely known as a *wedgie*. True enough, when the question was first posed in the early 1990s, the responses were (almost) "overwhelming": specifically (as discussed in detail in Chambers, 2012, pp. 471–473) there were four main responses and at least a dozen minor ones – and almost everybody over 50 left it blank. What a mess, we thought at the time. But when we replicated the survey 10 years later, the results were stunningly different: this time, there was only one word for it. Almost everyone (93 percent) called it *wedgie*, including many old-timers. What had happened in the 10-year interval is that the *wedgie* had entered general consciousness. It had previously existed in the semi-literate subculture of grade-schoolers, but suddenly it was known to almost everyone. The word *wedgie* showed up in dictionaries, and it was called that by teachers, parents, and some grandparents as well as by schoolchildren. The shift from the profusion of responses in the first survey to the focusing of the later one documents "a prototypical standardizing change" (2012, p. 471), one of the best yet documented. The real-time evidence of the second survey illuminated the profusion of minority responses that formerly seemed overwhelming. Without it, Boberg may be right in saying that open-ended questions may sometimes yield more information than we know what to do with.

The sheer volume of data when we study language as it is used by real people in real situations was one of the chronic problems of dialectology. As Kretzschmar, Schneider, and Johnson (1989, p. v) put it some years ago: "The development of dialect studies, whether geographical or sociolinguistic, has always been hampered by a superfluity of data." This statement appeared in one of the pioneering introductions to computer applications in dialectology, and so it set up the problem of "superfluity" in the context of its solution. Data-handling is no longer the overriding problem that it once was. Our discipline has made striking advances in storing, manipulating, and processing data. These aspects are to some extent inseparable from other matters and find their way into virtually every chapter of the book. They come to the fore especially in the chapter on technology (Chapter 7 by Paul De Decker and Jennifer Nycz) and the vignettes associated with it, and in the chapter on preserving an accessing data (Chapter 12 by Tyler Kendall) and its vignettes.

This collection provides a balanced, judicious, forward-looking summation of the ways in which we collect, access, and process the data that are the foundation of our enterprise. In its format and its tone, it has the feeling of a symposium involving a select group of sociolinguists sharing their personal experiences as well as their collective wisdom. It is an invaluable sourcebook for researchers and students and also for veteran fieldworkers in the diverse situations we face on entering the community.

References

Chambers, J. K. (2012). Homogeneity as a sociolinguistic motive in Canadian English. *World Englishes, 31*, 476–477.

Kretzschmar, W. A., Jr., Schneider, E. W., & Johnson, E. (Eds.). (1989). *Computer Methods in Dialectology*. Special edition of *Journal of English Linguistics, 22*.

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

Acknowledgments

First and foremost, we thank the many scholars, colleagues, and friends who have contributed to this volume and, in so doing, have made it a premier resource on data collection in sociolinguistics. We also appreciate the hard work of the Routledge editorial team, including Louisa Semlyen, Elysse Preposi, Ivy Ip, and Michael Andrews; Rita J. Turner, who helped proofread the manuscript; and the anonymous reviewers whose excellent suggestions on our proposal helped refine the scope of this volume.

We particularly thank Ana Carvalho, who piloted this volume in her graduate research methods course at the University of Arizona. We also benefited from the feedback we received at our (2010) workshop, "Participants, Purpose, and Process: An Interactive Workshop on Data Collection in Variationist Sociolinguistics," at the New Ways of Analyzing Variation 39 conference.

Parts of this book were informed by material we have published elsewhere. We acknowledge the (2011) article by Becky Childs, Gerard Van Herk, and Jennifer Thorburn, "Safe Harbour: Ethics and Accessibility in Sociolinguistic Corpus Building," part of the special issue "Corpus Linguistics and Sociolinguistic Inquiry" of the journal *Corpus Linguistics and Linguistic Theory, 7*(1). Additionally, Chapters 15 and 17, by Christine Mallinson and Anne H. Charity Hudley, are informed by work supported by the National Science Foundation under Grant No. 1050938/1051056.

Finally, all three co-editors thank our respective universities – UMBC, Coastal Carolina University, and Memorial University of Newfoundland – for the sabbaticals that allowed us to complete this project. We are grateful for the support.

Copyright Acknowledgments

Table 11.1, “Categorization of Text Types According to Their Proximity to Speech,” was originally published on p. 73 in “Investigating Variation and Change in Written Documents,” by Edgar W. Schneider, in the *Handbook of Language Variation and Change* (2002), ed. by J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, published by Wiley-Blackwell.

Part I

Research Design

This page intentionally left blank

1 Research Design

Christine Mallinson

Part I of this volume, “Research Design,” addresses two central concerns in relation to sociolinguistic data collection: research design and ethics. First, the chapters and vignettes in this section provide guidelines, offer suggestions, and troubleshoot challenges that can arise when asking research questions, choosing frameworks and paradigms, and designing a study, all of which directly affect what data are to be collected and how. Second, while many authors throughout the volume discuss ethics, the authors of the chapters and vignettes in this section grapple with specific challenging ethics-related questions that are particularly, though not exclusively, relevant to sociolinguists conducting research with human subjects. How should we represent research participants? What issues should we consider when working with vulnerable populations, who may need more protection than ethics boards would normally require? What sort of ethical dilemmas face scholars who work with written documents? How should our ethical decision-making protocols be adapted when conducting research online? As the authors in this section assert, these questions should be considered not only at the beginning but throughout the research process.

In Chapter 2, Barbara M. Horvath emphasizes the diverse frameworks, topics, and methods that are included under the umbrella of sociolinguistic research. Sociology, anthropology, geography, psychology, and other disciplines have influenced sociolinguistics, leading to diversity in research design, methods, and the linguistic and social phenomena to be investigated. As Horvath says, the connection between the linguistic and the social is inseparable and requires the studying of both. As a result, sociolinguists often employ qualitative, quantitative, and mixed methodological approaches, as questions about language change in progress (generally quantitative) are frequently tied to questions about what variability means to speakers (generally qualitative). At the same time, while an array of research frameworks, paradigms, designs, and methods is available to sociolinguists, the central concern of sociolinguistics as a field is the nature of language variation and how it relates to social contexts, factors, and outcomes. Within that scope, researchers must decide which aspects of language variation, change, and social meaning to foreground, which to background, and how to do so as they plan and conduct their research.

In Vignette 2a, Marcia Farr extends the conversation about research paradigms and design to multidisciplinary sociolinguistic studies, in which a researcher draws from multiple disciplines to inform the research at hand. Farr illustrates the benefits and challenges of this approach with her own study of transnational Mexican families, which drew from linguistic anthropology, sociolinguistics, history, sociology, and cultural studies. Multidisciplinary research requires a deep understanding of the concepts being borrowed from other fields. It also requires researchers to be flexible, patient, and open to the evolution that research drawing from multiple theoretical and empirical traditions often necessitates – a process that, Farr says, is demanding but ultimately rewarding.

Three subsequent vignettes address the topic of variables in sociolinguistic data collection. In Vignette 2b, "How to Uncover Linguistic Variables," Walt Wolfram notes the importance of examining linguistic variables that may be outside of the "canonical set" but that may nevertheless provide important insight into sociolinguistic variation. Two such variables – *a*-prefixing in Appalachian English and the "call oneself" construction in African American English – are useful case studies in the methodological and analytic challenges that can arise when uncovering, describing, and analyzing linguistic variables.

In Vignettes 2c and 2d, James N. Stanford and Rania Habib respectively discuss complex social variables, which are often imbued with local and contextualized meanings of which researchers may initially be unaware. In his work in southwest China, *clan* emerged as a meaningful social variable only after Stanford became engaged in the community, interacted with a range of residents, and learned the cultural knowledge required to interpret its relevant social structures. In her research in one rural and one urban speech community in Syria, Habib intended to investigate the role of social class on language variation. Unlike in Western contexts, however, where education, occupation, and income are often good proxies for social class, in these communities income and residential area proved to be the relevant class indicators. Both Stanford and Habib note the limitations of assuming that social variables operate in the same way across different contexts. Rather, researchers must acquire in-depth knowledge of the community to determine which social variables are relevant and how they relate to sociolinguistic variation.

Conducting ethical research with those from whom we collect our data has long been recognized as a critical goal for sociolinguists, especially for those who conduct field- and community-based research. In Chapter 3, "Social Ethics for Sociolinguistics," Sara Trechter provides readers with grounding in both normative ethics, which focuses on establishing criteria for right and wrong actions, and applied ethics, which considers how to act in specific situations. She notes that while sociolinguists have traditionally done well in considering applied ethics, we have paid correspondingly less attention to normative ethics, such that the broad concepts of ethics and ethical engagement remain undertheorized in sociolinguistics. As such, Trechter challenges sociolinguists not only to think about the real-life decisions that must be made while conducting research, but also to articulate and debate our philosophical standards and models for ethical reasoning that guide our judgment. Doing so will allow sociolinguists to establish effective, consistent recommendations for how to conduct ethical research.

The major themes that arise in Chapter 3 center on the roles of researchers and participants in sociolinguistic studies and the power relations already present in researcher–participant dynamics. Drawing on her experience as a member of the Linguistic Society of America committee that developed the official LSA Ethics Statement (2006–2009), as well as several case studies from the sociolinguistic literature, Trechter identifies important considerations that sociolinguists may face when determining ethical obligations to research participants and communities. How much involvement should a researcher have in a community, given that sociolinguistic research tends to hinge upon engaging with members of a community in order to obtain data? How should the needs of a given community be assessed, and how should research participants be represented? How might research affect and be affected by the sociopolitical contexts in which participants and communities are situated? How might a researcher's status as an insider vs. outsider (or something in between) affect and be affected by her or his ethical obligations – not only to the community and to the participants, but also to her- or himself as a researcher? What do we, as researchers, hope to gain from our work and how exactly do we benefit, particularly at different stages of our careers? What are the rights and roles of various academic stakeholders in the power relations that occur before, during, and after a researcher has engaged with a community or with participants? While answers to ethical questions are generally neither immediately evident nor clear-cut, Trechter advocates that researchers reflect on ethics, sociopolitical relationships, and power dynamics not simply at one or two points but rather at every stage: from the point at which we begin to design our project and continuing throughout the course of the research process.

Following Chapter 3 are five vignettes, each of which provides examples of ethical considerations and challenges that sociolinguists have faced. The first three vignettes illustrate the fact that to do ethical research, sociolinguists must consider how we relate to our research participants, whether we know them and are in close contact with them or not.

Sometimes, what researchers think is trivial, research participants may find harmful. In Vignette 3a, "Responsibility to Research Participants in Representation," Niko Besnier discusses his own research in the Central Pacific. While the study of gossip is a relatively common topic in sociolinguistics, it proved to be a sensitive one for the research participants, and community members had different concerns about how their linguistic practices might be represented locally and abroad. As Besnier explains, the intention not to harm research participants does not necessarily ensure that harm is not done to them or experienced by them; in addition, the concept of "do no harm" does not prioritize ways of giving back to participants to ensure that they benefit from the research they agreed to participate in.

In other situations, individuals may not recognize the full potential for harm that they may face if they consent to participate in research, and even the regulations of ethics boards may not ensure that full protection of research participants is secured. In Vignette 3b, "Conducting Research with Vulnerable Populations," Stephen L. Mann describes the dilemmas he faced in his research observing at a public drag talent show held in a gay bar and interviewing self-identified gay and queer men in the US South. On the basis of his own understanding of the potential for harm that might result from his study, Mann decided to adopt a stricter stance toward anonymity than even his university ethics board would have required or his participants themselves had requested.

In some cases in which a researcher is using secondary data, consent from research participants may not be required, but ethical issues may nevertheless arise based on how data are represented and individuals are portrayed. In Vignette 3c, "Ethical Dilemmas in the Use of Public Documents," Susan Ehrlich discusses her research on the discourse of women who have been complainants in rape trials. Because she works with public documents, the women whose language data she is analyzing are not active participants in the research; furthermore, there is the potential for their data to be read and interpreted by others in ways that objectify and sexualize the women. For Ehrlich, questions linger as to how to protect participants from misrepresentations and how to use research to benefit them in the face of the potential, however indirect, for research to cause harm.

Vignette 3d, "Real Ethical Issues in Virtual World Research," by Randall Sadler, deals with similar themes related to the domain of conducting online research. As Sadler discusses, ethical challenges and temptations can arise when collecting data in virtual worlds. With examples from research conducted in Second Life, Sadler provides recommendations for how to assign pseudonyms, obtain informed consent, evaluate participants' accessibility, consider how participants perceive privacy, and assess risk to participants in order to help maintain high ethical standards when conducting online research. In sum, technological change can affect the quality, type, and scope of language data that sociolinguists collect; it can also affect research participants, who on the one hand may be increasingly comfortable with technology, access to media, self-publication, and self-revelation, but on the other hand may be less aware of the potential for risk when agreeing to take part in online research.

As the authors of these chapters and vignettes suggest, as sociolinguists we should plan our research carefully and strategically in advance in order to maximize our potential for conducting effective and ethical research. Open-mindedness and flexibility are also needed for researchers to be able to integrate relevant frameworks from other disciplines, spot new variables, adjust how we conceptualize and operationalize traditional variables, and adapt our research plans as necessary to fit the local context of the research situation at hand.

The research process can raise a host of dilemmas related to power dynamics, inequality, authority, authenticity, empowerment, advocacy, access, risk, and privacy. These complexities require us to recognize how research is socially embedded and to interrogate the consequences of engaging in research for both researchers and participants. We must consider the participants behind our data just as carefully as we consider the data themselves, seeking to understand throughout the research process how the questions we ask and the data we aim to collect to answer our questions have bearing on real life and real lives.

2 Ways of Observing

Studying the Interplay of Social and Linguistic Variation

Barbara M. Horvath

I first learned the importance of dialect variation when I moved at the age of 15 from a Catholic girls' high school in Providence, Rhode Island, to a large public high school in Los Angeles, California. Within days of starting school, kids I didn't even know were stopping me in the hallways and asking me to say "park the car." Most of the students, recent migrants from Missouri, Texas, and Oklahoma, had never heard anyone speak with such a strange accent. I stood out like a sore thumb, but I remember quite clearly making the decision that I would remain loyal to my roots and would never change the way I spoke. However, by the end of the school year, I sounded like and wanted to be a Californian. My parents, by contrast, were taken for Rhode Islanders for the rest of their lives. And repeating their experience, I am always recognized as an American by Australians despite having lived in Sydney for over 35 years.

This little bit of my linguistic history illustrates the importance of two facts about language that are central to sociolinguistics. A language can be strikingly variable in its pronunciation and can very quickly reveal something about the speaker that she or he may or may not want others to know. Migration, peer pressure, wanting to belong, stage in the life cycle, and changes in features like the pronunciation of post-vocalic /r/ are intimately intertwined.

Social and Linguistic

The inherent variability of language and the social and linguistic interpretations it makes possible form the basis of the discipline of sociolinguistics. Given such a general base, it will not be surprising to find that studies in sociolinguistics cover a wide variety of topics and are influenced by a number of relevant social science disciplines, primarily geography, sociology, and anthropology but also history and psychology. Initial interest in the relationship between linguistic variability and language change in progress was sparked by the iconic article by Weinreich, Labov, and Herzog (1968). What has become known as Labovian or variationist sociolinguistics grew out of dialect geography but differed in a number of ways: it was urban centered rather than rural, and it was more interested in where a dialect was heading than where it had been – hence speakers of all ages were part of the data sample, not just older people from small towns. The notion of "place," a geographical category referring to the embodiment of interacting sociocultural practices in a locality, remains central to most studies of linguistic variability, although the use of the concept of place in the explanation of linguistic variability is not often invoked (see Horvath & Horvath, 2001).

The Social Dimension

There are two ways to understand the influence of the other social sciences on the study of linguistic variation. The social sciences provide the descriptive categories that have been found to be important in the structure of society and that play important roles in the social distribution of wealth, education, power, and influence. Researchers trained primarily as linguists have tended to borrow the descriptive categories and practices of geography and sociology. For instance, from geography they borrow regional studies, maps, and the concept of place, and from sociology they borrow community studies, social survey methods, and social network analysis. Of most importance, sociolinguists have overwhelmingly used social class, gender, age, and ethnicity to capture the social structure of linguistic variation within a geographic locale. Labov's (1972) New York City study, the model still widely followed in dialect description today, was initially informed by sociology. Labov used the results of a sociological survey of New York City as a basis for selecting speakers for his study. Those very familiar social characteristics used by variationist researchers have been widely regarded as sufficient to reveal the social structure of linguistic variation and change. Milroy (1987) introduced social networking to sociolinguistic studies, and many researchers have followed in her footsteps. Speaker selection in network studies is based on selecting people who frequently communicate with each other; the groups generally consist of family, workmates, and friends.

Anthropology, on the other hand, does not so much influence the study of language variation as permeate it. Anthropology presents a more complex case than the other social sciences since the interests of anthropological linguistics overlap extensively with linguistics. Anthropological linguists are trained to study both linguistics and culture. When Labov was doing his work in New York, Dell Hymes was also exploring ways of extending the study of language from the writing of descriptive grammars to studying the social uses of language. Moreover, anthropology had developed its own ways of studying culture and society. Ethnography as a field method was developed in cultural anthropology for the collection of data in geographical and cultural contexts where the investigator was a complete outsider. What is of immediate relevance to data collection in sociolinguistics, however, is the difference it makes to a research design whether one takes an ethnographic approach (see Levon, Chapter 5) or a broader, survey-type approach (see Boberg, Chapter 8).

Anthropological linguistics entered into the study of linguistic variability with a broad understanding of social structure and an interest in linguistic variability that emerged from its own disciplinary interests in language. Ethnographies of indigenous peoples attempt to describe their social structure by discovering the social categories that are meaningful to them. It stands to reason that anthropologists would be interested in language variability from the point of view of discovering what meanings the variability has for its speakers. Unlike sociologists and other social scientists, anthropologists do not use preconceived social categories such as social class, gender, or even age unless those categories emerge as meaningful in the specific social context they want to describe. From a data collection perspective, an ethnographic approach would have a variationist sociolinguist first study the social structure of a community to discover the meaningful social groups and then collect the linguistic data that mark group membership. On the more linguistic side of sociolinguistics, the ethnography of communication introduced by Dell Hymes (see, for example, Hymes, 1964) focuses on ways of speaking or interacting in the speech community, encompassing how language is used in interaction as well as describing the social construction of interactions.

Even this glossing of the roots of sociolinguistics illustrates only some of the influences from the social sciences that have an impact on the selection of speakers to be included in a sociolinguistic study. All of the social sciences have played a role in creating the rich tapestry found in sociolinguistic studies today.

The Linguistic Dimension

I would venture to say that phonological, morphological, and syntactic features – in that order – are the linguistic phenomena most frequently studied by quantitative sociolinguists. The linguistic contexts that constrain variability can include the surrounding phonological environment, the morphological or syntactic status of the variable, or any other linguistic structures that are deemed to be relevant. The social constraints are associated with factors such as the speaker's social class, gender, age, and ethnicity. The method of data collection is generally interviews that are specially designed to simulate as closely as possible a relaxed conversational style, although many other approaches to data collection have been developed (as the chapters and vignettes in Part II of this volume demonstrate). The data consist of counts of the occurrence of the sociolinguistic variable, noting the constraining linguistic and social environments. The interviews themselves are generally not the object of scientific inquiry in variationist studies. Overwhelmingly, variationist studies use statistical methods in the analysis of the data, which set strict requirements on the nature of the speaker sample and the type of linguistic variable open to quantitative analysis.

Even though the variationist sociolinguistic interview is not usually studied per se, discourse certainly is of interest to sociolinguists. The discourse may be a particular type of genre, such as narratives, or may describe speech events or ways of verbally interacting associated with particular subgroups. The study of variability in discourse is associated, although not exclusively, with an ethnographic approach and with qualitative methods.

The divide between the quantitative and the qualitative is not just a simple matter of how to do sociolinguistics. The choice of whether to study phonological or morphosyntactic features or discourse is partially determined by the research question. The quantitative approach is primarily concerned with the question of language change in progress and asks how the phonological system accommodates variability and change. It assumes that the social structure of the variability, particularly the age/social class/gender structure, will provide the variability required to be able to observe the language change as it is progressing through the linguistic system and through the speech community. The quantitative approach often is used to show geographic variability in the spread of a sound change. The qualitative approach takes a speaker's position and asks what speakers mean (whether they are aware of it or not) when they use one variant sound or the other or whether a discourse type is appropriate to one social occasion but not another. This research question is best investigated by observing the linguistic variability in its "natural" setting – that is, in the interaction between members of the speech community.

From the beginning, researchers who trained in linguistics departments have tended to emphasize the phonological/morphosyntactic aspects of language and have used quantitative methods in their studies. Researchers trained in anthropological linguistics have been oriented to qualitative methods based on ethnography and on discourse. The social categories of importance to a variationist analysis, often a combination of social class, age, sex, and ethnicity, have been more or less taken to be uncomplicated and straightforward, at least in urban contexts. Anthropologically trained linguists have begged to differ and have brought to the sociolinguistics enterprise an appreciation of the complex nature of social categories and their relationship to linguistic variation.

Although we can divide sociolinguistic studies into two approaches, the quantitative and the qualitative, or what have historically been seen as the sociologically influenced and the anthropologically influenced, it is much more difficult to divide sociolinguists themselves along these lines – that is, by how they do sociolinguistics. The boundaries between variationist sociolinguistics and anthropological linguistics are seen more in the breach than in the observance, and we can find Labov being as well known for his more ethnographic work on narrative as for his more sociological study of phonological variation in New York City English or his geographically influenced account of linguistic variability in *The Atlas of North American English* (Labov, Ash, & Boberg, 2006). In fact, his study of the sociolinguistic variability in department stores associated with job status or job location (which floor one worked on) and his study of the social groups on Martha's Vineyard are early uses of an ethnographic method.

This bipartite division of sociolinguistics into the quantitative study of language change in progress and the qualitative study of the use and the meaning of linguistic variability in its social context is also not sufficient to describe all of the work sociolinguists do. It is possible to take both a quantitative approach to the description of a sound change in progress and a qualitative approach to the meaning of that variability to the speakers. The work that illustrates best the combining of quantitative and qualitative methods is Eckert's (1989) study of the Jocks and Burnouts in a Detroit high school. Through detailed ethnographic work she was able to identify two social groups: the Jocks, who subscribe to the ethos of the school, and the Burnouts, who are in opposition to it. Her linguistic analysis involved detailed quantitative phonological analysis of the vowel systems of these two groups. Her descriptions of the social structure and the linguistic structure of this local speech community are equally thorough.

The constant in all of this for the study of language variability is that studies in sociolinguistics are overwhelmingly empirical. The actual language of speakers – not the intuitions or casual observations of the investigator – constitutes the body of data that forms the basis of a sociolinguistic analysis.

Research Design: Quantitative, Qualitative, or Both?

The decision to use quantitative or qualitative methods in a research project is largely dictated by the research question and not insignificantly by the research tradition of the investigator. If the research question is about language change in progress and the question is primarily about how it is constrained by the linguistic system, then it makes sense to follow the quantitative approach. The social, as always, is important because variability cannot be observed outside of its social context. If, on the other hand, the research question is primarily about language use and/or identity, then an ethnologically based qualitative approach is more appropriate. The question may also take the researcher beyond these two, as when studying how the linguistic system constrains language change as well as how speakers create new meanings by using the potential of language to be variable. In this case, perhaps both quantitative and qualitative methods are in order.

The Quantitative Approach

Why is it that studies of language change in progress require a quantitative approach? The basic task of linguists of all kinds is to discover patterns in the linguistic data, and when it is change that is focused on, the patterns are subtle and can only be seen in the gradual increase over time in the use of the incoming sound or morphosyntactic feature. It is not a matter of the presence or absence of some linguistic feature, nor is it a case of observing a change that has been completed – like the Great Vowel Shift in English. For a number of reasons, phonological change is of primary importance to variationist sociolinguists. First of all, a phonological feature will be frequent in the data sample so the researcher can be confident of obtaining a sufficient number of instances from all speakers to be able to conduct quantitative analysis. Identification of the feature and its variants in the data is generally straightforward. There are well-known quantitative analytical techniques (e.g., Goldvarb, Rbrul, R) for handling data, and comparative studies of the same phenomena are likely, so that eventually studies can get beyond description and make significant generalizations about linguistic constraints on language change.

The general requirements for data collection in the quantitative approach are fairly well established but still open to insightful creativity: (1) select the variable linguistic features; (2) select speakers from significant social groups and from across the age spectrum; and (3) design a relevant interview. All of these requirements are of course more complex than these simple steps would suggest; however, there are many models to select as exemplars. Sometimes the researcher may be well acquainted with or may even be a member of the speech community so that questions of what features to study and what social groups to select speakers from will be more or less straightforward. However, this is often not the case. The possibilities for getting the information needed to begin the study in such a case are numerous. For example, Sylvie Dubois is a French Canadian who studies the Cajun speech community in Louisiana. She acquired the background information needed to design a quantitative study of language change in progress by first conducting a social survey in a number of locales around the state (Dubois & Melançon, 1977). She discovered not only relevant information about the language preferences of the Cajuns surveyed but also a great deal of social information, such as the prevalence of bilingualism and attitudes to both French and English. She developed a good understanding of the social forces at work in the Cajun community, which was later used in interpreting the social patterns of linguistic variation.

In my own case, I came to Sydney from the United States and learned as much as I could after arriving by observing and keeping records of the speech I heard around me, and I also profited by the work that had already been done on dialect variation in Australian English. I set out to use the tried and true social characteristics of social class, age, gender, and ethnicity but found only a very general sociological account of social class in Australia. I also encountered the widespread belief in Australian English studies that social class was not a factor in dialect variation. My approach was to use the basic occupation classes in the sociological account where I could, but also to report the difficulties of assigning social class status to the speakers in the sample. In addition, I used a quantitative technique introduced to me by David Sankoff, called principal components analysis, to delve further into the question of whether social class was in fact associated with variable pronunciation of English in Sydney (Horvath & Sankoff, 1987). In addition, most research on Australian English had avoided including newly arrived non-English-speaking migrants from Europe, but it seemed to me that the large numbers of migrants entering the speech community constituted a potential source of language change in progress. I included speakers who spoke English as a second language as well as the children of migrants who were either bilingual or spoke only English, and I worried about when "ethnicity" stopped being a relevant social category. Once again, the principal components analysis helped to show that these distinctions were important in explaining the patterns of linguistic variation that I found (Horvath, 1985).

The objective in quantitative sociolinguistic research design is to select linguistic features that are variably distributed in a speech community and that are constrained by both linguistic and social factors. It is also important that the selected features occur frequently so that in talking to a speaker for less than an hour, there will be sufficient data to analyze statistically. For instance, in my work, two features of Australian English, [f] or [v] substituted in *thorn* or *weather* and sentence-final *but*, were infrequent in the data and could not be studied quantitatively. The speaker sample must also be representative of the speech community, and there needs to be a sufficient number of speakers. Many researchers have taken five speakers for each combined set of social characteristics to be a minimal size for analysis. In my study of Sydney English, I included social class (working/middle), age (adult/teenager), gender (male/female), and ethnicity (Anglo-Celtic/Italian/Greek). A minimal sample size of 120 speakers (2 × 2 × 2 × 3 × 5) was required. The sociolinguistic interviews lasted between 40 and 60 minutes and were usually conducted in the speaker's home at a time convenient for them. Usually this meant during the evening after dinner for the adults, although some teenagers were available at more convenient times. The data collection process for this kind of urban sociolinguistic study clearly consumes a great deal of time, energy, and resources. There are many examples of much more constrained research designs where perhaps only one or two phonological or morphosyntactic variables are studied and the number of social constraints is limited by leaving out social class, or perhaps limiting ethnicity to only one group.

The Qualitative Approach

The sociolinguistic focus on ethnic and working-class dialects has meant that applied sociolinguistic research, such as that associated with the Center for Applied Linguistics, is of particular relevance to the teaching and learning of reading, to teacher–student interactions in the classroom, to standardized testing, and to educational and public policy. Therefore, under qualitative approaches I would include much of the applied sociolinguistic research done in educational settings and in advising on public policy. I was involved in doing a sociolinguistic evaluation of a reading test for young children in Washington, DC. I tape-recorded one-on-one sessions with African American third graders reading out loud short passages from a standardized reading test and responding to the questions about the passage. I then asked them why they had chosen a particular answer and, more often than not, they had a reasoned answer even if it was not the "correct" answer.

With respect to public policy, I was asked to write a short paper defining standard language so that a government-owned community language broadcaster in Australia would have an objective criterion for hiring only those presenters who spoke the standard language of their country of origin. After some negotiation, it was decided that I would instead take the responsibility for producing a short report on each of the languages under consideration, and the broadcaster could at least make decisions based on a better understanding of the language issues in a given country. Having started out with a policy of one country–one presenter, those responsible decided later that for some countries this policy was not viable. For instance, in the case of Yugoslavia (which was only one country at the time), the political and linguistic situation there meant that both Serbian and Croatian presenters would be required.

In qualitative sociolinguistic research, the social system is moved toward a more central position, and the relationship between linguistic variation and identity is prominent. Although all sociolinguists accept that linguistic variation is associated with identity, in qualitative research designs speakers are frequently studied because they share an identity more narrowly defined than the broad categories of social class or ethnicity.

Rather than selecting a sample of speakers who represent a speech community in a "place" like New York City, Sydney, or Pittsburgh, Pennsylvania, in a qualitative approach the speakers are more narrowly conceived as groups or types of speakers who share a social identity that distinguishes them from other speakers. Their membership in the group is reflected in the way they speak. When the ethnographic method is more strictly followed, the social divisions or groups that are meaningful to the members themselves are identified after a period of participant observation. The social identities can be locally defined, like the Jocks and the Burnouts in Eckert (1989) or the Homegirls in Mendoza-Denton (2008), or they can be more widely inclusive, like Texas women (Johnstone, 1999). The ethnographic approach can vary, and some researchers may already be closely involved in the community, or may even be members of the community. They may already have identified the subgroups they want to study. If not, then a great deal of time is required to observe patterns of social and interactional behavior in some locale, like a school or social club. A qualitative analysis removes the strict requirements set for quantitative studies such as the size and representativeness of the speaker sample. The speaker sample can be small, for instance only two speakers or a single individual representing a type or subgroup.

Whereas the quantitative approach was built on a well-developed descriptive and theoretical literature on phonology and morphology, the linguistic analysis of discourse was in its infancy at the time of Labov's narrative work. There was not then, nor is there now, an agreed-upon set of descriptive linguistic categories that can be coded, counted, and replicated with the same assurance available to phonological or morphological studies of variation. Labov's linguistic description of the structure of narratives followed the well-developed approach in descriptive linguistics in which recurrent patterns are observed in the narratives and then are given descriptive labels that can be confirmed, expanded, or contested in further research.

The data for qualitative sociolinguistic research are of widely diverse types, but labeling qualitative data as "language in use" perhaps captures a coherent element in the diversity. There is much more concern about revealing the social context under which the data were produced: who was speaking to whom; what was the setting; what was the relationship between the interlocutors; what roles in the group do the interlocutors have; and any other aspects of the occurrence of the utterances that are considered to be relevant to the analysis. The data for the analysis are often taken from interviews, much as in the quantitative approach. However, in qualitative studies, extracts that have been taken (and usually transcribed) from the recorded interviews or from conversations between speakers, one of whom may be the investigator, are provided as evidence for the linguistic claim being made. Holmes' (2012) Language in the Workplace project is a good example of such an approach.

Quantitative and Qualitative

There are a number of examples of sociolinguistic studies in which both quantitative and qualitative approaches are used. The fact that recorded interviews have been the most widely used form of data collection in sociolinguistics means that the data are available for phonological or morphological studies as well as for discourse studies. There are many ways to combine quantitative and qualitative approaches. Schilling-Estes (2004), for instance, uses one interview, with two speakers from two different ethnic backgrounds, taken from a large-scale study of Robeson County in North Carolina. She does a quantitative analysis of phonological and morphosyntactic features and then uses excerpts from the interview to qualitatively demonstrate how identity is constructed in the interaction between the two speakers.

Having studied Cajun English for a number of years, Sylvie Dubois and I (Dubois & Horvath, 2002) undertook a study of written and performed versions of that variety. We collected data from children's books by Cajun authors written both for a local readership and as souvenirs for tourists. One of the books is a retelling of the story of Little Red Riding Hood using Cajun English. We also obtained data from a commercial tape recording accompanied by a transcript of the telling of amusing anecdotes. The narrator was from Alabama and made a clear dialectal distinction between remarks addressed to the audience and the telling of tales he had heard from his grandfather, a Cajun from Louisiana. In telling the tales, he assumed the identity of his grandfather and spoke "Cajun English." We were then able to compare the "commercial" use of Cajun English with the features of Cajun English that we had described in our quantitative study.

As I mentioned earlier, Eckert's (1989) study and subsequent work stand out for their far-reaching potential for bringing together the study of the interplay of social structure and the structure of linguistic variation. The ethnographic approach to uncovering subgroups within a speech community whose identity is marked by their use of particular phonological variants combined with a quantitative analysis of those variants is an important step forward in sociolinguistics. Together with insights from geography, we can begin to set our sights not only on the possible origins of language change but also on its spread from highly differentiated subgroups to the whole speech community. *The Atlas of North American English* (Labov et al., 2006) marks the beginning of yet another approach to sociolinguistics.

Conclusion

What holds sociolinguistics together is the understanding that language is inherently variable and that variability is available to mark, for better or worse, social divisions in the speech community. Language variation and change cannot be empirically observed unless both linguistic and social structures are viewed at once. This connection is inseparable, and the understanding of how language works, including how it changes, requires both to be studied. When language is foregrounded, the subtle patterns of language change and the movement through the linguistic system are revealed. When the social is foregrounded, the subtle patterns of how a linguistic variant becomes a meaningful aspect of the identity of social groups are revealed.

References

Dubois, S., & Horvath, B. M. (2002). Sounding Cajun: The rhetorical use of dialect in speech and writing. *American Speech, 77*, 264–287.

Dubois, S., & Melançon, M. (1997). Cajun is dead – long live Cajun: Shifting from a linguistic to a cultural community. *Journal of Sociolinguistics, 1*, 63–93.

Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. New York: Teachers College Press.

Holmes, J. (2012). Language in the workplace. Retrieved from http://www.victoria.ac.nz/lals/lwp/resources/publications.aspx

Horvath, B. M. (1985). *Variation in Australian English: The sociolects of Sydney*. Cambridge: Cambridge University Press.

Horvath, B. M., & Horvath, R. J. (2001). A multilocality study of a sound change in progress: The case of /l/ vocalization in New Zealand and Australian English. *Language Variation and Change, 13*, 37–57.

Horvath, B. M., & Sankoff, D. (1987). Delimiting the Sydney speech community. *Language in Society, 16*, 179–204.

Hymes, D. H. (1964). Introduction: Toward ethnographies of communication. In J. J. Gumperz & D. H. Hymes (Eds.), *The ethnography of communication* (pp. 1–34). Washington, DC: American Anthropologist.

Johnstone, B. (1999). Uses of Southern-sounding speech by contemporary Texas women. *Journal of Sociolinguistics, 3*, 505–522.

Labov, W. (1972). *Language in the inner city: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, W., Ash, S., & Boberg, C. (2006). *Atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.

Mendoza-Denton, N. (2008). *Homegirls: Language and cultural practice among Latina youth gangs*. Malden, MA: Blackwell.

Milroy, L. (1987). *Language and social networks* (2nd ed.). London: Routledge.

Schilling-Estes, N. (2004). Constructing ethnicity in interaction. *Journal of Sociolinguistics, 8*, 163–195.

Weinreich, U., Labov, W., & Herzog, M. I. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics* (pp. 95–188). Austin: University of Texas Press.

Vignette 2a
Multidisciplinary Sociolinguistic Studies

Marcia Farr

Sociolinguistics itself originated as an *interdisciplinary* endeavor to study language in social context, combining either anthropology with linguistics (as done by Dell Hymes and John Gumperz) or sociology with linguistics (as done by William Labov, Joshua Fishman, Erving Goffman, and others). Interdisciplinarity, however, is different from *multidisciplinarity*: the former purposefully integrates methods and underlying assumptions from different disciplines to create a new discipline, whereas the latter uses methods or approaches from different disciplines to enhance research findings but does not attempt to create a new discipline out of the combination. An example of a multidisciplinary endeavor is the field of language socialization, which draws from both linguistic anthropology and cognitive psychology, but these disciplines remain separate.

My long-term study of transnational Mexican families (Farr, 2006) is multidisciplinary, although it is primarily grounded in linguistic anthropology, a field closely aligned with sociolinguistics. In fact, contemporary linguistic anthropology partially derives from the combining of anthropology and linguistics as led by Dell Hymes and John Gumperz during the time when sociolinguistics was also emerging; it also, of course, carries forward much from Franz Boas and early US anthropology. Given their close relationship and heritage, it should not be surprising that researchers recently have called for combining linguistic anthropology and sociolinguistics to do "sociocultural linguistics" (Bucholtz & Hall, 2008, p. 1). Indeed, sociolinguists increasingly use ethnography, and some linguistic anthropologists utilize sociolinguistic principles and patterns. Moreover, the move from analyzing language from interviews only to analyzing language as it is used in social networks (Farr, 2006; Milroy, 1987) and in communities of practice (Eckert & McConnell-Ginet, 1992; Meyerhoff, 2008) not only combines sociolinguistics and linguistic anthropology but also draws concepts from literary and cultural studies, as well as from social theory (see, for example, Coupland, Sarangi, & Candlin, 2001; Hanks, 2005). A key example of such borrowing is the concept of gender as performed and thus socially constructed (Butler, 1990), which led to extensive work on the linguistic construction of gender (Bucholtz, Liang, & Sutton, 1999; Eckert & McConnell-Ginet, 2003; Hall & Bucholtz, 1995). The notion of performative language now is used in analyzing the construction of all aspects of identity via the selective use of linguistic styles (Coupland, 2007; Eckert & Rickford, 2001; Jaspers, 2010). Other examples of borrowing from literary and cultural studies are Bakhtin's (1981) concepts of dialogism and heteroglossia, which have had similarly widespread influence (e.g., Tedlock & Mannheim, 1995).

Although my study of speech and identity among transnational Mexican families was framed as linguistic anthropology, my graduate education in sociolinguistics deeply influenced it. For example, I explored the rural Mexican dialect that the families spoke, and I included quantitative analysis, for example of levels of schooling and of the use of the informal and formal "you" pronouns, *tu* and *usted*. Since traditional sociolinguistic studies of language variation are based on standardized interviews and therefore are generally more structured than anthropological studies of language, my background in sociolinguistics implicitly guided me in structuring the (sometime) chaos of ethnographic field experience and fieldnotes, as well as recordings of daily speech. The simple expectation that there would be structure in the linguistic and cultural data led me to look for persistent patterns and themes and to organize them for analysis. A second aspect of sociolinguistics, the assumption of meaningful variation in language, also influenced my ethnographic research by providing a central understanding: that the language (and the culture) of the people I was studying would vary in patterned ways – for example, according to various aspects of speaker identities such as gender and age.

In my multidisciplinary work I also drew from history, sociology, and cultural studies. Social histories of both sites of the transnational community (Chicago and northwest Michoacán) provided historical depth to the contemporary ethnography, and historical studies of gender in Mexico (Stern, 1995) provided a conceptual framework for my discourse analyses that illuminated gender relations. Studies of migration and of race/ethnicity from sociology informed my indexical analysis of racial discourse. Finally, to provide a deeper and broader understanding of the *ranchero* identity of the families I also delved into the etymology of the words *rancho* and *ranchero* and relied on studies of Mexican cinema and music that focused on this subculture of the rural Mexican population. Such work provided a rich background for, and a deeper contextualization of, understandings from my own study.

Doing multidisciplinary work is not always easy. One must borrow ideas and concepts carefully, bearing in mind their philosophical assumptions. Without a background in the discipline borrowed from, a researcher may not fully understand the implicit contexts within which the concepts are embedded. Such problems can make some syntheses unworkable, leading to superficial results. Alternatively, researchers attempting to fully understand concepts from other disciplines can be overwhelmed with the amount of reading and work involved in using them appropriately. Finally, and more positively, sometimes such borrowed concepts are challenged by analyses of empirical sociolinguistic data, revealing their limitations. Erickson (2001), for example, cautions against taking the social theories of Bourdieu and others "too far," leaving insufficient room for speaker agency, local interactional contingencies, or cultural variation. Instead, he reverses the usual direction of borrowing and argues that sociolinguistics can contribute to the development of social theory, rather than just the other way around.

Although many textbooks assume that researchers always "design" their studies ahead of time, many studies evolve as the researcher becomes more grounded in the research context. In my experience, doing multidisciplinary research has required much flexibility, in terms of both time schedules and what and how I learn during the research process. Ethnography especially (as part of linguistic anthropology) can take the researcher into unexpected areas and findings, but that is its delight and its promise. For example, as a participant-observer, of necessity I gathered data according to my participants' schedules, not my own. Yet this patience and receptivity was richly rewarded with trust and therefore honest and candid language, continually reinforcing my own stance of openness and approachability. Being constrained to gather data according to the willingness and time frames of others can be demanding, to say the least, but the resulting linguistic data are culturally richer and more naturalistic than those gathered through interviews alone.

References

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. M. Holquist (Ed.). C. Emerson & M. Holquist (Trans.). Austin: University of Texas Press.

Bucholtz, M., & Hall, K. (2008). All of the above: New coalitions in sociocultural linguistics. *Journal of Sociolinguistics, 12*(4), 401–431.

Bucholtz, M., Liang, A. C., & Sutton, L. A. (Eds.). (1999). *Reinventing identities: The gendered self in discourse*. Oxford: Oxford University Press.

Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. New York: Routledge.

Coupland, N. (2007). *Style: Language variation and identity*. Cambridge: Cambridge University Press.

Coupland, N., Sarangi, S., & Candlin, C. N. (Eds.). (2001). *Sociolinguistics and social theory*. Harlow, UK: Longman, Pearson Education.

Eckert, P., & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology, 21*, 461–490.

Eckert, P., & McConnell-Ginet, S. (2003). *Language and gender*. Cambridge: Cambridge University Press.

Eckert, P., & Rickford, J. R. (Eds.). (2001). *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.

Erickson, F. (2001). Co-membership and wiggle room: Some implications of the study of talk for the development of social theory. In N. Coupland, S. Sarangi, & C. N. Candlin (Eds.), *Sociolinguistics and social theory*. Harlow, UK: Longman, Pearson Education.

Farr, M. (2006). *Rancheros in Chicagoacán: Language and identity in a transnational community*. Austin: University of Texas Press.

Hall, K., & Bucholtz, M. (Eds.). (1995). *Gender articulated: Language and the socially constructed self*. New York: Routledge.

Hanks, W. F. (2005). Pierre Bourdieu and the practices of language. *Annual Review of Anthropology, 34*, 67–83.

Jaspers, J. (2010). Style and styling. In N. H. Hornberger & S. L. McKay (Eds.), *Sociolinguistics and language education* (pp. 177–204). Bristol: Multilingual Matters.

Meyerhoff, M. (2008). Communities of practice. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *Handbook of language variation and change*. New York: John Wiley.

Milroy, L. (1987). *Language and social networks*. Oxford: Blackwell.

Stern, S. J. (1995). *The secret history of gender: Women, men, and power in late colonial Mexico*. Chapel Hill: University of North Carolina Press.

Tedlock, D., & Mannheim, B. (Eds.). (1995). *The dialogic emergence of culture*. Chicago: University of Illinois Press.

Vignette 2b
How to Uncover Linguistic Variables

Walt Wolfram

In the beginning was the linguistic variable. The heuristic utility of unifying a set of fluctuating linguistic variants within the structural construct of a linguistic variable and correlating the relative use of different variants with co-varying social and linguistic factors has now been reified in a full range of sociolinguistic studies. There is even now a canonical set of phonological and morphosyntactic variables that are traditionally investigated in variation studies, so that variables such as unstressed *-ing* fronting, syllable-coda cluster reduction, copula/auxiliary absence, and inflectional *-s* absence have earned honor-roll status as paradigmatic linguistic variables. At the same time, the search for linguistic variables cannot be limited to the replication and refinement of established variables if we seek to extend our understanding of systematic variation. Furthermore, variationists need to be sensitive to variation that is not as readily transparent and as convenient for coding and predetermined quantitative analysis as those examined in early variation studies.

Labov's timeless observation stands at the foundation of analyses based on the construct of the linguistic variable:

> [E]ven the simplest type of counting raises a number of subtle and difficult problems. The final decision as to what to count is actually the final solution to the problem at hand. This decision is approached only through a long series of exploratory maneuvers.
>
> (1969, p. 728)

While time-honored variables may serve to validate variationist sociolinguistics, it is important to understand that there are other variables that stretch the traditional variationist paradigm.

In this vignette, I focus on two quite different kinds of methodological challenges in uncovering variables: one readily quantifiable variable that illustrates extraction and coding challenges, and one that is a more elusive, rarely occurring form that presents a primary qualitative challenge for uncovering linguistic variables. The case of *a*-prefixing in Appalachian English (Wolfram & Christian, 1975; 1976) represents a variable that requires substantial preliminary manipulation to determine the parameters of systematic variation. In contrast, the case of the "call oneself" construction ([NP$_i$ *CALL* NP$_i$ V-*ING*]) as in *He calls himself* *dancing* represents a rare but subtly significant form that eluded attention for decades in the description of African American English (AAE).

Notwithstanding the refinements in methods of collecting spontaneous speech data over the decades, there is still a residue of problems in the extraction and analysis of data from natural conversation. The case of *a*-prefixing in items such as *He was a-huntin’ and a-fishin’* illustrates at least three kinds of issues that confront the description of systematic variation. First, there is the issue of semantic equivalency, an assumption underlying the specification of the variants of a variable. To capture authentic variability, we have to ask whether *a*-prefixed and non-*a*-prefixed forms mean the same thing or whether the *a*-prefix denotes a unique aspectual meaning (as originally postulated by Stewart, 1967). We can refer to this as the *equivalence issue*, since it must be assumed that all of the variants of a variable will be denotatively equal in order to meet the conditions of "inherent variability." Only after establishing and following an elaborate set of preliminary procedures to examine independently the meaning of *a*-prefixed and non-*a*-prefixed constructions (Wolfram & Christian, 1975) was it determined that *a*-prefixed and non-*a*-prefixed variants were semantically equivalent, though a later investigation of the stylistic-pragmatic significance of *a*-prefixing (Wolfram, 1988) indicated that there was, in fact, an "intensifying" reading associated with the use of the *a*-prefixed form that might distinguish its use from a non-*a*-prefixed form.

A second critical issue for examining systematic variation is the *countability issue* – that is, determining permissible structural contexts for variation. As with any variable, there are cases where variation may not occur for one reason or another, the so-called no-count cases (see Blake, 1997, for a similar discussion with respect to the copula). Notwithstanding Krapp’s (1925) assumption that "in popular speech, almost every word ending in *-ing* has a sort of prefix *a-*" (p. 286), there was an indication that there were *-ing* constructions in which the *a*-prefix was *not* permissible. Accordingly, these cases could not be counted as potential occurrences for the *a*-prefix. As it turns out, particular syntactic and phonological constructions need to be excluded from any count of systematic variation in *a*-prefixing because they are not eligible for the attachment of the *a*-prefix. These include nominal and adjectival uses of *-ing* (e.g., **He likes a-fishin’*; **The a-charmin’ dog entertained her*), post-prepositional position (e.g., **They made money by a-fishin’*), and initial unstressed syllables (e.g., **They were a-producin’ movies*) because of a universal phonetic constraint prohibiting successive unstressed syllables at the beginning of a word. Conclusions about the no-count cases for *a*-prefixing were established only through an extensive series of procedures that tapped speaker intuitions about the grammaticality of particular *a*-prefixing constructions (Wolfram, 1982).

Finally, there is the *performance issue*. Like other natural speech phenomena, the actual performance of vernacular speakers does not contain flawless discourse and grammatically well-formed utterances; it is full of hesitations, false starts, and other flawed production. As it turns out, the phonetic production of the *a*-prefix is a schwa [ə], phonetically quite similar to the filled, mid-central hesitation vowel. As we listened to utterances such as *That was [ə] interesting* we needed to use a full range of phonetic and discourse cues about performance to determine whether a particular case was a genuine case of *a*-prefixing or simply a filled hesitation phenomenon. The case of *a*-prefixing thus illustrates the range of theoretical-descriptive and practical data extraction procedures that have to be considered in applying the notion of the linguistic variable, but just about every linguistic variable we have investigated has encountered a comparable set of "subtle and difficult problems," as Labov noted, that attend the extraction and coding of data for variable analysis.

The case of the counterfactual *call oneself* V-*ing* construction, as in *He calls himself dancing* or *She calls herself acting*, presents a completely different kind of descriptive and methodological challenge. In this instance, we have a relatively obscure form that, to my knowledge, no sociolinguists had described in their early accounts of AAE. This construction rarely shows up in sociolinguistic interviews, and when it does it is not socially salient – either to speakers in the community or to sociolinguists describing the morphosyntax of AAE (Wolfram, 1994). To add to the challenge, the construction does not appear to be quantifiable in terms of the traditional methods of variation analysis. So, there is good reason that it went unnoticed for the first several decades of descriptive attention to AAE. I first noticed its use and started collecting examples from some of my African American colleagues and students at the University of the District of Columbia over more than a decade of everyday interactions during the 1980s. Its description as a camouflaged form relied on a series of structural elicitation tasks (Wolfram, 1994) that were constructed after I had collected a number of examples from ordinary conversations that were needed to complement my collection of examples. To confirm my emerging hypotheses, a set of African American and European American subjects were given a productive structural elicitation task (e.g., "Think of three ways you might complete the following sentence: *He calls himself*…") and a receptive co-indexing task in terms of a created scenario illustrated by the following:

> Suppose a woman who has a dog would like to think that the dog is listening to her when in fact he is not listening at all. Choose only one response to describe this situation.
>
> 1. Look at that, he calls himself listening.
> 2. Look at that, she calls him listening.

African American speakers overwhelmingly (18 out of 23) chose the co-referential response (*he calls himself listening*) while European American speakers' responses were a mirror image, choosing the non-referential response (*she calls him listening*) in 18 out of 23 cases. This differential response indicated a significant difference in the interpretive task by the two groups of respondents (Wolfram, 1994, p. 347). These results led to the descriptive conclusion that many African Americans were familiar with and productively used a structural extension of the *call oneself* constructions with a V-*ing* complement, whereas European Americans were not familiar with the [NP$_i$ *CALL* NP$_i$ V-*ING*] construction, even though they shared the counterfactual semantic-pragmatic reading of the form. This finding led not only to the inclusion of this construction in more recent descriptions of AAE (e.g., Green, 2002; Wolfram & Schilling-Estes, 2006) but also to the recognition of the notion of semantic camouflaging as an extension of Spears' (1982) original description of syntactic camouflaging.

The two cases presented in this vignette offer dramatically different instances of linguistic variables, including both quantitative and qualitative challenges in uncovering and describing linguistic variables. One is a relatively salient structure in which the challenge is to arrive at a valid account of its inherent, systematic variability; the other illustrates how elusive and subtle variables can sometimes be in the search for meaningful sociolinguistic variation. They both point to the need for adequate preliminary qualitative, structural data as the starting point for quantitative analysis. The important lesson to be learned from such explorations is how diverse and complex – and ultimately rewarding – such discoveries can be if we are open to and creative in our pursuit of meaningful sociolinguistic variation in the name of the linguistic variable.

References

Blake, R. (1997). Defining the envelope of linguistic variation: The case of "don't count" forms in the copula analysis of African American Vernacular English. *Language Variation and Change, 9*, 57–80.

Green, L. J. (2002). *African American English: A linguistic introduction*. Cambridge: Cambridge University Press.

Krapp, P. (1925). *The English language in America*. New York: Frederick Unger.

Labov, W. (1969). Contraction, deletion and inherent variability of the English copula. *Language, 45*, 715–762.

Spears, A. (1982). The Black English semi-auxiliary *come*. *Language, 58*, 850–872.

Stewart, W. A. (1967). *Language and communication problems in southern Appalachia*. Washington, DC: Center for Applied Linguistics.

Wolfram, W. (1982). Language knowledge and other dialects. *American Speech, 57*, 3–18.

Wolfram, W. (1988). Reconsidering the semantics of a-prefixing. *American Speech, 63*, 247–254.

Wolfram, W. (1994). On the sociolinguistic significance of obscure dialect structures: The [NPi *CALL* NPi V-*ING*] construction in African American Vernacular English. *American Speech, 69*, 339–359.

Wolfram, W., & Christian, D. (1975). *Sociolinguistic variables in Appalachian dialects*. (Final Report, National Institute of Education Grant No. G-74-0026).

Wolfram, W., & Christian, D. (1976). *Appalachian speech*. Arlington, VA: Center for Applied Linguistics.

Wolfram, W., & Schilling-Estes, N. (2006). *American English: Dialects and variation*. Malden, MA: Blackwell.

Vignette 2c
How to Uncover Social Variables

A Focus on Clans

James N. Stanford

> "By the way, my mother's dialect is not quite the same as mine," a Sui friend casually mentioned to me. "That's how it is for most children growing up in Sui villages."

Little did I know that this tidbit of information would eventually develop into my dissertation and then further research on clan as a sociolinguistic variable. But I first had to become engaged in the Sui community, let go of prior assumptions, and learn from the perspectives of cultural insiders. After all, the purpose of data collection is to gain new knowledge, and sometimes that new knowledge involves new variables. If we design our data collection on the basis of old knowledge, we may miss the chance to uncover meaningful variables.

I arrived on the field trained in a set of classic, time-tested principles about how language varies with respect to socioeconomic stratification, age, gender, ethnicity, geographic region, and other factors. Many of those principles are primarily based on major world languages or well-known minority languages and varieties, and they tend to be studied in the context of Western industrialized societies (e.g., Labov, 1972; 1994; 2001; Trudgill, 1974). Other research settings may be dramatically different, such as indigenous minority languages (Stanford & Preston, 2009). In the rural Sui villages of southwest China, for example, clan is a crucial social variable that far outweighs socioeconomic status. Likewise, Sui research on gender, regional variation, child dialect acquisition, and other topics would be incomplete without considering clans.

Entering the Sui World

The indigenous Sui people of Guizhou Province, China, follow a strict custom of clan exogamy: husbands and wives must not be members of the same clan, and the wife moves permanently to the husband's village at the time of marriage. As a result, these patrilineal villages are complex sociolinguistic environments involving a wide range of clans, many of which have distinctive dialect contrasts. Though mutually intelligible, Sui clan-level dialects have striking linguistic contrasts that include lexical variants (such as high-frequency words like the first singular pronoun), diphthong variants, and tone variants. Even so, each village has a single dominant dialect: the dialect of the men, unmarried women, teenagers, and older children. Married women maintain the dialect features of their original home villages to a high degree, even after decades in the husband's village (Stanford, 2008a).

In this way, Sui speakers make use of dialect features to reflect and construct their loyalties to clans. The clans can be viewed as "communities of descent" (Stanford, 2009b): social groups that are constructed around local notions of shared lineage and often perceived as being rooted in the ancient origins of the society. Lifelong membership in a community of descent is socially assigned at birth according to local understanding of each infant's lineage, and so these communities are best described from an emic sense of ancestral descent. By contrast, the notion of "community of practice" is typically applied in situations that have a much stronger sense of emergent, evolving, negotiable identities and memberships, rather than ancestral lineage (Eckert & McConnell-Ginet, 1992; Meyerhoff, 2002; Wenger, 1998). Yet language has an important constitutive role in both cases. In-married Sui women express a strong sense of stable, lifelong loyalty to their communities of descent despite being separated from their home villages, and this is clearly evident in their dialect choices.

The role of clan as a social variable is also important in Sui child dialect acquisition. Prior work in other societies has suggested that children acquire the dialect features of their peers rather than their parents (Labov, 1972, p. 304). But in Sui society, the key distinction lies along clan lines, not parent-versus-peer. As a result of patrilineal clan exogamy, most Sui children are raised in households where the mother speaks an "outsider dialect" (matrilect), while the father and older siblings speak the local dialect (patrilect). Sui children rapidly learn to distinguish these clan-related dialects; young children may speak a mix of matrilect and patrilect, but older children and teenagers are almost fully patrilectal (Stanford, 2008b). A child's most important dialect choices are therefore focused on clan distinctions.

Uncovering Social Variables

The Sui examples above illustrate the value of seeking local perspectives from the very beginning of a project – before data collection begins. Johnstone (2004) suggests that we seek "the local knowledge that motivates and explains the behavior of a particular group" (p. 76). This perspective is clearly important when studying a culture like that of the Sui, and it is just as important when investigating one's own culture. Here are a few suggestions to help uncover locally meaningful social variables:

1. *Be engaged with the community and personally involved in local life as much as possible.* Be a participant-observer, not a detached scholar. Make genuine friendships, join local activities, and find a way to make a positive contribution as a stakeholder in the community. For me, this involved building long-term friendships with Sui families and assisting in educational and development projects in the village. In this way, ethnography becomes a natural part of community participation, rather than an afterthought or supplement to the "main" research. Rather than viewing these social activities as a task that must

be completed so that the "real linguistic research" can begin, recognize that personal interactions are a key part of the research itself. Without such experiences and knowledge, my sociolinguistic data and understanding of Sui would have been inaccurate. From the beginning, have the goal of engaging with the community and understanding local social meanings:

> Variationists who are interested in the local meanings of variation have to be willing to *start* with ethnography, using ethnographic research methods to decide what the possible explanatory variables might be in the first place, rather than starting with predefined (and presumably universally relevant) variables and bringing in ethnography only to explain surprising findings or statistical outliers.
>
> (Johnstone, 2004, p. 76)

2. *Let go of prior assumptions*. For example, at first I wrongly assumed that the boundary lines on Chinese maps of the Sui region, such as townships, counties, and other administrative entities, were central to the Sui experience of place. I later found that indigenous Sui notions of place do not match those administrative boundaries. The Sui notion of place – and dialect features – is constructed by local understanding of clans, indigenous toponyms, and surnames (Stanford, 2009a; 2009b).
3. *Depend on the insights of cultural insiders*. Have the attitude of a learner, not an "outside expert." Rather than simply collecting linguistic data from the language consultants, learn from them. They have a lifetime of knowledge of the community.

At the same time, it is good to remember that local consultants (like all of us) may overlook the significance of some of their own behavior, viewing it as simple "common sense." For example, when I asked why different Sui people within the same village use different first singular pronouns, some respondents thought it was a foolish question: "Each speaks their own way [of course] … People surnamed Lu speak like the Lu place. We people surnamed Pan speak like people surnamed Pan [even though we live here in the Lu place]" (Stanford, 2009b, p. 295). Overlooking the significance of everyday behavior can also occur when the researcher is a cultural insider studying her or his own community. Therefore, it is always wise to gain perspectives from a wide variety of community members of various ages and diverse backgrounds.

Cultural insiders can also produce valuable moments of performance speech (Schilling-Estes, 1998). For example, if someone mentions that another person "speaks differently," the researcher might ask, "Can you imitate that speaker for me or describe the way that person talks?" Even though performed speech features are typically exaggerated, they often contain valuable clues for later research on natural speech. Those performances are especially valuable in the early research stage when some of the basic social and linguistic variables are unknown. But note that the researcher should first determine whether such activities are culturally appropriate for the specific community.

"To See What Is in Front of One's Nose"

When my Sui friend first told me about his mother's dialect, it was the initial step in a long but fascinating process of learning about the linguistic complexity of Sui society. During that process, it was necessary to become involved in the community, let go of prior assumptions, and depend on cultural insiders. In this way, I was eventually able to make progress in uncovering locally meaningful social variables.

George Orwell once wrote, "To see what is in front of one's nose needs a constant struggle." For sociolinguists and the communities we study, this struggle involves seeing past our own familiar ways of thinking and perceiving the world. This is certainly important for Sui research. It might seem less important when the research site is more similar to our own culture, but actually the underlying challenges and opportunities are the same. If we can see beyond our noses, we should be able to gain new sociolinguistic insights about any particular community and about human society in general.

References

Eckert, P., & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as a community-based practice. *Annual Review of Anthropology, 21*, 461–490.

Johnstone, B. (2004). Place, globalization, and linguistic variation. In C. Fought (Ed.), *Sociolinguistic variation: Critical reflections* (pp. 65–83). New York: Oxford University Press.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1994). *Principles of linguistic change: Internal factors*. Malden, MA: Blackwell.

Labov, W. (2001). *Principles of linguistic change: Social factors*. Malden, MA: Blackwell.

Meyerhoff, M. (2002). Communities of practice. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 526–548). Malden, MA: Blackwell.

Schilling-Estes, N. (1998). Investigating "self-conscious" speech: The performance register in Ocracoke English. *Language in Society, 27*, 53–83.

Stanford, J. N. (2008a). A sociotonetic analysis of Sui dialect contact. *Language Variation and Change, 20*(3), 409–450.

Stanford, J. N. (2008b). Child dialect acquisition: New perspectives on parent/peer influence. *Journal of Sociolinguistics, 12*(5), 567–596.

Stanford, J. N. (2009a). Clan as a sociolinguistic variable. In J. N. Stanford & D. R. Preston (Eds.), *Variation in indigenous minority languages* (pp. 463–484). Amsterdam: John Benjamins.

Stanford, J. N. (2009b). "Eating the food of our place": Sociolinguistic loyalties in multidialectal Sui villages. *Language in Society, 38*(3), 287–309.

Stanford, J. N., & Preston, D. R. (Eds.). (2009). *Variation in indigenous minority languages*. Amsterdam: John Benjamins.

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.

Vignette 2d
How to Uncover Social Variables

A Focus on Social Class

Rania Habib

While social variables are critical for any sociolinguistic study, their inclusion or exclusion as well as uncovering their meaning and significance can sometimes be challenging. For example, social class has been widely investigated in sociolinguistic studies, as it is one of the social variables that often play a significant role in the linguistic choices that speakers make. However, is the application of social class grouping always necessary or possible? How can we tell whether we are grouping people accurately into specific social classes?

The difficulty of grouping speakers into social classes has led some researchers to adopt different methods of grouping such as social networks (Milroy, 1987), life-modes (Højrup, 1983), and communities of practice (Eckert and McConnell-Ginet, 1992). In countries such as the United States and the United Kingdom, sociolinguists have often relied upon well-defined sociological indices for social classes that are based on three main socioeconomic indicators: education, occupation, and income (Labov, 1966; Wolfram, 1969; Trudgill, 1974). In some countries, for example Syria and Egypt, such indices are not readily available to researchers, but this challenge did not preclude some researchers from using similar socioeconomic indicators to classify speakers into social classes (see Haeri, 1996; Habib, 2010a; 2011a).

When I conducted my own sociolinguistic fieldwork in an Arabic-speaking community, I found out first-hand, using sociological methods of testing social class indictors, about the complexities of social class division and how such division may differ from one place to another and should be based on different indicators (Habib, 2010b). In some cases, social class may not be apparent or may not matter. Through knowledge of the community and experience, the researcher should be able to discover during his or her fieldwork whether social class is an influential variable or not.

Rural Social Uniformity vs. Urban Social Plurality

In one of my research projects (Habib, 2011b), I investigated the spread of urban sounds and the variation and change of the variables (q) and (e) in child and adolescent language in the village of Oyoun Al-Wadi in Syria. In this study, I determined that social class should not be included as a social factor in the analysis. During data collection, I observed that class differences among children are unnoticeable. The strong social ties among the village's people diminish the socioeconomic differences. A child whose father owns a major restaurant, a beautiful house, and other property in the village is a very good friend of the son of a baker. All the schoolchildren, regardless of their parents' incomes, go to school in taxis in the winter months. All children go to the same public schools and participate in the same events in the village.

In small communities, such as Oyoun Al-Wadi, researchers may have to take different approaches to social class because although people may vary socioeconomically, they live in a small area, are all in contact with each other, and may be related to each other. They follow the same rhythm of life. Most of the people who live in Oyoun Al-Wadi have worked overseas and accumulated some wealth or have relatives working overseas and supporting them financially. Very few families have fewer resources. Even those families seem to copy the lifestyle of those with more resources so they do not appear different, as in the case of sending their children to school by taxi despite the short distances in the village. They all participate in a local funeral. They all are invited to the same wedding. Thus, linguistic differences may not depend on their socioeconomic status as much as they may depend on the people they are in constant contact with, on the social identity they intend to reflect, or on the strength of attachment to their locale. For example, a fifth grader told me that his mother changed her pronunciation toward the urban forms after using the rural forms throughout her life because her husband's restaurant business required her to be in contact with many people from urban centers. She developed friendships with many women from outside the village and thus changed her speech to sound urbane. The case of his uncles' wives is similar. His sister mentioned that she is amazed when she hears her mother and her uncles' wives switch suddenly from urban to rural when they are having a closed conversation away from strangers. She did not know how they could do it.

Because I noticed the absence of clear social class distinction in the community, I focused on other social factors, such as age, gender, residential area, mother's origin, and degree of contact with urban features through various quantitative measures. I also included in my research design, besides quantitative analyses, an in-depth ethnographic investigation of the social meanings of the variables under investigation to reveal the hidden aspects of the observed variations and their relation to the identities the speakers intend to adopt or reflect. Thus, I observed closely the community's and the participants' attitudes toward the urban and rural features when they are used by boys/men and girls/women.

Social Class Identification

In my (2010a) and (2011a) studies of the variable use of the voiceless uvular stop [q] and the glottal stop [ʔ] in the speech of rural migrants to the city of Hims in Syria, social class was included as a variable, following Western methodology, more specifically Labovian methods and social variables. However, a question emerged regarding how to classify speakers according to social class, as there were no commonly used indices to refer to, unlike those used in the United

States and the United Kindom. My profound knowledge of the community and how its members regard each other helped me classify speakers into lower-middle and upper-middle classes. I looked at the social criteria that are usually indicative of one's social status. The appearance of speakers, the way they dress, their lifestyles, what they own, how they are talked about in the community, and the people they most often associate or socialize with are some of the criteria that allowed me to determine their social class assignment. Furthermore, during the interviews I asked questions about how they would socially regard or classify certain persons. Their responses supported and confirmed my intuitive social class classifications.

However, it was important to examine this classification of speakers into two classes based on the general view of the community. Tests of the strength of association between social class and its indicators showed that social class identifiers in this rural migrant Arabic-speaking community are indeed different from those in the West (Habib, 2010b). While education, occupation, and income play major roles in social class identification in the West, income and residential area emerged as major class identifiers; occupation played a minor role only, regarding the government employees' category; and education played no role. In addition, social class as a variable did not necessarily show the same influence as individual social class indicators did, when I examined their effects on the linguistic variables under investigation. For instance, while social class emerged as statistically *insignificant* regarding the variable use of [q] and [ʔ], the strongly associated social class indicator, residential area, emerged as statistically *significant*. These findings resulted from including these indicators as separate variables in a multivariate test to explore their main effects on the variable use of [q] and [ʔ]. The findings speak to the need to explore the influence of social factors such as education, occupation, income, and residential area on linguistic variation, not to test the influence of social class as a composite variable alone.

Conclusion

Sociolinguists must consider different options when it comes to deciding whether or not to include social class in their analyses, as social class may differ from one country to another and from one community to another. While urban centers, even in Arab countries, may have more defined social classes, rural areas may lack such clear division. Furthermore, while social variables in general are highly important in variationist studies, during analysis one may discover that some social variables in particular play no role in the variation in question. In this case, the researcher has to look for other factors and methods to explain the variation. While social variables may appear important initially, sometimes the data may inform us otherwise.

References

Eckert, P., & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology, 21*, 461–490.

Habib, R. (2010a). Rural migration and language variation in Hims, Syria. *SKY Journal of Linguistics, 23*, 61–99.

Habib, R. (2010b). Towards determining social class in Arabic-speaking communities and implications for linguistic variation. *Sociolinguistic Studies, 4*(1), 175–200.

Habib, R. (2011a). New model for bilingual minds in sociolinguistic variation situations: Interacting social and linguistic constraints. *International Journal of Psychology Research, 6*(6), 707–760.

Habib, R. (2011b). Meaningful variation and bidirectional change in rural child and adolescent language. *University of Pennsylvania Working Papers in Linguistics, 17*(2), 81–90.

Haeri, N. (1996). *The sociolinguistic market of Cairo: Gender, class, and education*. London: Kegan Paul International.

Højrup, T. (1983). The concept of life-mode: A form-specifying mode of analysis applied to contemporary Western Europe. *Ethnologia Scandinavia*, pp. 15–50.

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

Milroy, L. (1987). *Language and social networks* (2nd ed.). London: Routledge.

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Washington, DC: Center for Applied Linguistics.

3 Social Ethics for Sociolinguistics

Sara Trechter

Through its focus on language, a social phenomenon, consideration of ethics in linguistic research is necessarily framed within a language community and consistently foregrounds issues of shared rights, obligations, and responsibilities. For sociolinguistics (the subdiscipline focused on the dialogic construction of the self, other, and society through ideology that is created, performed, and manifested through social interaction), attention to personal and community obligations is even more essential. Such obligations naturally lend themselves to a research process that is sometimes outside "scientific" norms. Rather than glorying in the complexity that their field of research engenders, some sociolinguists have regarded it as particularly problematic because it invokes the object of study (the language of speakers), the research subjects, and also the voice or presence of the researcher. Labov (1972) has famously referred to this situation as the "observer's paradox," in which "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation" (p. 209). Assuming a positivist or realist paradigm in which the object of study must be an authentic vernacular that is corrupted by observation has led to a number of methodological machinations, which are sometimes less than ethical and sometimes more. These include surreptitious recording, requests for emotional stories such as the "danger of death narrative," the assumption of a "natural" community role to obtain data deceitfully, employing community members for elicitation, training community members in linguistics, etc.

Recent introductions to the methodological approaches to sociolinguistics have continued to include discussions of the ethical problems engendered by the positivist approach and reification of the "observer's paradox" (Ball, 2005; Coupland & Jaworski, 1997; Mesthrie, Swann, Deumert, & Leap, 2000). These discussions address the common ethical concerns such as those listed above and accompanying methodological concerns, and they give appropriate advice to the novice researcher or student whose initial approach to sociolinguistics is based in the received patterns of quantitative, variationist work. Rather than questioning the underlying theoretical assumptions, an explicit line is drawn between the scientific enterprise and the future ethical obligations of the linguist engendered by research interactions. The implicit recommendation is that she or he *should*, but may or may not choose to, use her or his knowledge of the linguistic rules and social context to advocate for the speakers of a vernacular, following the example of previous well-known advocates (Labov, 1982; Rickford, 1997; Smitherman-Donaldson, 1988).

Often, researchers view language communities from a theoretical viewpoint that is ideal, seeking "authentic" vernacular and variation, where any outside observation or interaction, especially with a researcher, is considered unnatural interference. This view contrasts with a less objectivized and idealized assumption: that within any community, the reflection on and observation of language and interaction with outsiders is somewhat normal. The tension between these two views underlies the themes of ethical research explored in this chapter. At what point does an observer become a quasi community member? In terms of ethical obligations, an approach that assumes that the observer is causing unnatural interference is less likely to allow research participants to assume different roles and degrees of agency. Yet assuming that observation and participation by an outsider in the community is a normal part of human language behavior is less likely to produce a familiar or academically recognizable scientific research plan from the viewpoint of academic stakeholders: granting agencies or college ethics committees (Institutional Review Boards – IRBs – in the United States). Discussion around sociolinguistic fieldwork ethics therefore continues to include: (1) the problem of any sociolinguistic research, which necessarily invokes the object of study and participants in the research process; (2) the presupposed separation of the roles and goals of researcher from those of the speech community participants once she or he has already engaged within the community, and the concomitant role of ethics committees in enforcing these power relations; and (3) the place of the researcher not as part of normal community interaction but as an outsider with rights and obligations.

Background

The consideration of ethics as an essential part of the research plan in linguistics has gained an unusual, though not unwarranted, amount of attention in the past 20 years. Aware of the increasing difficulty involved in explaining linguistic research to ethics boards, and building on postmodern anthropological discussions that encourage reflexivity in the research process and an understanding of the effect that the researcher has in her representation of others (Kroskrity & Field, 2009), the Linguistic Society of America vetted and produced (2006–2009) an official statement concerning research ethics and created a permanent committee on Ethics in Linguistics (Linguistic Society of America, 2009). Though begun 35 years subsequent to its sister field's code, the LSA Ethics Statement draws broadly on the 1998 Code of Ethics of the American Anthropological Association (revised from 1971 and 1986) and also on the 1988 Statement of Ethics of the American Folklore Society. Like these codes, the LSA Ethics Statement broadly emphasizes the obligations of members of the linguistics profession to various groups and principles (which I have numbered here to match the LSA Statement): (6) to the public, in terms of accessibility and social and political implications of research; (5) to professional standards of honesty in scholarship;

(4) to students and colleagues, in terms of respect and attributing their contributions to scholarship; (3) to the needs and desires of language communities, especially where the community has an investment in language research; and (2) to the protection of individual research participants. Although the preface to the statement acknowledges its breadth and that it may conflict with other statements, and part 1 acknowledges that every field situation is different, it does not acknowledge that the general recommendations and responsibilities to research communities and the profession within the statement may conflict with each other in the reality of actual research.

Assuming that the linguist has obtained research rights as an autonomous agent, each of the areas above delineates specific obligations incurred in the research enterprise to the protection of the rights of human subjects and a responsibility for fairness and accuracy when dealing with the profession and its members. Item 3 above proposes additional responsibilities to linguistic communities. Such obligations, according to Garner, Raschka, and Sercombe (2006), "highlight the tension between public ethics concerning major social issues, such as the legal rights of minorities, and individual ethics, which relate to issues of professional responsibility and personal conscience" (p. 62). These responsibilities force an individual to grapple with the questions of who undertakes research and in whose interest, who the research belongs to, who writes and gets credit for authorship, how public the findings are, and what effect and status they have (Davies, 1999).

In fact, as a member of the LSA ad hoc committee charged with drafting and vetting the LSA Ethics Statement, I found that the areas that required the most careful language and that were the most difficult to negotiate with fellow linguists were ethical obligations to the "other": the individuals and communities whom we treat as research subjects. Indeed, within the field of linguistics, the less powerful and more "other" a community or individual is perceived as being, the more attention linguists tend to pay to the possible ethical ramifications within that community. Thus, the largest body of work on ethical concerns in linguistics deliberates either field research on minority languages or sociolinguistic research on minority populations, often within or in connection to a majority language educational system. In fact, these are both the focus of a thoughtful 2006 issue of the *Journal of Multilingual and Multicultural Development*.

Although this bent in the anthropology and linguistics literature may appear to be just more emphasis on the "other" in an effort to understand and position ourselves, consideration of a variety of communities with a focus on their differences is not necessarily without practical motivation. In particular, academics from native communities have a long-standing commitment to self-determination and cultural sovereignty (Champagne & Goldberg, 2005; Deloria, 1969). This commitment results in a literature that has repeatedly exposed the naïveté of academic researchers from the cultural majority who come to a minority, impoverished community, often with a "noble" goal of advocacy, but with little real knowledge and experience of the historical and cultural complexities providing context for the research. Their vague assurances of progress and frequent subsequent failure to provide tangible benefits to the language community create even more cynicism regarding the academic enterprise and the continued presence of researchers (see Trechter, 1999). Yet most linguists, if they acknowledge the complexity of different on-site contexts, often do so only in passing. Bowern (2008), for instance, warns linguists of the culturally grounded ethical complexity of different situations:

> I have talked about ethics as though there is just one ethical way to behave in research, and one system to satisfy. That is not true. Ethics are strongly a function of culture, and what may be considered ethical in one community would be unethical in another.
>
> (p. 150)

This is an excellent warning, but it is unclear whether Bowern is implying that general moral principles (normative ethics) are different from situation to situation or if their application differs according to culture.

Case Studies

Linguists typically approach ethics from an applied standpoint, informed by a vague sense of social justice, but we do not often draw on the philosophical standards in the ethical literature to articulate the normative reasoning (general moral principles) underlying our approaches. Without a shared understanding of the basis for ethical reasoning, we are likely to have difficulty arguing for the benefits of one approach over another. Instead, we have avoided philosophical argumentation and, particularly in the United States, focused on application without shared reasoning. Indeed, circumstantial variation has led sociolinguists to focus on very different recommendations when considering research ethics, and it is worth briefly examining three case studies.

First, Rickford (1997) undertakes a description of the reciprocal obligations that linguists have incurred to the African American speech community after years of research and contributions by this community to the study of linguistics. Despite some prominent examples of advocacy for specific communities, Rickford convincingly argues that very little has been done in terms of producing a new generation of African American students and faculty within the field of linguistics (cf. LSA Ethics Statement, pp. 3–5, and item 5 above).

In addition, research in disparate African American communities (including, but not limited to, research from the field of linguistics) has too often centered on analyzing the activities and behaviors of street-wise, often gang-affiliated males and even highlighting their sexual exploits. While perhaps "sexy" or "exotic" to the outsider academic consumer, these biased data have skewed the picture of language variation by gender and social class within the larger African American linguistic community (Morgan, 1994). Such a mistaken focus on the search for the most authentic vernacular speaker can lead to potentially racist or otherwise inaccurate misrepresentations of community variation and diversity (cf. LSA Ethics Statement, pp. 3–4, and items 3 and 6 above).

Rickford ultimately reminds us of the injunctions of a number of senior researchers, including Labov (1982), who states that we should give back by using our skills as linguists or by helping in community projects:

> An investigator who has obtained linguistic data from members of a speech community has an obligation to use the knowledge based on that data for the benefit of the community, when it has need of it. Perhaps as a start we might demand from ourselves and our students one hour of community service or applied work for every hour of tape collected, or every hour spent on theoretical and descriptive issues.
>
> (p. 173)

In a second context, Wolfram, Reaser, and Vaughn (2008) invoke Wolfram's "principle of linguistic gratuity" (1998) to communities, particularly in the context of the Southern US dialect communities where they work. They summarize a variety of contributions, including school curricula, documentary film projects, museum events, and inclusion of local communities in the design of and input for dictionary projects. Such work is fraught with both practical and principled concerns. For example, representing Appalachian voices of the community in the documentary *Mountain Talk* without expert commentary or correction demonstrates the potential conflict between scientific representation and "empowerment" of speaker communities (pp. 15–16). Community ideologies concerning language, dialect, and their historical roots are well represented, but may be gratuitous to a linguist or granting agency focused on scientific accuracy.

The ethical considerations invoked in both of these contexts arise out of community needs. Rickford focuses on the African American community at large, and Wolfram considers community recognition and respect at a regional level within different Southern dialect communities. Wolfram et al. also note that in some instances, as with the Lumbee Indians, linguistic recognition may affect federal recognition of their tribal status.

In a third and different context of Francophone Canadians, Heller (1999) sees the role of the ethnographic sociolinguist who recognizes local language use and ideology as being inextricably tied to the changing world of linguistic minorities: where a minority language was once defined in juxtaposition to nationalism, it is now defined in the context of hypermodernity, global capitalism, and world language alliances. Her detailed study of a Francophone Toronto high school considers the sociolinguistic detail of this microcosm associated with broader sociopolitical matters, such as transnational trends and economic commodification of language. The ethics of her research, though not discussed explicitly as such, are situated within the broader political context and within considerations of social justice.

Broader Ethical Models

Although researchers may focus on ethical issues as they perceive them at the forefront of their own research sites, there are also broader, applied ethical models that may be useful for linguists to consider before going into the field or beginning their own individual research projects.

In their oft-cited work *Researching Language*, Cameron, Frazer, Harvey, Rampton, and Richardson (1992; 1997) offer a generalized ethical paradigm that is not context dependent, recognizing that both the research agenda and the kind of ethical research that one can engage in is governed by the project and the participants at any site. In contrast to positivism or postmodern relativism, they argue for a realist approach to research, one which maintains that reality stands outside both the observer and the ideology of the observed and therefore is difficult to describe definitively. This stance underlies their ethical model, which because of its conceptual influence is worth considering in some detail.

Cameron et al. (1992) recognize three broad types of ethical research, which are not mutually exclusive within any research context. These are ethical, advocacy, and empowering – research *on*, *for*, and *with* the human subjects and/or their community, respectively:

> In ethical research … there is a wholly proper concern to minimize damage and offset inconvenience to the researched, and to acknowledge their contributions.… Human subjects deserve special ethical consideration, but they no more set the researcher's agenda than the bottle of sulphuric acid sets the chemist's agenda.
>
> (pp. 14–15)

> [T]he "advocacy position" is characterised by a commitment on the part of the researcher not just to do research *on* subjects but research *on and for* subjects. Such a commitment formalises what is actually a rather common development in field situations, where a researcher is asked to use her skills or her authority as an "expert" to defend subjects' interests, getting involved in their campaigns for healthcare or education, cultural autonomy or political and land rights, and speaking on their behalf.
>
> (p. 15)

> We understand "empowering research" as research *on*, *for* and *with*. One of the things we take that additional "with" to imply is the use of interactive or dialogic research methods, as opposed to the distancing or objectifying strategies positivists are constrained to use. It is the centrality of interaction "with" the researched that enables research to be empowering in our sense; though we understand this as a necessary rather than a sufficient condition … we [propose three] programmatic statement[s] …:
>
> (a) Persons are not objects and should not be treated as objects.
> (b) Subjects have their own agendas and research should try to address them.
> (c) If knowledge is worth having, it is worth sharing.
>
> (pp. 22–24)

Although Cameron et al. (1992) oversimplify the reality of the role that human subjects play in determining the agenda of all research, including ethical research, and despite the difficulties of doing any sociolinguistic or ethnographic work without researching *with* human subjects, these three broad categories have provided a useful tool for linguists to process their own ethical strategies. Invoking the notion of "empowering" research, both Rickford and Wolfram describe their efforts to give back to communities as fulfilling a promise and trying to address their subjects' own agendas, as with the efforts of Wolfram et al. (2008) to render the knowledge gained from the research process publicly accessible through dictionaries, documentaries, school curricula, and museum projects. Cameron et al. (1997) also discuss Rampton's use of feedback techniques in which information was shared with the subjects to close the loop of informed consent and validate results, Cameron's research with youth to produce a video encapsulating their views on racism, and Frazer's combined approach, where the subjects did both. By taking on the agenda of the researched – in each of these cases the subjects were all youth – the researchers illustrate how their accounts of empowerment play out in real contexts.

However useful Cameron et al.'s model, it is nevertheless problematic because it defines the sociolinguistic field primarily in terms of power relations, but these are not then sufficiently "disturbed" by the recommendations of the model. Although Cameron et al. recognize that each of the ethical models may not be attained when working with human subjects, there are a number of practicalities associated with advocacy and empowerment that are vitally missing. For one, the notion of empowerment on the theoretical level is often discussed as compensatory, or after the fact (Edwards, 2006). Though this is the likely reality of some research situations, Edwards argues that such a model actually reinforces difference especially within the majority/minority situations described in so much of sociolinguistic ethics research; in comparison, real power or empowerment is typically taken or negotiated rather than given.

Garner et al. (2006, p. 68) further problematize the complexity of relationships in the field or sociolinguistic site, arguing that an empowerment model reifies relationships only in terms of power, the very thing that it tries to undermine. Although Cameron et al. (1992) recognize the multiple identities of the linguist, not all of which are powerful, by "offering a voice" or "alternative understandings" that the subjects did not necessarily take up, the researcher is inevitably placed in the subject or "powerful" position with little done to disturb this notion. Instead, Garner et al. argue that a model grounded in a methodology of social relations such as Fiske's (1992) better captures the ethical concerns of sociolinguistics.

Fiske (1992, p. 690) posits that in all cultures, people attune to four relational models to generate all interaction, associated affect, and evaluation. Within interactions, social relations among people are progressively (re)defined along two axes: equality/inequality and independence/interdependence. For a sociolinguist, these axes generate the ever-changing complexities of ethical fieldwork and the research site.

A relationship characterized by *inequality* and *independence* is typical of a paid study in which the researchers hire individual subjects, such as in the case of experimental studies. Much quantitative research in sociolinguistics, even non-experimental research, is similarly characterized by *inequality* and *interdependence*. Recognizing the need for consultants' cooperation within a community and understanding the unequal situation, many researchers therefore find themselves ethically obligated to "give back."

In contrast is a research relationship characterized by *equality* and *independence*, which, as the literature on ethics in sociolinguistics reveals, is rare because it implies complete reciprocity and an agreement concerning such reciprocity at the outset. My own fieldwork situation with the last Mandan speaker and the affiliated tribe of the Mandan, Hidatsa, and Arikara (MHA) may be an example of such. The tribe hired me as part of a grant project for language preservation to do a specific task, paying for my travel and lodging, with the understanding that textual materials produced by the project would be copyrighted by the local school but that grammatical analysis would be under my purview and control. This relationship quickly changed to one of *equality* and *interdependence*: although the participants (co-researchers) in the Mandan language work have access to different types of knowledge, there is reliance and an expectation of mutual reciprocity. It would be difficult to analyze this situation merely in terms of ethics, advocacy, or empowerment.

In fact, analyses of ethical obligations are often best grappled with in light of the combined theoretical contexts, individual relationships, and sociopolitical relationships necessitated by the field in which a scholar is working. Sociolinguistics is necessarily social, and analysis of relationships of power is of vital consideration if we are to appropriately assume obligations to speakers and communities within our field of study. In an excellent example of the changing social relations in the field and a rare consideration of linguistic ethics that privileges both the voice of the researcher and of the consultant, McLaughlin and Sall (2001, p. 189) consider representations of themselves and each other based on their experience of working with each other in Senegal (Sall is the local participant, while McLaughlin is the fieldworker). They do so using the metaphors of grammatical relations, the primary focus of the fieldwork. Challenged by the task they set for themselves, they cite the following conversation, which summarizes the real difficulty in addressing such issues:

SALL: Somehow one has the impression that we are always the object and never the subject. We are the "material" that *toubabs* [white foreigners] come to study.

McLAUGHLIN: But this time, by presenting your own narrative, don't you think you have the opportunity to be the subject instead of the object?

SALL: (laughter) I'll talk about myself, but only at your initiative. So where does that put us?

Even when researchers are perceived by the community or individuals as having power by virtue of their potential identities as Westerner, white, middle-class, academic, male, standard speaker of a national language, etc., other facets of identity such as "struggling student," as in McLaughlin's case, may also strongly affect the degree to which the researcher can fulfill and sustain ethical obligations of advocacy and empowerment or social relations of interdependence. Considerations of ethics in sociolinguistics and endangered language studies have been largely silent on this point. Students are advised to create sustainable relationships with communities and to give back through their access to knowledge and institutions of higher education, but continuing beneficial relationships are dependent on a researcher's ability to sustain her or his own subsequent career through continued academic employment and/or grants. The practical exigencies associated with student status, tenuous job possibilities, and ethical obligations sometimes put the realistic student in limbo at best, and carrying an unfair burden at worst.

Empowerment itself is at the very least ethically problematic without careful consideration and exploration of all research relationships; not doing so could lead to naïve action or belief that individual reciprocal obligations have been addressed. For example, in the MHA language revitalization project I have already referred to, fulfilling obligations to the broader community was part of my initial agreement. Yet it was only after a year of work that I understood through my interactions with the only fluent speaker of Mandan that he had not been asked or had not thought his voice was an important factor in deciding what to highlight for the next generation. By changing my interactions so that we emphasized his conceptual expertise and cultural knowledge, we were able to create a productive project for the community.

Contrariwise, as a prominent sociolinguist stated, when evaluating initial drafts of the LSA's Ethics statement, "What if I don't want to advocate for or empower these people because they are corrupt, I don't like them, and they already have too much power?" It is not difficult to imagine scenarios in which we might work with groups for whom the kind of involvement implied by a model of ethics including advocacy and empowerment is at the very least distasteful, if not against our own personal ethics. As researchers interacting with communities, we are in a constant process of negotiating and choosing our obligations. This does not mean, obviously, that they can be ignored, but our ethical principles and obligations must be repeatedly and thoughtfully engaged even as they may change.

A social relations model such as that advocated by Garner et al. (2006) could have the effect of letting a researcher "off the hook" as she or he becomes absorbed with all of the emerging detailed relationships and ethical pitfalls, leading to little action because the goal is on the constant (re)negotiation of relationships. For the ethical researcher, there is no one clear direction with a social relations model. There is no normative ethics to which we appeal for guidance. There is no clear understanding that "empowering" research is at the pinnacle of ethics, but such an approach does reflect the deeper reality of research that is by its very nature grounded in the social.

College/University Ethics Boards

No matter how we negotiate our ethical position, the human subjects research process typically positions the researcher in control within an independent, unequal relationship in which subjects give up their rights to remain unimpeded by participating in the research. Ethics of a specific discipline are backgrounded to the larger federal good and a biomedical research paradigm. For many, ethics boards, including Institutional Review Boards (IRBs) in the United States, give rise to the largest number of practical concerns with regard to sociolinguistic research, especially that which takes into account interdependence between the researcher and the subjects and a developing field relationship. *American Ethnologist* issue 33(4) (2006) discusses and debates the long-standing problem for ethnographic research and the (mis)understanding and (mis)placement of social science research under the aegis of (mis)informed ethics boards.

Additionally, there is great deal of institutional variance concerning the enforcement of human subjects code. For instance, a number of institutions have determined that ethnographically oriented sociolinguistic or descriptive fieldwork focusing on minority languages or dialects does not fall under the definition of "research" and therefore is exempt from review. Research is defined by the federal regulatory code 45 CFR 46 as "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge" (DHHS 2005, known since 1991 as the "Common Rule"). How "generalizable" is sociolinguistic variation within one small community? Some institutions, on the opposite end of the spectrum, hold to a different interpretation and considerable mission creep, examining linguistic research from the view of psychological or medical research.

No matter what the institutional bent, it is vital to assure protection of human subjects with fair treatment, honesty concerning remuneration and possible effects of the research, anonymity, informed consent, and no anticipated harm. However, each of these considerations gives rise to a number of questions on its own, even without the notion of researcher–researched interdependence in the mix. Informed consent is a complex issue, especially if subjects are illiterate, are unwilling to give recorded consent because of a history of government discrimination, and/or simply feel that the concept of "informed" is ambiguous. If these concerns can be anticipated and addressed, subjects may be informed of all of the possible risks and benefits to them, but it is rare that the researcher informs participants of the full value of the research to the scholar in assuring her or his own academic position, obtaining grants, providing indirect costs to support the academic institution, and other details.

As many linguists move toward a model that assures that subjects benefit socially from linguistic research, anonymity is similarly problematic and ultimately processual. Where subjects speak a minority language with a strong folkloric tradition, or in situations in which they wish to be recognized as an expert contributor and leave a legacy, anonymity is anti-ethical. Childs, Van Herk, and Thornburn (2011) delineate the planning and procedural difficulties within a small community in Newfoundland, Canada. They grapple with how best to balance anonymity for research participants, building a corpus for other researchers (eliding identifying information that may be discomforting), and allowing the community as a whole to be recognized for its distinctiveness. Granting access both to community members and to academics in varying degrees as requested and negotiated by the participants and the researchers was vital to the success of their project. Such detail is difficult to put into a human subjects form at the outset, and it is the type of flexibility necessary for a dialogically constructed ethical approach with one's consultants.

In effect, for an ethical sociolinguist the ethics board may be both too lenient (because it does not consider community responsibilities) and too strict (because of its focus on complete anonymity and destruction of identifying records). The solution as recommended by many right-minded linguists is that we get involved with our local ethics boards. It is our ethical obligation to educate the academy on the norms and ethical issues at the forefront of our field. Though this recommendation is laudable, on July 26, 2011 the Department of Human and Health Services proposed the implementation of stricter guidelines for human subjects protection in order to standardize the response of IRBs across institutions. The proposal includes the standardization of informed consent forms to a greater degree, increased security provisions and standardization of rules concerning anonymity and confidentiality of data, and the extension of the federal rule to all institutions that receive any federal research funding. Consideration of these proposals is still ongoing, and the social, interdependent and developmental nature of sociolinguistic ethics is likely to continue to be fraught within our home institutions. As one reviewer of this chapter commented, it is ironic that while the United States is potentially standardizing ethical review of research, the Tri-Council policy in Canada has just moved away from strict human subjects protection for research in the social sciences and humanities. Either approach has its challenges. Ethics boards often ask the difficult human subjects questions that have been overlooked because of past practice or inadequate research preparation. They play an important role in helping students and faculty make their research and the ethics of their field understood by the public as a whole. A standardized model will likely miss the subtlety that many ethics boards have, but are often accused of lacking.

Conclusion

I have argued that ethics in sociolinguistics is inextricably linked to an interactive, dialogic speech community and to socio-methodological considerations that arise from the inception of a project. The nature of the social relationship with a speech community and with individual speakers changes over time and continually calls our methodological assumptions into question, forcing us to reflect more thoroughly during the creation of sustainable projects.

Although I have concentrated on subjects of research, the ethical gaze cannot just look toward the community of speakers and their representation in light of a larger social vision. For ethics, thorough reflexivity goes beyond what we bring to the representation of the other. Our social interactions also necessarily bring the weight of our institutions that benefit from the indirect costs produced by our grant-supported research, and we as scholars see our own gains from engaging in such research through jobs, grants, or career advancement. Because human subjects guidelines fix our attention on the protection and rights of those studied, we are not required to disclose what we hope to gain for ourselves or have gained for our institution. Our lack of frankness regarding our own benefits has long been regarded suspiciously by minority communities even if we do "give back," and giving support to a few minority students in a grant is no longer considered an adequate guarantee of broader impact. In fact, in reviewing recent grants for the National Science Foundation (NSF), I have noticed the glaring omission of indirect costs in some, which would have added at least $100,000 to the modest amount of requested money. By eliminating the academic institutional middleman, and by hiring linguists as consultants to help with archiving and research, communities themselves now have the ability in some situations to empower themselves. Our positions as scholars who benefit greatly from research we conduct on, and sometimes for or with, communities therefore may not continue to be so privileged.

Our own ethical obligations and responsibilities may also differ depending on our position and status with the academic community. For a senior, endowed professor at a major research institution to ask students to take on long-term responsibilities for a community – obligations that require academic employment or the ongoing possibility of doing research and interacting with a community – does not recognize the reality of our past and current economic climate. Likewise, it is important for senior academics to acknowledge their privileged position as one of relative power when assuming a commensurate amount of responsibilities or obligations. Because sociolinguistic ethics are most appropriately grounded in the social, the roles of all speech participants, as well as their institutions, culture, and history, play a vital role in determining the best methods for thoughtful, sustained interaction. Ignoring this first ethical responsibility puts us at risk of replicating a hegemonic research paradigm, which is simultaneously as anti-science as it is insensitive.

References

Ball, M. J. (Ed.). (2005). *Clinical sociolinguistics*. Malden, MA: Blackwell.

Bowern, C. (2008). *Linguistic fieldwork: A practical guide*. Basingstoke, UK: Palgrave Macmillan.

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., & Richardson, K. (1992). *Researching language: Issues of power and method*. London: Routledge.

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., & Richardson, K. (1997). Ethics, advocacy and empowerment in researching language. In N. Coupland & A. Jaworski (Eds.), *Sociolinguistics: A reader* (pp. 145–162). New York: St. Martin's Press.

Champagne, D., & Goldberg, C. (2005). Changing the subject: Individual versus collective interests in Indian country research. *Wicazo sa Review, 20*, 49–69.

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.

Coupland, N., & Jaworski, A. (1997). *Sociolinguistics: A reader*. New York: St. Martin's Press.

Davies, A. (1999). Ethics in educational linguistics. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 21–25). Oxford: Elsevier Science.

Deloria, V., Jr. (1969). *Custer died for your sins: An Indian manifesto*. New York: Macmillan.

Department of Health and Human Services (DHHS). (2005). Code of Federal Regulations, Title 45 "Public Welfare," Part 46 "Protection of Human Subjects." Retrieved from http://www.hhs.gov/ohrp/policy/ohrpregulations.pdf

Edwards, J. (2006). Players and power in minority-group settings. *Journal of Multilingual and Multicultural Development, 27*(1), 4–21.

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review, 99*(4), 689–723.

Garner, M., Raschka, C., & Sercombe, P. (2006). Sociolinguistic minorities, research, and social relationships. *Journal of Multilingual and Multicultural Development, 27*(1), 61–78.

Heller, M. (1999). *Linguistic minorities and modernity: A sociolinguistic ethnography*. New York: Longman.

Kroskrity, P. V., & Field, M. C. (2009). Revealing Native American language ideologies. In P. V. Kroskrity & M. C. Field (Eds.), *Native American language ideologies* (pp. 3–28). Tucson: University of Arizona.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1982). Objectivity and commitment in linguistic science: The case of the Black English trial in Ann Arbor. *Language in Society, 11*, 165–201.

Linguistic Society of America. (2009). Linguistic Society of America Ethics Statement. Retrieved from http://lsadc.org/info/pdf_files/Ethics_Statement.pdf

McLaughlin, F., & Sall, T. S. (2001). The give and take of fieldwork: Noun classes and other concerns in Fatick, Senegal. In P. Newman & M. Ratliff (Eds.), *Linguistic fieldwork* (pp. 189–210). Cambridge: Cambridge University Press.

Mesthrie, R., Swann, J., Deumert, A., & Leap, W. L. (2000). *Introducing sociolinguistics*. Amsterdam: John Benjamins.

Morgan, M. (1994). The African-American speech community: Reality and sociolinguists. In M. Morgan (Ed.), *The social construction of identity in creole situations* (pp. 121–148). Los Angeles: Center for Afro-American Studies, UCLA.

Rickford, J. R. (1997). Unequal partnership: Sociolinguistics and the African American speech community. *Language in Society, 26*, 161–198.

Smitherman, G. (1988). Discriminatory discourse on Afro-American speech. In G. Smitherman-Donaldson & T. A. van Dijk (Eds.), *Discourse and discrimination* (pp. 144–147). Detroit: Wayne State University Press.

Trechter, S. (1999). Contextualizing the exotic few: Gender oppositions in Lakhota. In M. Bucholtz, A. C. Liang, & L. A. Sutton (Eds.), *Reinventing identities: The gendered self in discourse* (pp. 101–119). New York: Oxford University Press.

Wolfram, W. (1998). Scrutinizing linguistic gratuity: A view from the field. *Journal of Sociolinguistics, 2*, 271–279.

Wolfram, W., Reaser, J., & Vaughn, C. (2008). Operationalizing linguistic gratuity: From principle to practice. *Language and Linguistics Compass, 2*(10), 1–26.

Vignette 3a
Responsibility to Research Participants in Representation

Niko Besnier

Since the late 1980s, like other researchers who focus on people's behavior and social practices, sociolinguists have been increasingly required to abide by principles of ethical responsibility and to demonstrate their commitment to these principles in a formal way. The precept that underlies codes of ethics can be summarized by the seemingly simple injunction "Do no harm." Underlying this simplicity, however, lies a host of entanglements, many of which derive from the fact that researchers' intention not to harm people does not necessarily ensure that harm is not done. In addition, people who are under research scrutiny are increasingly asking researchers, "In what way does research benefit us?" While the concern that harm be avoided applies today to all research, the concern over creating value is increasingly important, particularly when research focuses on people whose rights and welfare have historically been undermined.

The recognition that research must be ethically grounded arose, in the decades following World War II, outside the social sciences and the humanities, principally in biomedical research. But important questions of continued relevance also emanated from an entirely different source: the postcolonial critique of Western-based knowledge making of the 1980s and 1990s. Inspired by Foucault's (1980) argument that knowledge is always infused with power, Said's (1979) epoch-making critique of the Orientalist intellectual tradition of the 19th and 20th centuries demonstrated that intellectual endeavors such as philology, epigraphy, and other humanistic pursuits were complicit with imperialist domination in its multiple forms (military, economic, moral, etc.), even when such complicity was not intended. Following this critique, researchers could no longer assume that the production of knowledge was a politically neutral endeavor and that scientific imperatives overrode the concerns of those researched.

In comparison to its kindred disciplines, sociolinguistics was slower in its uptake of these principles, although the activism of such scholars as William Labov and Shirley Brice Heath in mobilizing against the marginalization of speakers of non-standard dialects, principally in educational institutions, stands out as particularly important early engagements with ethical responsibility (Heath, 1983; Labov, 1972a). Today, sociolinguists must address the implicit responsibility that our research be made available, in one fashion or another, to the people who produce the data. While early sociolinguists assumed a clear boundary between researcher and research subject, contemporary researchers can no longer do so, as "subjects" today, even among geographically remote groups, seek to have their voices heard in academic conversations about them. The influence of contemporary cultural anthropology can be felt here, particularly in the recognition that "data" are the intersubjective product of an encounter between the researcher and the people whose lives are under scrutiny, who themselves may occupy differing positions. An early version of this insight was already present in Labov's (1972b) methodological discussion of the "observer's paradox," but it has now dispelled myths of objectivity and authenticity in research on language in its social context (Bucholtz, 2003).

Beneath these seemingly straightforward recognitions, however, lies an array of questions that defy simplistic treatment. One central issue is the extent to which research subjects are materially disadvantaged and the role that language practices play in this position of disadvantage. The canonical example is that of sociolinguistic research in situations of language endangerment. Speakers of endangered languages are overwhelmingly small-scale indigenous populations (e.g., Native Americans, Australian Aborigines, indigenous peoples of Siberia) that are often multiply disadvantaged: socially, economically, politically, and culturally. In many cases, the disadvantage results from a history of marginalization and oppression in the hands of a majority group, in which the very linguistic denigration that produced the endangerment, such as children being forbidden to speak their native language in schools, operates as one of many forms of subjugation. Some have argued that researchers in such situations are under no obligation to produce work that benefits the groups in question (other than paying informants for their time), although this perspective is today generally viewed as deeply problematic. Most researchers agree that in such situations the design of any (socio)linguistic research must explicitly address how the results of the work will encourage linguistic revival. At the same time, researchers must consider the power dimensions of such efforts, which may end up being viewed by speech communities as paternalistic actions that simply reproduce the hegemonic past (Hale et al., 1992; Walsh, 2005; Whiteley et al., 2003).

As sociolinguists have amply documented, language may be involved in the marginalization of groups in numerous ways. Such is the case of socioculturally devalued non-standard dialects. One aspect of researchers' responsibility in this case that has received some sustained attention is the potential consequence of the choices researchers make when transcribing speech, which can tacitly reproduce power asymmetries and social inequalities. The seemingly innocuous arrangement of speakers' turns on a page, for example, visually privileges certain speakers (Ochs, 1979), and the use of non-standard orthographic conventions, while constituting a powerful expressive tool, may reinforce the social devaluation of non-standard ways of speaking (Bucholtz, 2000; Jaffe, 2000; Preston, 1985).

More generally, problems arise when researchers become interested in social practices, including linguistic practices, that some members of the society may find problematic. Such was the case with my own work on gossip on a small atoll of the Central Pacific (Besnier, 2009). While islanders find gossip morally reprehensible, they also take enormous pleasure in engaging in it, and it is through gossip that much consequential political action takes place. When it gains enough momentum, however, gossip can be deeply damaging to the lives of people that it targets, and this damage can be further aggravated by the suppression of gossip in public contexts because of its moral taint.

This example raises several points. One is the fact that, even in small-scale, close-knit, and seemingly clearly bounded social groupings, "community" is a problematic category, as a number of anthropologists and cultural theorists have argued (e.g., Creed, 2006; Joseph, 2002). Even when they have everything to gain from presenting a face of communal harmony to the outside world, including the researcher, social groups can be fraught with deep interpersonal conflicts and divergent opinions over such matters as culture, morality, and the future. The second point is the consequence of the first: in such situations, who should decide how the group's social and linguistic practices should be represented remains an open question. If researchers opt to conform to the image preferred by those in power, they risk eliding alternative representations and potentially aggravating structures of inequality. If researchers align their representation with those of the oppressed, they may ameliorate the latter's position – or on the contrary provide additional tools for their oppression. Moreover, researchers cannot simplistically assume that "returning their results" to those in power will satisfy the ethical requirement of ensuring that their research is useful or that it will ensure that research is used for the improvement of the human condition.

While no simple ethical guideline applies to all situations, one can suggest a rule of thumb: prior to beginning a study, researchers must think through potential power dynamics and ethical conflicts that may arise in the field, and all decisions about the nature, form, and circulation of research must be made with recognition of the complexities of the situation and the instability of power relations. While researchers have the power to animate certain voices but not others, and to privilege certain representations over others, social groups and individuals also have the power to accept, resist, and reject representations of their social practices, particularly in a world in which the boundary between researcher, object of research, and audience is no longer straightforward (Allen, 1997; Brettell, 1993). Ironically, while they may benefit from greater insight into the workings of the community in which the research takes place, "insider" researchers are in a potentially even more difficult position and can become the object of severe criticism for bias by fellow community members. In short, all sociolinguists must consider carefully questions of ethics broadly defined, taking into account not only issues of consent, but also power, scale, representation, subjectivity, and positionality.

References

Allen, C. (1997). Spies like us: When sociologists deceive their subjects. *Lingua Franca, 7*(9), 31–39.

Besnier, N. (2009). *Gossip and the everyday production of politics*. Honolulu: University of Hawai'i Press.

Brettell, C. B. (Ed.). (1993). *When they read what we write: The politics of ethnography*. Westport, CT: Bergin & Garvey.

Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics, 32*(9), 1439–1465.

Bucholtz, M. (2003). Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics, 7*(3), 398–416.

Creed, G. W. (Ed.). (2006). *The seductions of community: Emancipations, oppressions, quandaries*. Santa Fe, NM: School of American Research Press.

Foucault, M. (1980). *Power/knowledge: Selected interviews and other writings, 1972–1977*. C. Gordon (Ed.). New York: Vintage.

Hale, K., Krauss, M., Watahomigie, L. J., Yamamoto, A. Y., Craig, C., Jeanne, L. M., & England, N. C. (1992). Endangered languages. *Language, 68*(1), 1–42.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. Cambridge: Cambridge University Press.

Jaffe, A. (2000). Nonstandard orthography and nonstandard speech. *Journal of Sociolinguistics, 4*(4), 497–513.

Joseph, M. (2002). *Against the romance of community*. Minneapolis: University of Minnesota Press.

Labov, W. (1972a). *Language in the inner city: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1972b). Some principles of linguistic methodology. *Language and Society, 1*(1), 97–120.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York: Academic Press.

Preston, D. R. (1985). The Li'l Abner syndrome: Written representations of speech. *American Speech, 60*(4), 328–336.

Said, E. (1978). *Orientalism*. New York: Vintage.

Walsh, M. (2005). Will indigenous languages survive? *Annual Review of Anthropology, 34*, 293–315.

Whiteley, P., Errington, J., England, N. C., Friedman, J., Bulag, U. E., Haviland, J. B., & Maurer, B. (2003). Language ideologies, rights, and choices: Dilemmas and paradoxes of loss, retention, and revitalization. *American Anthropologist, 105*(4), 712–781.

Vignette 3b
Conducting Research with Vulnerable Populations

Stephen L. Mann

For the past several years, I have been conducting research in communities that are often defined as "vulnerable populations" by ethics boards, including Institutional Review Boards (IRBs) in the United States. In 2004, my fieldwork took me to a public drag talent show held weekly in a gay bar (Mann, 2011a). From 2009 to 2010, as part of my dissertation research (Mann, 2011b), I interviewed eight self-identified gay men in the US South and conducted focus groups with an additional eight self-identified gay and queer men. Some readers might assume that the decision to provide pseudonyms for all of my research participants and field sites across both projects was made for me as a result of rules set out by the IRB at my affiliated university at the time. IRB requirements did, of course, play a major and often primary role in the decision-making process; there were, however, several questions that I still had to address.

To What Extent Are Data Part of the Public Sphere?

My drag queen research (Mann, 2011a) was conducted in a public venue during a public performance. Both the bar and the performer are well known by local members of the community. I am fairly confident that, should community members read my article, they would easily be able to determine the actual identity of both the field site and the drag queen hostess whose language became the focus of my analysis. (She uses the gender-marked 'hostess' rather than 'host' to refer to her role in the performance.) Additionally, while I tried to be as unobtrusive as possible, audience members present at the time of the recording would have seen me with my audio recorder collecting data. Finally, sociolinguistic analysis of public figures is commonplace, and protecting the identity of public figures is rarely, if ever, deemed to be necessary (cf. popular hip hop artists, as studied in Blake & Shousterman, 2010; Margaret Cho, as studied in Chun, 2004; and Anita Hill and Clarence Thomas, as studied in Mendoza-Denton, 1995).

Why, then, did I ultimately decide to identify the city in which I conducted my fieldwork as simply "a mid-sized city in the southeastern United States" (Mann, 2011a, p. 795) and to provide pseudonyms for both the bar ("Jay's") and the hostess ("Suzanne")? In the end, I determined that providing a real name and location would not increase the strength of my argument. Nor would their inclusion help readers better understand my analysis, which might have been true if the hostess were a nationally recognized public figure, as in Barrett's analyses of RuPaul's language (e.g., Barrett, 1998). As a result, I opted to provide a level of protection to veil (albeit thinly) the identity of both the field site and the participant.

What Level of Anonymity Do Participants Want?

My university's IRB required me to use pseudonyms to protect the identity of my interview and focus group participants in my dissertation study (Mann, 2011b). At the beginning of each interview and focus group session, I gave each man the option of providing me with a pseudonym of his own choosing or letting me choose a pseudonym for him. All of the focus group participants and nearly half of the interview participants told me that I did not need to use a pseudonym; instead, they preferred that I use their real names, because they were so personally invested in the stories they were telling. At that point, I was faced with an ethical dilemma. If I used participants' real names, then I would be invalidating the approval already granted to me by my institution's IRB. If I did not use participants' real names, then I was not fully meeting the requests of the community in which I was conducting my research.

One option for me at the time was to submit an amendment to my IRB application and ask permission to use real names for participants who specifically requested that I do so. My experience requesting IRB approval to conduct sociolinguistic research with gay men left little doubt that such a request would have been denied. An earlier project for which I requested approval, for example, had been approved under the condition that I destroy all recordings after completing my data analysis – a condition I later successfully fought to have removed. I decided for the dissertation project, however, not to pursue permission to use participants' real names. Rather, I used pseudonyms as originally planned because of the question that I consider in the next section.

Might Any Unforeseen Risks to Participants Arise after the Informed Consent Process Has Been Completed?

The initial informed consent document that interview participants signed for my project stated that "there is a minor level of risk associated with participating in this study. Discussing the coming out process can trigger bad memories for some people" (Mann, 2011b, p. 186). About halfway through each interview, participants were informed that the primary focus of my research was language. I then asked them to provide consent to use a short sample of their speech, drawn from the recorded interview, in an online attitude and perception study. At that point, participants were warned, "There is a minor level of risk associated with including a sample of your speech in a survey. It is possible that a person … could recognize your voice from only a 20–30 second sound clip" (p. 191).

During the focus group sessions, participants listened to each speech sample and discussed their attitudes toward each clip. Participants were brutally honest in their discussions. One participant, for example, described a speaker as "an overly confident dumb ass who really shouldn't be confident" (p. 70). As a researcher, I appreciated the extent to which the focus group participants presented their attitudes openly and honestly. I realized, however, the potential for unforeseen embarrassment for speakers in the event that a connection could be made between the speaker and the negative commentary made about him in the focus groups – or simply from a speaker gaining insight into listeners' potential negative attitudes toward his own linguistic practices.

There is, unfortunately, no complete solution to this dilemma, but I did incorporate protections to minimize potential harm to study participants. First, I made a list of counselors available to participants in the event that their participation, including dissemination of study findings, were to cause them any psychological harm. Second, as discussed earlier, I protected participants with pseudonyms despite some individuals' requests to use their real names.

At What Point Do Even Pseudonyms Provide Insufficient Protection for Participants?

Up until now, I have been talking about whether or not to use pseudonyms to protect field sites and fieldwork participants. In one portion of my study I provide long interview excerpts in which participants discuss their relationships with their parents and how those relationships have changed over the years (Mann, 2011b, pp. 105–109). The information that participants provided was very personal and had the potential for worsening what, for many of my participants, were already strained family relationships. The information in the excerpts was important for the argument I was making at the time, but the specific link between the speaker and the excerpt was not necessary. I decided, therefore, to provide participants with another level of protection by omitting speaker names entirely rather than introduce more risk unnecessarily.

Working with "vulnerable populations" specifically and human subjects more generally poses many challenges for sociolinguists. As I have discussed here, researchers must meet the requirements for confidentiality that are set out by our institutions' IRBs, and in some cases may also need to go beyond them to provide additional levels of protection. Because language is intimately tied to speakers' social identities, we as sociolinguists need to consider fully such factors as the extent to which linguistic data can be considered "public," the relevance of field site and participant name disclosure to the argument being made, the needs and expectations of the participants themselves, and, most importantly, the extent to which disclosure introduces additional and/or unnecessary risk, not only for individual participants but also for their families and communities.

References

Barrett, R. (1998). Markedness and styleswitching in performances by African American drag queens. In C. Myers-Scotton (Ed.), *Codes and consequences: Choosing linguistic varieties* (pp. 139–161). New York: Oxford University Press.

Blake, R., & Shousterman, C. (2010). Diachrony and AAE: St. Louis, hip-hop, and sound change outside of the mainstream. *Journal of English Linguistics, 38*, 230–247.

Chun, E. (2004). Ideologies of legitimate mockery: Margaret Cho's revoicings of mock Asian. *Pragmatics, 14*, 263–289.

Mann, S. L. (2011a). Drag queens' use of language and the performance of blurred gendered and racial identities. *Journal of Homosexuality, 58*, 793–811.

Mann, S. L. (2011b). Gay American English: Language attitudes, language perceptions, and gay men's discourses of connectedness to family, LGBTQ networks, and the American South. (Unpublished doctoral dissertation). University of South Carolina, Columbia, SC.

Mendoza-Denton, N. (1995). Pregnant pauses: Silence and authority in the Anita Hill–Clarence Thomas hearings. In K. Hall & M. Bucholtz (Eds.), *Gender articulated: Language and the socially constructed self* (pp. 51–66). New York: Routledge.

Vignette 3c
Ethical Dilemmas in the Use of Public Documents

Susan Ehrlich

Over the past decade or so, I have investigated the discourse of trials and judicial decisions in sexual assault cases in both Canada and the United States. My work has shown that, despite progressive reform of rape and sexual assault statutes in both countries since the 1980s, sexist rape mythologies continue to inform judicial decisions and circulate within the trial discourse of these cases in ways that disadvantage complainants and protect defendants. In other words, I have attempted to show how language use in these contexts is crucial to understanding the way that gendered inequalities are created and reproduced in the legal system. So, while I believe that my work has positive implications for women in general, the ethical dilemma that I face involves the way that my research represents and positions the individual women in my data – that is, women who have been complainants in rape trials.

Because I rely on public documents (e.g., trial transcripts and judicial decisions) and audiotapes (e.g., of trials) for my data, securing the informed consent of speakers is not necessary. As Johnstone (2000) points out, "anyone can quote or refer to" forms of public discourse such as "transcripts of speeches by public officials, tapes or transcripts of public meetings of governmental organizations, and published materials of most kinds (though permission to quote copyrighted material is often required)" (p. 56). Thus, in the sense that the language of trials is produced in a public forum – that is, in a courtroom, which is generally open to the public – where speakers know they are being recorded or transcribed for subsequent public "hearings," trial talk can be said to be "designed for public hearing" (Cameron, 2001, p. 25). What this means, however, is that I am not obligated to, and generally do not seek, the "active cooperation" of those that I study, a research situation that Cameron, Frazer, Harvey, Rampton, and Richardson (1992, p. 23) suggest can lead to the objectification of research subjects.

Even more problematic than the potential objectification of my subjects is the fact that they are, arguably, represented in my data as sexualized objects. While the invasion of individuals' privacy is not supposed to be an issue when speech is produced in the public domain (as opposed to the private domain), complainants in rape trials are typically under pressure to expose private details of their lives in ways that seem to break down the distinction between the public and the private. McElhinny (1997, pp. 107–108) argues that, while the public–private dichotomy may be useful for understanding middle-class lives, especially those of middle-class men, there are other segments of the population for which the distinction is not relevant. Indeed, the kind of interaction exemplified below, from the cross-examination of a complainant, JL, in an American rape trial, shows that public speech events do not preclude the possibility of individuals having to expose intimate and graphic details of their lives.

CE: (1.0) By the way, aside from your private areas, did you ever have any bruises or cuts or scrapes on any part of your body?
JL: Uh – no.
CE: (5.0) Michael Wilson was trying to put his penis in you?
JL: Yes.
CE: Were your underpants up or down?
JL: Uhm, they were down.
CE: Do you know how your underpants got down?
JL: M—Mike pulled them down.
CE: Were they ripped in any way?
JL: N—no.
CE: Then what happened?
JL: Uhm, he tried to put it in, but he ended up putting it in the wrong place and I said, "Ow."
CE: To Michael?
JL: Yes. (1.0) And that's when he started laughing. And then he told Maouloud to get out of the car.
CE: (2.0) When he told Maouloud to get out of the car had – to your knowledge, was his penis out at all?
JL: Mhm, no.
CE: (2.0) So, in your recollection by the time Maouloud got out of the car, he had touched you where you said, but he didn't try and put his penis in you.
JL: No.

Historically, in both Canada and the United States, sexual violence cases were subject to "special evidence rules": strict and unique rules of evidence that focused far more attention on the complainant's behavior than was possible in other kinds of criminal cases (Busby, 1999; Schulhofer, 1998). While such rules are no longer encoded in statutes in either country, complainants in rape trials seem to be subject to a higher standard of proof than is typically true for other kinds of criminal cases – for example, those involving theft. This situation can be seen in the above excerpt, where intrusive and personal questioning from the cross-examining lawyer subjects the complainant to a kind of scrutiny that would probably not extend to an alleged victim of theft. Thus, although the kind of interaction that we see in the above excerpt takes place in the public domain, the seeming collapse of the "public" vs. "private" distinction in this context raises questions about whether we can consider this interaction as "designed for public hearing." In fact, some feminist critics have argued that the rape trial forces rape victims to participate in a "pornographic vignette" (Smart, 1989, p. 39) to the extent that it "*gives pleasure* in the way that pornography gives pleasure" (MacKinnon, 1987, cited in Smart, 1989: 39, emphasis in original).

My research is significant to understanding how gendered inequalities are produced in the legal system, but it may be the case that for some audiences my data help reproduce the very gendered inequalities that I am trying to expose. The word "dilemma" is defined in at least one dictionary as "a problem that seems incapable of a solution." While I have no solution for the particular ethical dilemma that I face in my research (i.e., how my data position and represent complainants in rape trials), I think it raises more general issues about ethics in sociolinguistic research. First, in the same way that I have shown how complainants in rape trials may be required to reveal private aspects of their lives in public settings, McElhinny (1997) argues that the poor, who typically depend on state aid, "are forced to open themselves up to state scrutiny" in ways that also make the public–private distinction irrelevant (p. 108). Thus, because the public–private distinction does not seem to be meaningful for certain (disadvantaged) groups of people, perhaps it should not be assumed a priori that speech produced in the public domain is necessarily "designed for public hearing" and, by extension, does not require the informed consent of research subjects. Second, as Cameron et al. (1992) argue, researchers need to be attentive to the *representation* of their research subjects, as representation "inevitably involves the recontextualization of utterances" (p. 132), and, as Bauman and Briggs (1990) point out, recontextualization inevitably involves transformations in meaning. That is, according to Bauman and Briggs, once a stretch of talk is "lifted out of its interactional setting" and turned into a "text," it may bring something from its earlier context, but may also take on different meanings as it is "recentered" in a new context (pp. 73–75).

As researchers, we need to think about the effects of representing our research subjects in an academic context – a context that may serve to caricature or stereotype them, or, as in my research, may produce them as sexualized objects. Rampton's solution to this kind of problem, as reported in Cameron et al. (1992), was to seek informed consent for his "representations." In general, perhaps the primary implication of my ethical dilemma is that the securing of informed consent in sociolinguistics needs to be extended to research situations and research dimensions that are not typically understood as requiring it.

References

Bauman, R., & Briggs, C. L. (1990). Poetics and performance as critical perspectives on language and social life. *Annual Review of Anthropology, 19*, 59–88.

Busby, K. (1999). "Not a victim until a conviction is entered": Sexual violence prosecutions and legal "truth." In E. Comack (Ed.), *Locating law: Race/class/gender connections* (pp. 260–288). Halifax, Nova Scotia: Fernwood Publishing.

Cameron, D. (2001). *Working with spoken discourse*. Thousand Oaks, CA: Sage.

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., & Richardson, K. (1992). *Researching language: Issues of power and method*. London: Routledge.

Johnstone, B. (2000). *Qualitative methods in sociolinguistics*. New York: Oxford University Press.

McElhinny, B. (1997). Ideologies of public and private language in sociolinguistics. In R. Wodak (Ed.), *Gender and discourse* (pp. 106–139). Thousand Oaks, CA: Sage.

Schulhofer, S. J. (1998). *Unwanted sex: The culture of intimidation and the failure of law*. Cambridge, MA: Harvard University Press.

Smart, C. (1989). *Feminism and the power of law*. New York: Routledge.

Vignette 3d
Real Ethical Issues in Virtual World Research

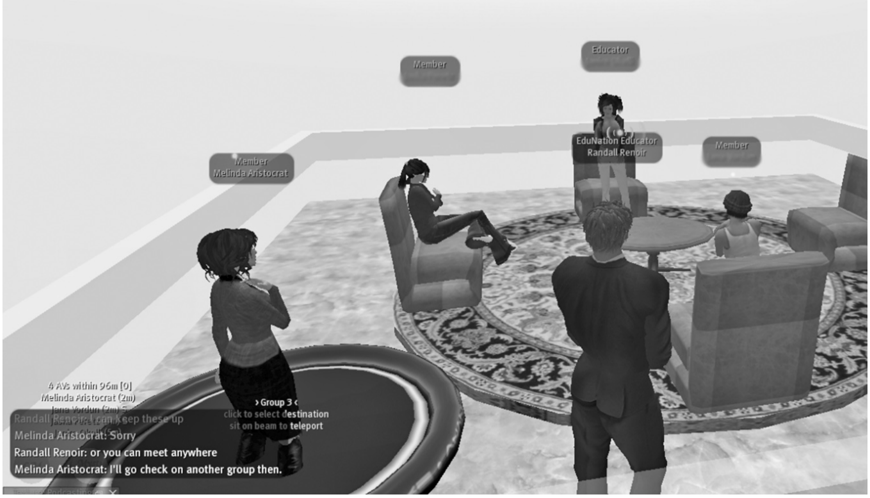
Randall Sadler

Since I began teaching and researching in virtual worlds (VWs) in 2006 using Second Life (SL), there has been a rapid increase in the use of VWs among both language educators and scholars. VWs are online 3D environments inhabited by avatars controlled by their real-world (RW) users. While a detailed examination of VWs goes beyond the scope of this vignette, *Virtual Worlds for Language Learning: From Theory to Practice* (Sadler, 2011) is a good source for more details. There are many reasons for educators to use VWs. They may provide a "rich range of collaborative social activities around objects" (Brown & Bell, 2004, p. 350) and allow educators "to develop learning activities which closely replicate real-world learning experience" (Childress & Braswell, 2006, p. 189).

From a researcher's perspective, perhaps the most attractive aspect of VWs is that they are the sites of enormous amounts of language production, both in users' native language(s) and in others they wish to practice (Sadler, 2011). The ease of data collection in these worlds makes them even more alluring – though ethically challenging. While a full discussion of online research ethics is beyond the scope of this vignette, one excellent overview is provided in the June 1996 issue of *Information Society*. In particular, King (1996) heavily influenced my own designs for ethical research discussed below, and the discussion by Thomas (1996) on the disastrous "Rimm 'cyberporn' study" should be required reading for anyone doing research online. Although the technology we use today has evolved greatly since 1996, many of the ethical concerns remain the same.

To illustrate these issues, I'll center this discussion on Figure 3d.1, which shows a meeting between two groups of students, one based at a US university and the other group located in Spain. These students (seated in the image) were collaborating in the VW in order to create jointly produced podcasts to use with their RW student colleagues in Spain and the United States.

Although the researchers are present in this image (positioned in the foreground are myself and Melinda Dooly – Randall Renoir and Melinda Aristocrat, respectively, in SL), there were actually six groups talking simultaneously in locations separated by hundreds of virtual meters, so we were only with each group briefly. We collected data in three ways. First, we asked the participants to activate an SL feature that saves a copy of all public text chat onto their computers. We then asked the students to email us copies of these texts. Since the majority of the communication took place via oral chat, we also asked at least one student



Randall Renoir: I can keep these up
Melinda Aristocrat: Sorry
Randall Renoir: or you can meet anywhere
Melinda Aristocrat: I'll go check on another group then.

Figure 3d.1 A Group Meeting in Second Life.

per group to screen record the interaction using tools like Fraps or Camtasia. We thus obtained videos of all the interaction – text chat, oral chat, and non-verbal communication. Finally, we acted as participant observers whenever present. Participants were aware from the beginning that we were engaging in action research involving a range of data collection techniques. In addition, a human subjects review process had been undertaken by both researchers at our respective universities, and students gave informed consent. When the data were utilized for publications, all avatars were given pseudonyms and, in the case of images from SL, avatar names were blurred (by default, each avatar's name appears above their figure on-screen).

While following these procedures is the proper way to conduct research in VWs, there is great temptation to collect more ethically ambiguous data, owing to the ease of recording in these environments. For instance, although this group platform that we were studying was 600 meters above one of the EduNation Islands in SL, I could have simply disabled the constraints on my SL camera while down on the island and then cammed up to this group while remaining on the ground. This would have allowed me to spy on the group and use a screen recorder to document all interactions (this unethical version assumes no consent has been given). Recording text chat could have been even easier since I could have placed an invisible chat recorder out on each meeting table to record all text chat and then retrieved the data the next time I came online. I could also have set out avatar trackers to tell me which avatars visited the meeting areas and how long they stayed. The challenge, as discussed by Hair and Clark (2007), is that "ethical decision making is heavily technology dependent and often subject to a 'technology lag' where ethics is often seen to play catch-up to the multitude of methodological options available to the researcher" (p. 784).

While this case is relatively simple, it is useful to consider the wide range of environments in VWs where either research is already being performed or that potential exists for sociolinguists, such as in SL schools (both formal and informal), dance clubs (ranging from ballroom to hip hop), beaches (both nude and clothed), horse-riding sims, role play regions, shopping malls, re-creations of famous RW locations (e.g., London), private homes, space stations, sex clubs, etc. While getting informed consent in some VW environments is both required and easy (e.g., in classroom research), how do researchers determine what to do elsewhere?

Happily, there are a growing number of works that discuss the ethical challenges of VW research. Joshua Fairfield, a professor of law, has an excellent overview of the field in *Avatar Experimentation: Human Subjects Research in Virtual Worlds* (2010). Several articles in the *International Journal of Internet Research Ethics* (*IJIRE*) also discuss these issues, including Grimes, Fleischman, and Jaeger (2009), McKee and Porter (2009), Reynolds and de Zwart (2010), and Rosenberg (2010). All of these works are well worth referencing before conducting VW research.

As with other forms of online (and RW) field research, there are three key components that we must consider: group accessibility, public versus private spaces, and perceived privacy. Group accessibility refers to how easy or difficult it is to gain entrance into a group or to enter into the space occupied by that group. In general, the larger the group, the easier it is to gain access. Many VW systems, like SL, also have group functions built into the program (e.g., in SL you can join groups related to almost any interest, or create your own). VW groups may be designated as open or closed enrollment, and if someone owns land, they can even set access to that property so that only group members may enter. Of course, not all the groups we wish to research may belong to documented groups.

Public spaces in a VW are not always easy to identify. In SL, almost all the land is privately owned, so there are very few *truly* public spaces. However, *private* spaces in SL are often quite public in nature. The two islands that make up EduNation, for example, are open to any visitors. This is true of many other locations, as already discussed. However, there are also many private homes or entire regions in SL where uninvited visitors would certainly not be welcome and where ban lines may prevent anyone who is not on the permitted list from entering.

Perceived privacy is a topic that has been extensively discussed, but one of my favorite examples comes from Waskul and Douglass (1996), who ask you to imagine having a private discussion with a friend on an RW park bench. When you turn your head and realize that someone has been secretly audiotaping your conversation, and you confront them in outrage, they explain that as a researcher it is their right to do so since this is a *public* park. But in this case, there was a high perception of privacy that was violated by the researcher. This might also happen in a VW where two avatars in a sim that recreates ancient Rome are engaged in a private discussion and discover another avatar listening in. At least in RL it was evident that the bench sitters were being recorded. In SL, however, there will likely be no recorder visible at all – another violation of perceived privacy.

Our goal as researchers should be to *do no harm*. All research carries risk, so our goal should be to lessen the risk to our participants as much as possible, but some types of VW research are much less risky for participants than others. For example, researching the number of visitors to certain regions in SL and what countries those avatars are from carries almost no risk. Collecting language samples from groups of avatars in order to investigate the use of *requests* in VWs also carries little risk, provided the data are sanitized. On the other hand, visiting a support group in SL for survivors of RW sexual abuse is high risk.

As seen in Table 3d.1, groups that are highly accessible (including being located on land that is open) and with a low perception of privacy are lower risk. On the other hand, groups with very low accessibility (e.g., closed membership and on restricted land) that also have a high perception of privacy would be very high risk.

In all studies, researchers must follow the rules and regulations laid out by their research institutions. However, in cases with lower accessibility, in private spaces, and with a high perceived privacy, the first question we must ask ourselves is whether the research is essential. If the research is high-risk, then the potential harm may outweigh the good. If the research is not overly risky, then there are several steps to be followed that are essential when conducting ethical research in VWs:

1. Check the research policies of the VW under study. Some have specific research policies while others have none (but that does not mean you are off the ethical hook).
2. Contact the owner or head of the group you would like to research (if it is an existing group) for permission.
3. Contact the landowner (who may or may not be the group owner) for research permission.
4. Gain informed consent from the individuals you wish to participate.
5. If you are denied at any of the previous stages, your research should not proceed.

All research should maintain high ethical standards. It is sometimes easy to forget that the avatars that inhabit VWs are all being controlled by RW individuals, but we must keep in mind that many of those people not only have large

Table 3d.1 Risk Analysis in VW Research

How Accessible	*How Private*	*Risk*
+ accessible	– perceived privacy	Lower risk (*not* “no risk”!)
+ accessible	+ perceived privacy	↓
– accessible	– perceived privacy	
– accessible	+ perceived privacy	Higher risk

financial investments in their avatars (through VW fees, changes in appearance, clothing, housing, etc.) but also have strong emotional investments. Harming an avatar most definitely can harm the individual behind it. As ethical VW researchers, we must avoid that.

References

Brown, B., & Bell, M. (2004). CSCW at play: "There" as a collaborative virtual environment. Paper presented at the ACM Conference on Computer Supported Cooperative Work. Chicago, IL.

Childress, M. D., & Braswell, R. (2006). Using massively multiplayer online role-playing games for online learning. *Distance Education, 27*(2), 187–196.

Fairfield, J. A. T. (2010). *Avatar experimentation: Human subjects research in Virtual Worlds.* Retrieved from http://works.bepress.com/joshua_fairfield/1

Grimes, J. M., Fleischman, K. R., & Jaeger, P. T. (2009). Virtual guinea pigs: Ethical implications of human subject research in virtual worlds. *International Journal of Internet Research Ethics, 2*(1), 38–56.

Hair, N., & Clark, M. (2007). The ethical dilemmas and challenges of ethnographic research in electronic communities. *International Journal of Market Research, 49*(6), 781–800.

King, S. A. (1996). Researching Internet communities: Proposed ethical guidelines for the reporting of results. *Information Society, 12*(2), 119–128.

McKee, H. A., & Porter, J. E. (2009). Playing a good game: Ethical issues in researching MMOGs and virtual worlds. *International Journal of Internet Research Ethics, 2*(1), 5–37.

Reynolds, R., & de Zwart, M. (2010). The duty to "play": Ethics, EULAS and MMOs. *International Journal of Internet Research Ethics, 3*, 48–68.

Rosenberg, Å. (2010). Virtual world research ethics and the private/public distinction. *International Journal of Internet Research Ethics, 3*, 23–37.

Sadler, R. (2011). *Virtual worlds for language learning: From theory to practice*. Bern: Peter Lang.

Thomas, J. (1996). When cyber research goes awry: The ethics of the Rimm "cyberporn" study. *Information Society, 12*(2), 189–198.

Waskul, D., and Douglass, M. (1996). Considering the electronic participant: Some polemical observations on the ethics of on-line research. *Information Society: An International Journal, 12*(2), 129–139.

Part II

Generating New Data

This page intentionally left blank

4 Generating New Data

Becky Childs

In Part II of this volume, "Generating New Data," we take an in-depth look at methods and processes in creating data and corpora for sociolinguistic analysis. The chapters and vignettes in this section address the various ideological frameworks and applied methods for creating and dealing with new sociolinguistic data, and readers will note the predominant theme of creating data that speaks best to one's research questions and strengths as a researcher.

In Chapter 5, Erez Levon discusses ethnographic data collection. He presents a comprehensive chronological overview of the use of ethnographic data collection in sociolinguistic research and walks readers through the strengths and weaknesses of the method, with reference to specific studies. Focusing on four guiding principles for conducting ethnographic fieldwork, the chapter moves readers through the important processes of accessing a community, interacting with participants, collecting data, and then, importantly, leaving the community. Levon provides specific examples from his own research in Israel to illustrate approaches and methods to be considered when conducting ethnographic research.

The vignettes that follow the chapter on ethnographic data collection present three studies in communities from around the world that have used an ethnographically informed framework for data collection. Vignette 5a, by James A. Walker and Michol F. Hoffman, looks at fieldwork in immigrant communities. They begin by identifying the characteristics of immigrant communities and then work through their study, which looks at the place of immigrant English varieties in the large metropolitan city of Toronto. Walker and Hoffman offer suggestions for ways to effectively collect a corpus in an immigrant community, from using in-community fieldworkers to university-affiliated fieldworkers, and focusing throughout on the importance of building rapport with a community. In Vignette 5b, Rajend Mesthrie describes his work in a migrant and diasporic community in South Africa. He recounts a range of experiences often encountered when engaged in ethnographic study, including how to locate speakers of language variety under study to the strange and at times humorous interactions and missteps that can occur when working with people in a community that is not your own. In Vignette 5c, "Fieldwork in Remnant Dialect Communities," Patricia Nichols describes her work in the Gullah community in the sea islands of South Carolina. Having spent a significant amount of time situating herself in the community before she began data collection, Nichols' experience as explained in this vignette illustrates the importance of gaining community insight prior to initiating a research study.

Chapter 6, "The Sociolinguistic Interview," by Kara Becker, looks intensely at the construct that has been the standard for data elicitation in the field. Becker carefully breaks down this method for data collection, beginning the chapter with a strict definition, moving on to look at the usefulness of this elicitation tool, and highlighting the utility of the sociolinguistic interview, especially when the research question involves an examination of different speech styles. Becker concludes her chapter with a reminder that any elicitation tool must work in hand in hand with one's research question and that the sociolinguistic interview is well suited for any study whose central questions fall within the Labovian variationist paradigm.

Four vignettes accompany Chapter 6, each showing the application of the sociolinguistic interview in a specific study. Vignette 6a, "Cross-cultural Issues in Studying Endangered Indigenous Languages," by D. Victoria Rau, looks at the use of the sociolinguistic interview in an endangered indigenous language, Yami. Rau worked within the Labovian variationist paradigm and adopted a traditional sociolinguistic methodology for data collection, using word lists, texts for intelligibility tests, tests of bilingual ability, and information about language use and language attitudes. While some of these methods were successful, others had to be adapted to fit the community, and Rau recounts these adaptations and the rationales. Finally, Rau leaves us with a four-step approach to data collection, especially if the purpose is to produce materials. In Vignette 6b, Ceil Lucas explains the process of conducting a sociolinguistic interview in Deaf communities and the concerns that linguistic work in communities that have historically had direct ties to education and educational policy may bring up. These considerations are also echoed in Vignette 6c, where Joseph Hill takes a close look at two major issues that arise in collecting sign language data. Specifically, he points to the role of the observer's paradox and the signer's sensitivity to the interlocutor's audiological status and ethnicity. While these issues are of specific concern in the Deaf community, issues of this type (interlocutor ethnicity, age, and social class, among others) are common concerns that ideally are considered before data collection begins. Vignette 6d, "Other Interviewing Techniques in Sociolinguistics," by Boyd Davis, takes a critical look at the sociolinguistic interview, provides alternative methods for data collection, and reminds that there is no "one size fits all" method that will answer any research question.

Although methodological issues such as interview style are important considerations in sociolinguistic data collection, we cannot overlook the methodological concerns related to the actual recording process. In the past 20 years, technology has changed the way in which data collection is done. Reel-to-reel recorders gave way to cassette tape recorders, and soon afterwards the era of digital recording began. In Chapter 7, Paul De Decker and Jennifer Nycz look at the technology of conducting sociolinguistic interviews. With the goal of obtaining a significant corpus of data of high enough quality for sociophonetic analysis, the authors cover appropriate recording equipment (including minimum requirements), technical concerns (e.g., sampling rate, group recordings), and interview storage. After reading this chapter, fieldworkers will be ready to conquer the "technical" aspects of the interview process. Vignette 7a, by Lauren Hall-Lew and Bartlomiej Plichta, gives real-life examples of recording challenges as well as practical considerations for equipment use and choice when in the field. The experiences and advice in this vignette are invaluable and are important to consider before entering the field.

While spoken language data has been a primary focus of many studies in sociolinguistics, the use of written language data, especially data collected by surveys, has also had an important place in the field. In Chapter 8, "Surveys: The Use of Written Questionnaires in Sociolinguistics," Charles Boberg looks at the role of surveys in sociolinguistics and the strengths and weaknesses of this method. Sociolinguistic surveys have the advantage of gathering a large number of participants with a relatively quick response and collection time; however, as Boberg points out, they are limited in the type of data that they can collect. Pointing to specific survey-based studies, especially those on Canadian English, Boberg shows how surveys can be used and discusses how different methodological choices can drive the selection of variants for a survey.

Vignette 8a, by Kathryn Campbell-Kibler, extends the survey method beyond linguistic variable collection to speaker attitude and evaluation study. She shows how speaker evaluation studies are a specialized form of survey that can work alongside other sociolinguistic methods, in order to provide a more robust picture of a language variety and speaker group. Campbell-Kibler moves us through the various steps of setting up a speaker evaluation study, from stimuli to tasks and context, all the while focusing on the goal of having accurate and understandable results. Vignette 8b, by Naomi S. Baron, looks at online surveys as a method for data collection, a viable and attractive method for data collection for many researchers. Baron discusses her work in collecting online data in communities from around the world and points out variation in cultural considerations. Most importantly, she notes that we must consider the responses of our participants when we design surveys, as we do not all have the same cultural assumptions.

Closing this second part of the book, Chapter 9 looks at the use of experiments as data-generating resources. Cynthia G. Clopper begins the chapter by noting that experiments and the data that they create can either stand alone or be used alongside data generated by other collection methods or collected from other sources. Focusing on both production and perception experiments, Clopper demonstrates the utility of each method and, more importantly, the types of research questions that they can answer. She walks us through the ways that experiments can be used in the field and covers the advantages of experiments, while also cautioning us as to the disadvantages of using this approach (including limitations on the naturalness of data and having access to equipment of the quality needed) that must be thought through.

The chapters and vignettes in Part II work together to provide an overview of the process of collecting sociolinguistic data, which can be overwhelming in the choices that are available to researchers and the detail needed to consider each methodological approach. Allowing the research question to drive the primary data collection method and even supplemental data collection will help researchers decide on a framework that is best suited to their study. With an understanding of the technological necessities for conducting ideal recordings and the potential missteps that can occur and that do occur when working with real language and real speakers, researchers will also be aware of what awaits us in the field and ready to respond, even when a miscue happens.

5 Ethnographic Fieldwork

Erez Levon

Since its inception as a field, sociolinguistics' primary goal has been to account for observed patterns of language variation and language change. To that end, sociolinguists have focused attention on understanding the properties of both the linguistic systems in which variation occurs and the broader social matrices in which those systems are embedded. The reason for this dual focus is that, from a sociolinguistic perspective, language never exists in a social vacuum. In the words of Labov (1963, p. 275), "one cannot understand language [variation and] change apart from the social life of the community in which it occurs." In this chapter, I discuss ethnographic fieldwork as one of the principal methods through which sociolinguists come to apprehend the social lives of the communities and community members they study. I begin with a brief overview of what the term "ethnography" can be taken to mean, before turning to a more practical discussion of the various methodological steps that conducting ethnographic fieldwork involves.

Introducing Ethnography

Selecting one definition from many, let us describe ethnography as the study of the "social organization, social activities, symbolic and material resources, and interpretive practices characteristic of a particular group of people" (Duranti, 1997, p. 85). In other words, ethnography is the study of how the members of a community behave and why they behave in that way. Ethnography is normally conducted through prolonged observation and direct participation in community life in the form of ethnographic fieldwork.

Blommaert (2007) describes ethnography as a *methodology* – a broad theoretical outlook that extends beyond the particular methods most often associated with it (such as participant observation). Blommaert maintains that, as a theory, ethnography is built on two crucial and interdependent assertions. The first is ontological: that all social events, including language use, are necessarily contextualized (spatially, temporally, historically, or otherwise) and potentially multivalent. Put another way, events are always connected to other events, and their meanings are multiple. This ontological assertion then gives rise to ethnography's second foundational principle, this time an epistemological one: that knowledge of these events is always situated within the individual, group, or community in which the event took place and is hence subjective. Ethnography therefore rejects the notion of an objective understanding of social action and instead insists that knowledge is, to a certain extent, always contingent. Ethnographic knowledge – that is, knowledge gained from ethnographic research – is always interpretive: it depends as much on the "reality" of the event as it does on the reality that was perceived by those who participated in and/or observed the event.

The inherently interpretive nature of ethnography is the methodology's greatest strength as well as its greatest potential weakness. By rejecting the idea that knowledge of social practice can exist independently of the people engaged in that practice, ethnography works to avoid the twin pitfalls of reductionism and essentialism that are endemic to much of the research in a so-called positivist paradigm (Cameron, Frazer, Harvey, Rampton, & Richardson, 1992). In other words, ethnography attempts to go beyond explanations that are based solely on externally identifiable characteristics, such as gender or social class, and instead pays close attention to how individuals engaged in social action construct and interpret their own practice. Yet, at the same time, too much reliance on practitioners' own understandings can push ethnography toward an untenably strong relativist position, one that gives individuals free range as agents and fails to recognize the larger social, institutional, and ideological forces that shape interaction. As Cameron et al. (1992) put it, "whatever they say, people are not completely free to do what they want to do [or] be what they want to be" (p. 10). It is therefore necessary to "tie ethnography down" (Rampton, 2007; Rampton et al., 2004) and make it accountable to individual subjective experience while simultaneously locating those experiences within a structured and independent social reality. This kind of approach is what Cameron et al. (1992) call *realism*.

For sociolinguists, engaging in realist ethnography means examining the use of locally meaningful linguistic forms within a community of speakers and then detailing how that use is inextricably linked to larger linguistic patterns and distributions in a given social context. Consider, for example, Eckert's (1996) analysis of the emergence of linguistic style among a group of pre-teen girls in a San Francisco Bay area middle school. Eckert describes how the girls adopt certain linguistic and other stylistic devices as a way of projecting more mature, teenager-like personae. Specifically, Eckert discusses the girls' pronunciations of the low front vowel /æ/, which they produce as consistently backed and lengthened when occurring before a nasal (as in *ban*). This practice is significant because it parallels what adult Chicano speakers in Northern California are doing with this vowel, where they avoid the majority Anglo pattern of raising /æ/ before nasals and instead produce a distinctive backed and lengthened version. What is interesting about what the girls that Eckert discusses are doing is that they are not using backed /æ/ as a way to perform Chicana identity (as outside observers might assume). Instead, Eckert argues convincingly that the girls are in effect borrowing the salient Chicano English /æ/ pattern and recontextualizing its meaning as a way of projecting social and sexual maturity. In other words, /æ/ is a salient social marker of Chicano identity in the broader social context of California. These girls, however, are reinterpreting /æ/ in their local practice and deploying it as a symbol of "tough" and "mature" femininity. The point is that the girls would not be able to use /æ/ to do this kind of very local identity work if the variable did not already carry the wider indexical connotations that it does. And, even more importantly for our purposes, it is only by adopting an ethnographic approach that Eckert is able to identify the complex layers of social meaning that account for the observed variation in the girls' linguistic practice.

With the preceding conceptual discussion of ethnography as background, let us now turn to a detailed description of how to go about conducting ethnographic fieldwork. While there is no set template for how fieldwork must be done, there are certain key stages involved in all ethnographic field projects, and the remainder of the chapter introduces and discusses these stages in turn. Before we get to that, I briefly mention some general methodological principles that are useful to bear in mind when planning and conducting ethnographic fieldwork.

1. *Be prepared*. Before you enter the field, develop as much prior knowledge about the community you plan to study as you possibly can. This knowledge can come from prior academic research on the community (whether from linguistics or other disciplines). It can also come from your own interactions with the community in a non-academic setting or from general information you are able to collect from reliable sources. You should use this background knowledge to develop a preliminary fieldwork plan. Doing so will help to ensure that you are able to make the most of what is normally a relatively limited amount of time in the field.
2. *Be adaptable*. No matter how much preliminary planning you do, you can never anticipate every situation you will encounter while in the field. For that reason, it is important to remain flexible and open to change. It could be the case, for example, that you do not gain access to a particular group of speakers or that you are unable to record a certain interaction. If this happens, you need to be able to think on your feet and adjust your data collection protocol accordingly. On a more interpretive level, it is also often the case that your participants will defy your initial expectations of them in some way. This is a normal part of the process of developing a more intimate insider's perspective (what is called an *emic* viewpoint). It is therefore important to remain open to the possibility that things are not how you originally imagined them to be when you were still an outsider (i.e., with an *etic* viewpoint). (See Harris, 1976; Hymes, 1974, for a foundational discussion of the etic–emic distinction.)
3. *Be mindful*. Observation is at the center of the ethnographic project. Your job as an ethnographer is to identify meaningful patterns in behavior that others take for granted. For this reason, it is crucial to remain attentive to even the most seemingly insignificant details of the interactions you observe. You never know which of the hundreds of little things that your participants do every day – such as tossing their hair or putting on eyeliner – will turn out to be culturally important. This also means that you need to adopt a very deliberate approach to keeping fieldnotes and other catalogs of your experiences.

4. *Be respectful.* Finally, you should never forget that your research participants deserve your respect and gratitude for allowing you privileged access to their lives. Be mindful of their time and of their feelings, and realize that they are the ones who ultimately control how much access you will have. Remember too that your participants' own opinions and interpretations of their practice are worthwhile and deserve to be considered (even if you end up disagreeing with them). Lastly, respect your participants enough to share your findings (as appropriate) with them. As Cameron et al. (1992, p. 24) state, "if knowledge is worth having, it is worth sharing." It is the least you can do for your participants to thank them for their generosity.

Accessing the Community

The first step in all fieldwork is to gain access to the community you want to study. Obviously, the details of how you do so will vary greatly depending on the community in question and your relationship to it. For some researchers, gaining access is relatively straightforward since either they already know members of the community or are members themselves. More commonly, however, researchers stumble across a community that they were previously unacquainted with and then have to find a way to orchestrate an introduction for themselves. Kulick (1998), for example, describes how he first got the idea to study Brazilian transgendered prostitutes, or *travesti*, when he saw a group of them from a bus he was riding on one night in Salvador. Kulick had never heard of *travesti* before (nor did he know at the time that that is what they are called), but something about what he saw from the bus that night piqued his interest. After that initial sighting, Kulick began to read up on previous research on *travesti* and to get in touch with the scholars who had conducted this work. Kulick also contacted a local LGBT organization to see whether they could provide him with any additional information. This preparatory research eventually led Kulick to be introduced to a member of the community, who then helped him arrange his fieldwork.

As Kulick's experience demonstrates, ethnographers often gain access to their communities via the "friend of a friend" method, where the researcher is introduced to a community member by a mutual friend or acquaintance (Milroy, 1987; Milroy & Gordon, 2003). This mutual acquaintance could be a friend or family member of someone in the community, or an academic or some other service provider who has worked with the community before. The idea is that this mutual acquaintance allows the researcher to make first contact with a community member. This initial contact person in the community then introduces the researcher to her or his friends, who in turn introduce the researcher to their friends, and so on. The gradual development of a participant population through social networks in this way is called *snowball sampling* (Goodman, 1961), and it is perhaps the most common method used in the social sciences. When I first began fieldwork in Israel, for example, I was lucky to have a number of politically active friends who were able to introduce me to members of some of the activist groups I ended up studying (Levon, 2010). These initial contacts invited me to come along to group meetings with them, where they introduced me to other group members. In addition to allowing me to gain access relatively quickly to the various groups I aimed to study, using mutual friends to arrange introductions meant that I was able to enter the groups as a known entity (“so-and-so’s friend”) rather than a complete stranger. This in turn meant that I was accepted more readily as a regular participant in the groups’ interactions.

While the friend of a friend method is a very good way to gain access to a community, it is by no means the only one. Another commonly used method is to go through a community’s official (or semi-official) “brokers”: people whose job it is to manage relations between the community and outsiders (Schilling-Estes, 2007). Brokers are often political or religious leaders, individuals who command some sort of community-wide authority. For this reason, they are usually very successful in encouraging community members to participate in a study that they deem worthwhile. There are, however, a couple of things to bear in mind when working with a broker. First, brokers may not grant a researcher access to all aspects of community life, especially if some of those aspects are considered less presentable than others. Brokers normally want to present their communities in the best possible light and so will sometimes deny access to those people or things that they feel might harm the community’s image. Second, and relatedly, community members may not behave as naturally with a researcher who was introduced to them by a broker as they would with a researcher introduced by a friend, often because of an increased sense of formality that the involvement of a broker can create. In spite of these concerns, however, brokers can be an invaluable resource for ethnographers working in an unfamiliar community. In certain cases, it may even be necessary to go through brokers before research can begin. Schilling-Estes (2007), for example, reports how all scholars wanting to conduct research on Smith Island, Maryland, in the 1980s had to coordinate their work through the island’s pastor, who acted as a sort of gatekeeper to the entire community.

Finally, sometimes researchers have neither friends nor brokers that they can rely on. In these situations, it is up to the researcher to gain access to the community on her own. Gaudio (2009) describes how when he first became interested in ‘*yan daudu*, Nigerian men who talk and act like women, he began to attend various ‘*yan daudu* events on his own (he was in Nigeria working on another research project at the time). After attending a number of these events, Gaudio managed to chat with and get to know some of these men, who in turn then introduced him to others in their community. Essentially, Gaudio adopted a snowball methodology as discussed above. He, however, was obliged to do it the hard way, without an initial entrée into the community.

No matter which method one ends up using, the goal is to gain as much access as possible to the community in question. For this reason, it is important to think carefully about what method is right for you and your project. Remember too that your initial contact can in many ways shape your positionality within the community for the duration of your fieldwork. We turn to a more detailed discussion of how ethnographers position themselves in relation to their participants in the next section.

Interacting with Participants

The role of ethnographer in a community can at times be a difficult one to negotiate. While your goal as a researcher is to become as much of an "insider" as possible, it is important to realize that most people are not accustomed to having someone observe and comment on what they do. And while you want to be able to get as close to the inner workings of a community as you can, it is also crucial to maintain a certain amount of analytical distance so that you can critically reflect on what you observe. Being an ethnographer is about finding a balance between "insider" and "outsider" status and making sure that your participants understand and are comfortable with your role.

Being viewed as an outsider can also have its advantages. Kulick (1998) describes how the fact that he is not Brazilian meant that he was not aware of the many negative cultural stereotypes regarding *travesti* that circulate in Brazil, which made his *travesti* participants more comfortable with him. In my work in Israel (Levon, 2010), I was initially perceived by my participants as an Israeli gay man (my name is recognizably Israeli, and I was able to converse with all of them in fluent Hebrew). While on the one hand this perception was positive, since it allowed me a sort of guaranteed insider status, on the other hand it also made some participants wary of me because they assigned me all the cultural baggage that goes along with the "Israeli gay man" moniker. In those cases, I worked hard to promote the "American academic" side of who I am (for example, by talking about how I grew up in the United States) since it allowed me to be perceived as distant enough from the Israeli social context to gain access to those communities as well.

This last point about how I was perceived in Israel is an important one, since it highlights the fact that, as ethnographers, we are observed and critically reflected upon by our participants as much as we observe and critically reflect upon them. Mendoza-Denton (2008), in her ethnography of Latina youth gangs in California, discusses how when she began interacting with the girls in her study, her participants tried to define her according to their locally meaningful cultural categories as either a *Norteña* ("Northern" Chicana) or a *Sureña* ("Southern" Chicana). Mendoza-Denton resisted this classification, however, since being read as either *Norteña* or *Sureña* would have meant that she would not have been able to have access to the other group. At the same time, it was crucial that Mendoza-Denton take part in at least some of both groups' cultural practices so that she could maintain a certain level of insider status with each. One of the ways in which Mendoza-Denton did this was to allow the girls from each group to do her makeup in their preferred style. She was therefore able to participate in at least part of both groups' social worlds while simultaneously inhabiting a sort of liminal space between completely "in" and completely "out."

Finally, it is in thinking about how we are perceived by our participants that issues of how we access a community in the first place become important. As mentioned above, if an ethnographer gains entry to a community via "official" channels (such as a broker or other community leader), it may be difficult to shake off the aura of officialdom that such an introduction may carry with it.

This issue becomes particularly salient when working with marginalized communities or with any group in which there is a powerful "us versus them" mentality. When beginning her fieldwork on the island of Martha's Vineyard, for example, Josey (2004) was sensitive to the fact that there was a prominent ideological divide between local, year-round residents of the island and those who vacationed there. In order to escape categorization as another outsider coming to the island, Josey decided to change her physical presentation of self – for example, by removing her makeup – and to offer her services as a babysitter for local families. In doing so, Josey was able to ingratiate herself with the local community and escape being labeled as another "big city person" coming to the island.

On the flip side, coming into a community as a "friend of a friend" or as a near-insider can also have its disadvantages. While this more intimate status certainly brings with it special rights and privileges, it can also bring certain obligations. As Schilling-Estes (2007) puts it, "if one capitalizes on people's friendships, it is typically expected .. that one will give something back in return" (p. 180). (For more discussion of researcher-community ethical models and of linguistic gratuity, see Trechter, Chapter 3; Nichols, Vignette 5c; Ngaha, Chapter 16; and Starks, Vignette 17c.)

In sum, as an ethnographer it is important to be aware of how you are perceived in the community you are studying and do everything you possibly can to ensure you are perceived in a way that is most conducive to collecting the data you need. How to go about collecting data is the subject of the next section.

Collecting Data

Once in the field, the most common methods for collecting ethnographic data are participant observation, individual and/or group interviews, and analysis of cultural artifacts. Conducting interviews is discussed elsewhere (see Becker, Chapter 6; Rau, Vignette 6a; Lucas, Vignette 6b; Hill, Vignette 6c; and Davis, Vignette 6d), so I will not devote much space to this topic here. Instead, I focus on some of the primary issues to consider with respect to participant observation and analyzing artifacts.

According to Richards (2003), ethnographers should concentrate on four main areas of social interaction: the physical setting of events, the systems and procedures that are followed at these events, the people who take part in these events, and the practices (including language) that are observed at these events. As in any other type of scientific research, the goal of observation is to identify systematic patterns of behavior that can be correlated to some external factor (such as the physical setting, for example). As I mention above, you never can be sure what will end up being significant. It is therefore advisable to cast your net as widely as possible and note as many details as you can of the various events you observe. When it comes to analyzing your data, you will be able to pare down your observations and hone in on what is truly culturally meaningful in what you observed. Providing a complete account of your observations will also assist you, after your fieldwork is complete, in writing engaging ethnographic descriptions, in which issues of character and setting will play a prominent role.

To do this, make sure to maintain detailed and well-organized fieldnotes. Since you will be not only observing interactions but participating in them as well, it may be difficult to take notes during the events in question. If this is the case, you should try to note down your recollection of the events as soon as possible (details tend to be the first things we forget). Even if you are able to record interactions, it is still a good idea to write down your observations from memory and then compare them to what you see or hear in the recording. This is because recordings, by definition, provide an "outsider" perspective and will lack some of the more emic (i.e., insider) details that you will have experienced as a participant. Also, to assist with your analysis further down the line, it is a good idea to come up with a standardized system for keeping fieldnotes. You could, for example, work according to a template, where you provide details of the setting, the people, the clothing, the behavior, etc. for each event in the same way. This procedure will allow you to compare and cross-reference your events more easily. Whatever method you choose for recording your observations, make sure that it is transparent enough for you to retrieve the necessary information months, or even years, after the event has taken place. Also make sure that you store your notes in multiple places, using multiple forms of media (handwritten observations, typewritten notes, audio recordings, etc.), and that you back up all files in multiple locations.

In ethnography, data collection and analysis are not as clearly separable as they are in other kinds of empirical methodologies. During participant observation, it is hugely beneficial to be analyzing your findings along the way. This process helps you to determine the more precise direction that your ethnography will take, for example by allowing you to identify particular practices or situations that you want to focus on. It will also help you to determine what the right time to conduct individual and/or group interviews is and will help you to write a more effective modular interview schedule. Unlike what is typically the case when one is conducting sociolinguistic interviews (see Becker, Chapter 6), ethnographic interviews are not normally conducted the first time you meet a research participant. Rather, they tend to happen once you have been participating in and observing a community for some time. In Israel, for example, I began conducting interviews with my participants approximately four to five months after I began my observations. Waiting allowed me to have a much greater degree of familiarity with my participants and their social worlds than would have been possible had I conducted the interviews earlier. At the same time, it meant that I already knew the answers to many of the questions that are asked in standard interviews (e.g., about interviewees' families, friends, activities). I therefore had to devise an interview schedule that was not only suitable for my research questions but also adapted to the level of intimacy and rapport I had already developed with my participants.

Related to this issue of a researcher's relative intimacy within a community is the ever-present question of how to explain the goals of your project in the first place (see also Trechter, Chapter 3, and Ngaha, Chapter 16). It is obviously important to be forthright with your participants and to tell them what it is you hope to obtain from them. You must, however, also be aware of any social or cultural sensitivities that exist and of the fact that by virtue of entering a community as a researcher, you bring with you a form of intellectual and institutional power (see the earlier discussion). Unfortunately, there is no magic recipe for how best to explain your work while ensuring that your participants remain at ease, and, more likely than not, you will encounter a number of relatively awkward "who are you and what are you doing here?" kinds of questions. In situations like these, it is usually best to try to answer as honestly and unassumingly as you can. As Schilling-Estes (2007) notes, "most people will tolerate a friendly stranger far more readily than someone who pretends to be 'one of them'" (p. 178).

In addition to participant observations and interviews, the final principal data source in ethnography is the collection and analysis of cultural artifacts, which are any physical materials, images, broadcasts or other media products that relate to your participants' lives. In addition to my direct observations and interviews in Israel, I also conducted daily media monitoring for newspaper articles, films, and television programs related to Israeli lesbian and gay life. This process allowed me to gain a fuller understanding of the broader context in which my participants were located and provided me with a more complete interpretive framework for understanding their social practice.

Leaving the Field

Before I left to begin my fieldwork in Israel, I asked Don Kulick, one of my supervisors at the time, how I would know when I had enough information to leave the field. He said that you know that you are done with your fieldwork when you start to know the answers to your own questions before your participants have a chance to reply. His advice is not meant to imply that we can ever hope to develop a "perfect" understanding of the communities that we study or that no questions will remain. Rather, all ethnographers reach a point (sometimes called the "saturation point") where they feel as though they have gained enough insight into a particular group to be able to come to certain generalizations about the group's systems of belief and associated cultural practices and norms. It is nevertheless always a good idea to run your conclusions by your participants to see what they think of them. You can gather participant feedback in a variety of ways, including follow-up interviews, playback sessions (e.g., Rampton, 1995), and non-technical summaries of your research findings. Even though these kinds of activities may not be your first priority after completing the fieldwork phase of your project, they represent an excellent way to give back a little something to the community you studied and will help to ensure that the channels of communication remain open if one day you wish to return to that community for further research.

Once you have finished collecting and interpreting your data, the next step is to prepare your findings for presentation. Unlike other types of social scientific reports, ethnographic descriptions tend not to include clearly divisible sections (such as "Methods" and "Results") and are in certain instances more similar to pieces of creative writing than to scientific writing. This is because of ethnography's insistence on maintaining an *emic* perspective (see p. 71), in which you as a researcher must work to “convince the reader [of your interpretation] through drawing him or her into the world of the participants and sensing the believability of that world” (Goldbart & Hustler, 2005, p. 17).

Ethnographic fieldwork is not the simplest way of obtaining sociolinguistic data. In fact, many researchers would agree that it is one of the most personally and intellectually challenging methods of data collection. It is also, however, one of the most rewarding. In this chapter, I have provided an overview of ethnography for sociolinguists, including a discussion of some of the practical issues that arise when conducting this type of fieldwork. Whether you end up using ethnographic methods in your work or not, I hope this discussion has succeeded in demonstrating the benefits of an ethnographic approach for developing a nuanced understanding of sociolinguistic meaning and ultimately for providing a robust account of variation in observed linguistic practice.

References

Blommaert, J. (2007). On scope and depth in linguistic ethnography. *Journal of Sociolinguistics, 11*, 682–688.

Cameron, D., Frazer, E., Harvey, P., Rampton, B. B. H., & Richardson, K. (1992). *Researching language: Issues of power and method*. London: Routledge.

Duranti, A. (1997). *Linguistic anthropology*. Cambridge: Cambridge University Press.

Eckert, P. (1996). Vowels and nail polish: The emergence of linguistic style in the preadolescent heterosexual marketplace. In N. Warner, J. Ahlers, L. Bilmes, M. Oliver, S. Wertheim, & M. Chen (Eds.), *Gender and belief systems* (pp. 183–190). Berkeley, CA: Berkeley Women and Language Group.

Gaudio, R. P. (2009). *Allah made us: Sexual outlaws in an Islamic African city*. Malden, MA: Blackwell.

Goldbart, J., & Hustler, D. (2005). Ethnography. In B. Somekh & C. Lewin (Eds.), *Research methods in the social sciences* (pp. 16–23). London: Sage.

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics, 32*, 148–170.

Harris, M. (1976). History and significance of the emic–etic distinction. *Annual Review of Anthropology, 5*, 329–350.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Josey, M. (2004). A sociolinguistic study of phonetic variation and change on the island of Martha’s Vineyard. (Unpublished doctoral dissertation). New York University.

Kulick, D. (1998). *Travesti: Sex, gender and culture among Brazilian transgendered prostitutes*. Chicago: University of Chicago Press.

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

Levon, E. (2010). *Language and the politics of sexuality: Lesbians and gays in Israel.* New York: Palgrave Macmillan.

Mendoza-Denton, N. (2008). *Homegirls: Language and cultural practice among Latina youth gangs*. Malden, MA: Blackwell.

Milroy, L. (1987). *Observing and analysing natural language*. Oxford: Blackwell.

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

Rampton, B. (1995). *Crossing: Language and ethnicity among adolescents*. London: Longman.

Rampton, B. (2007). Neo-Hymesian linguistic ethnography in the United Kingdom. *Journal of Sociolinguistics, 11*, 584–607.

Rampton, B., Tusting, K., Maybin, J., Barwell, R., Creese, A., & Lytra, V. (2004). UK linguistic ethnography: A discussion paper. Retrieved from http://www.ling-ethnog.org.uk

Richards, K. (2003). *Qualitative inquiry in TESOL*. Basingstoke, UK: Palgrave Macmillan.

Schilling-Estes, N. (2007). Sociolinguistic fieldwork. In R. Bayley & C. Lucas (Eds.), *Sociolinguistic variation* (pp. 165–189). Cambridge: Cambridge University Press.

Vignette 5a
Fieldwork in Immigrant Communities

James A. Walker and Michol F. Hoffman

For us as sociolinguists, fieldwork not only provides us with the raw data we work with but is also an important step in understanding the community we are studying. While most sociolinguistic studies follow the tradition of dialectology in focusing on established populations, more and more studies are paying attention to immigrant communities.

What do we mean by an "immigrant" community? From William Labov's earliest work on Martha's Vineyard, sociolinguists have investigated language in ethnically and linguistically diverse contexts, but there has been a tendency to exclude people who are not felt to be fully part of the speech community (see Horvath, 1985; Labov, 1966). At what point does an immigrant community become an ethnic group within the larger community? Can we even consider people who have emigrated from the same location or who speak the same minority language as forming a community? We are not posing these questions to confuse you (or ourselves); rather, we want to point out the importance, when studying "immigrant" communities, of bearing in mind the different contexts of the first generation and subsequent generations.

Immigration is of course nothing new, but recent improvements in access to mobility, coupled with social and economic persecution, wars, and economic hardship, have all brought about an unparalleled scale of global migration that has altered the ethnolinguistic landscape of many cities around the world. English-speaking cities in North America and Australia have historically been immigrant communities, but more recently, cities in the United Kingdom and elsewhere in Europe have become increasingly multiethnic and multilinguistic.

Choosing a Community

You may have different motivations for conducting a sociolinguistic study of an immigrant community. Considerations include:

- a particular language or ethnic group and their status in a diaspora context;
- concerns about difference from and (non-)assimilation to the majority group;
- robustness of representation relative to both the majority group(s) and other minority groups;

- settlement pattern(s) and community cohesion;
- other practices that set them apart or unite them with other groups.

Of course, because immigrant communities are not all the same, we would not expect to see the same sociolinguistic patterns from location to location.

For example, our study of ethnolinguistic variation in Toronto English (Hoffman & Walker, 2010) was motivated by public concerns about the status of the English spoken by children with home languages other than English. Since Toronto is a city characterized by a high degree of ethnic and linguistic diversity, trying to get a representative sample of all ethnic groups in the city in one go was too daunting a task. Instead, we chose to focus the first stages of our project on the communities with the most robust representation, who also happen to be the most socially salient. "Chinese" form the largest group, even though not all Chinese Canadians have the same heritage language or regions of origin. Given the history of Chinese settlement in Toronto, we decided to limit our sample to speakers who came from Hong Kong (or nearby Canton) and had Cantonese as their first or heritage language.

If we compare Toronto with other situations of immigrant communities, we start to see some important differences. In Toronto, different immigrant groups have tended to settle in particular geographic areas, leading to some identifiable "enclaves" (e.g., Little Italy, Chinatown). In Sweden, although there are neighborhoods in the largest cities (Stockholm, Gothenburg, Malmö) that are identified as "ethnic" or "immigrant," they are not dominated by any one ethnolinguistic group. Even in Toronto, not all groups settle in identifiable neighborhoods these days. For example, Spanish speakers are spread throughout the city.

These differences underline the importance of understanding the demographic characteristics of each community and each location. A good place to start is by looking at census data or similar data collected by governments and community research institutions. Non-governmental advocacy and support centers may also collect community-specific information and can be valuable resources not just at this stage, but also after you embark on your fieldwork. Some of this information is available online, but in many cases you may have to visit the institution or request documents. Depending on the nature of the census data, it may be relatively easy to identify patterns of immigration, both in general and for specific languages or regions of origin. You may also be able to find ethnic and linguistic information about specific neighborhoods, such as respondents' mother tongues, ethnic origins, and home languages. Bear in mind that this information is self-reported, and people will often respond to the same question in different ways. The wording of the question is not always the same in each country's census, so be careful to take that into consideration in interpreting the responses.

Familiarity with your community (or communities) before you enter it (or them) is important for a variety of reasons, relevant both to your research questions and to your fieldwork experience. This may seem obvious, but keep in mind that even seasoned fieldworkers can find themselves in situations they have not previously encountered. It is impossible to prepare for all contingencies before embarking on your fieldwork, so you will need to keep your eyes and ears open and be prepared to revise your strategy.

Entering the Community

Once you have collected some background data, your point of entry to the community will depend on your status and connection. If you are a member of the community, of course you will have an insider's knowledge and connections. However, you may be led to rely on your family and friends, which may not give you a representative sample of the entire community. This is especially true if there are deep differences between subgroups of the community that don't get along, and you happen to be from the wrong subgroup! Go beyond your extended social networks and be open to reexamining any assumptions about the community you may have made as an insider.

If you are not a member of the community, you may still be able to gain access, but the approach will be different. Having friends or acquaintances who are members of the community is an important resource. Apart from offering valuable information, they may give you credibility with others, as a "friend of a friend." As with community members, though, there is a danger of relying too much on your friends' extended social networks. You can also enter the community by approaching community centers, groups, outreach or advocacy organizations, religious institutions, or, depending on the age of your consultants, schools. Such groups can be good first points of contact, as they can facilitate introductions. You can begin by spending time at the organization and building connections with staff and volunteers. Volunteering can also be a good way to build connections.

Building trust is essential when doing work in any community, but it is even more important in immigrant communities with which you may be unfamiliar. Immigrant groups, especially more recent arrivals, are relatively vulnerable because of language barriers, socioeconomic status, discrimination in work and housing, legal status, and public safety. Keep in mind that immigrants likely find themselves in socioeconomic situations quite different from those of their countries of origin. If you present yourself in too formal or official a manner, it may accentuate the power imbalance and hinder the success of your research.

For example, Michol conducted her dissertation research on the Spanish of Salvadorian youth in Toronto, a minority language in an immigrant community (Hoffman, 2004). As a non-community member, she relied on help from community organizers and schools. The youth director of an outreach and advocacy organization introduced her to many consultants. She spent time in the community organization and schools, chatting and interacting informally with the youth. Throughout their interactions, she took care to convey the message that even though she was the researcher, they were the experts on their community. Although a non-Latina Anglophone, Michol was able to conduct successful sociolinguistic interviews in Spanish.

In some cases, the needs of your project may be better served by working with a community member as a research assistant or co-investigator. This approach lessens the control you have over the fieldwork experience, but if you are looking at language use in in-group contexts, it may be the only way to get the kind of data you want. In our project on ethnolinguistic variation in Toronto English and Michol's project on different varieties of Spanish spoken in Toronto, we wanted the type of language that people use when talking to other people who share their ethnic and linguistic background. We employed student fieldworkers who are themselves members of the communities.

Fieldwork in immigrant communities presents challenges not faced by fieldwork in more "mainstream" communities, but a willingness to question your assumptions and flexibility in your approach will go a long way.

References

Hoffman, M. F. (2004). *Sounding Salvadorean: Phonological variables in the Spanish of Salvadorean youth in Toronto*. (Unpublished doctoral dissertation). University of Toronto.

Hoffman, M. F., & Walker, J. A. (2010). Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change, 22*, 37–67.

Horvath, B. (1985). *Variation in Australian English*. Cambridge: Cambridge University Press.

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

Vignette 5b
Fieldwork in Migrant and Diasporic Communities

Rajend Mesthrie

From 1981 to 1983, I undertook fieldwork for my PhD thesis, which aimed to document the Bhojpuri language of South Africa as spoken in the province of Natal (which I will label KZN, since it is now known as KwaZulu-Natal). No previous work had been done on this language, and indeed no one had ever called it by that name; it was officially recorded as "Hindi" (and still is). My motivation was that this variety was sociohistorically very interesting. Rather than being a non-standard (others even said substandard or "kitchen variety") of Hindi, the prestigious and co-national language of India, the variety that had evolved in South Africa since 1860 was in fact aligned to a group of languages related to but different from Hindi. Moreover, the language in itself would prove something of an oral archive, documenting the history of Indian plantation workers and their descendants over their (then) 120-year history in South Africa, via loanwords, semantic changes, neologisms, etc.

My main aim, then, was to carry out informal interviews that would provide documentation of the structure of Bhojpuri, reflect the fact that it was a blend (or "koiné") of several closely related north Indian languages, and give evidence of more recent changes arising out of contact with the other languages of the province, viz. English (the language of the formal economy), Zulu (the indigenous language of the province), and Tamil (from south India). Lastly, I knew that Bhojpuri and other Indian languages of the country were on the decline, as English was becoming the language of younger children (often the sole language, together with the Zulu pidgin, Fanakalo). Since I was aware of some regional variation within the province, I aimed to elicit this systematically as well. The syntactic and phonetic differences between the uplands and the coast would turn up in the interviews themselves; for vocabulary items I had to prepare a special list. Add to all this the fact that most Bhojpuri speakers (and almost all I interviewed) were second- or third- (sometimes fourth-) generation speakers who had never been to India, and you will see that it was vital to collect good data to do justice to a little-known, unstudied, and unprestigious koiné from the time of its inception to its demise. What follow are my recollections 30 years on: I am not sure that with hindsight I would have done things exactly the same (in fact, the bureaucracies and ethics committees probably wouldn't allow it, anyway), but I hope my experiences would prove, if not cautionary, then certainly entertaining for my own serendipitous experiences.

Those were the days before the big research grant and generous scholarships to graduate students. The small grant I received from national funders paid my tuition, and I carried the costs of fieldwork from my own modest salary (did I mention I was doing the research part-time?). One advantage I had was that I could speak Bhojpuri, though not as well as the previous generation. It was also the heyday of apartheid, the notorious South African practice of separation of the races. There were two spin-offs for me, though: first, it was easier to find the areas where Indians resided; and second, the resultant close-knit nature of the province's Indian communities meant that I would be readily welcome in people's homes. For this reason, I often didn't make appointments. (Many homes didn't have phones, anyway.) All I needed to do was drive up and speak to people – well, almost. Sometimes it was necessary to make an initial contact who would introduce me to a cross section of Bhojpuri speakers: often this person was a distant relative (in both senses: kinship and mileage), sometimes a friend of a friend or a close relative. I avoided priests, as they were custodians of the formal Hindi that wasn't my focus, though my sample did include some priests, who it turned out were bilingual in Hindi and (although they wouldn't admit it) Bhojpuri too. I worked in some rural areas without an initial contact. The ethnic solidarity was further enhanced by my expressing an interest in Hindi (as I had to call the koiné).

But without an initial contact, how would one know which were the Bhojpuri homes among the Tamil, Telugu, Gujarati, Urdu, etc.? By and large, Bhojpuri speakers had red flags hoisted in their yards, signifying homage to the god Hanuman, whereas south Indians hung a string of mango leaves or marigolds at their doorway. Muslim homes hoisted green or white flags. I had to obey community dynamics when introducing myself, speaking first to the head or senior person of the house (in either English or Bhojpuri or both). The initial greeting "*namaste*," said with hands together and a slight bow, was then the standard greeting in Bhojpuri homes and again helped establish my bona fides. Most people were incredibly trusting. Their first reaction would be to send me to the local priest, as the expert on matters linguistic. People would be apologetic and say that they didn't really know the proper Hindi. So, my first task was to persuade them that I was after the ordinary Hindi that we all spoke and that I was interested in the family's history, not that of the priest alone. In only one home was there someone who refused to talk: an elderly male who was certain that I was a con man who was there to defraud him. Some were confident that I had really come to sell insurance, but after their initial disappointment took the visit well, giving me of their time and memories. And lots of tea. Community etiquette required putting a pot of tea on the stove first, and then returning to the sitting room (as we called it then) to converse. Community solidarity also meant that you were not asked about preferences over milk and sugar; all teapots came with generous amounts of sugar and milk in them. Since I often did 10 interviews a day, that meant about 20 spoons of sugar for the day. (And I'm not yet diabetic, the last time I checked.)

Most people's attitudes were of friendly indulgence and sympathy for poor university students who had to do all sorts of crazy things to get a degree, including working on a broken, kitchen language that no one took seriously (surely?).

One person later pointedly asked her neighbor, a colleague from my university, "*Oke kām nahi he, kā?*" ("Doesn't he have a proper job?"). Men were more reluctant to be interviewed and would often switch to English, as they didn't want to seem old-fashioned: I had to make a special effort to collar them before they quietly absconded. It is always good to offer favors in return, and I sometimes gave men lifts, or money for a "loose" cigarette or two if they had asked me for a cigarette. Elderly grannies had lots of time on their hands and made the best interviewees – and they knew Bhojpuri very well. They gave me snippets of information about first family migrants from India, the ship's journey, first jobs on plantations or mines, and so forth. I collected some folksongs from the more obliging ones. Etiquette required that I use a female intermediary, either the daughter-in-law of the house or my initial contact in the community, if female, to help when interviewing older females. Since our main topics revolved around migration and diaspora in a concrete informal family setting, discussions were always lively. When I turned at the end to the dialect list there was always good-natured humor involved. "Do you say '*dhapna*' or '*dhakna*'?" (for 'lid of a pot') was one that people always laughed at. Here was "kitchen Hindi" coming out into the open, and being written down for the first time in their experience. Of course we say "*dhakna*"; only a fool would say "*dhapna*" (or vice versa, depending where you were).

Other tips? Well, I don't advise what I did once or twice in some areas: turn up uninvited for an interview at eight in the morning in a country district. In my defense, the sun rises early in KZN (4 a.m. in summer, and we have no daylight saving time). I'm amazed now to think that the woman of the house – let's call her Kunti – who had just shooed off the kids to school, patiently said, "Wait one minute," put her broom down, went into the kitchen, switched off the stove plates preparing lunch (we used to eat lunch early in KZN), and put on the kettle for a cup of tea with two sugars, returned to the sitting room in two minutes, and politely chatted to me about local life, her grandparents and her schooling in Hindi. She said "*dhakna*," by the way.

Last tip: to be a good fieldworker, whether of a migrant or *in situ* community, you have to be unafraid of dogs. Fortunately for my career as sociolinguist, I love animals (except, possibly, for the larger snakes with whom we cohabit in KZN). Just as well. An enduring memory of my Bhojpuri fieldwork was of one of Kunti's neighbors coming to the gate, which was some distance from the front door in this ample country district nestling close to the foothills of the Drakensberg. Amid the gently swirling midday mist, I can still picture her walking to the creaky gate, accompanied by a happy and uninhibited train of about ten dogs and six cats…

Vignette 5c
Fieldwork in Remnant Dialect Communities

Patricia Causey Nichols

"Meet the people's agenda, and they will meet yours." This advice came from J. Herman Blake, then provost of Oakes College at the University of California, Santa Cruz, whom I consulted in 1974 before beginning fieldwork on a river island in northeastern South Carolina inhabited exclusively by African Americans since the eighteenth century. Blake himself had been taking undergraduate students to another coastal island further south in the 1960s to study Gullah culture as they performed community service of various kinds (Jackson, Slaughter, & Blake, 1974).

In his recent analysis of the sociolinguistic construct *remnant dialect community*, Wolfram maintains that such a community "retains vestiges of earlier language varieties that have receded among speakers in the more widespread population" (2004, p. 84). Many African American communities in coastal South Carolina and Georgia, as well as a few in North Carolina and Florida, retain vestiges of African languages spoken by their ancestors, particularly in some grammatical constructions of the creole language often used among themselves. Gullah, a creole language with source languages from West Africa, Angola/Congo, and England, was spoken in coastal South Carolina by enslaved Africans, who outnumbered Europeans on large plantations but had no single African language in common. Sometimes known as Geechee or Sea Island Creole, Gullah is now an "insider" language for its speakers, and many outsiders believe it has disappeared entirely (Nichols, 2009). The very young and very old are among those who use it most.

Although my original, quantitative research design focused on morphosyntactic features of Gullah, my struggles to identify and meet the people's agenda over the next five months brought a wealth of qualitative data as well. I obtained these data by following the next important piece of advice I received, which came from Shirley Brice Heath, who was then completing research for her landmark study of language use in black and white communities of the Carolina piedmont (Heath, 1983). She suggested volunteering as an aide in the school attended by children of the community that was the focus of my research. Having taught sixth grade in Virginia many years previously, I was both comfortable with that activity and competent in the eyes of the local school officials. Heath later helped me design activities for children who had trouble with reading, recommending the children's books of Ezra Keats as some of the few then available with African

American characters. Peg Walker, the local teacher who accepted me into her social studies and science classes, was taking a class in reading instruction at a local college from Dr. Sally Hare, learning about the Language Experience Approach for children with persistent difficulties. This approach uses students' own language and prior experiences in the reading activities and made a contribution to my own research agenda. Walker asked me to work with certain students in small groups, having them tell stories to each other as I recorded them, typed them up, and returned with typed versions for them to read back to the group. Since these were their own stories, they had no difficulty with any of the vocabulary. Here I discovered that children who used distinctive Gullah features in storytelling for their peers would always substitute more standard forms when they read back their own stories. I began to observe many situations in which children as young as eight or nine were aware of the morphosyntactic differences between Gullah and classroom English, and often acted as "interpreters" when teachers from outside of the community could not understand something containing Gullah constructions.

As a white woman born and educated in a nearby community on the Waccamaw River, I needed both patience and time to establish credentials through work that the African American community valued. In addition to the nearly 100 children I saw three days each week in the social studies and science classes, I sometimes worked with children in the lower grades as well. The children took home news of their storytelling with me, our work with the set of magnetic letters, even some Saturday fishing, walks along the beach, and visits to my tiny apartment fronting the ocean, all of which helped establish my credibility and presence in the community. Within a month, the Sandy Island children invited me to visit their island on the boat they took home each day from school. After three visits, the island leader permitted me to tape-record him, and after this initial session I mailed him my proposal for teaching a weekly pre-college writing class for island young people.

Through my conversations with elders and children, I had come to understand the value the island community placed on higher education. Since I had taught basic writing classes in California community colleges, I began to explore the idea of a pre-college writing class as an activity that might "meet the people's agenda" while meeting my own as well. The island leader and his wife told me I could discuss my proposed writing class at preaching service on a Sunday before Thanksgiving, which was well into my third month of fieldwork. I also talked with the young man who drove the daily school boat, who told me that girls might like such a class – an interesting perspective on who would attend. On the assigned Sunday, the preacher asked me to say a few words about the class at the end of the service, and several girls stayed behind to sign up. At the first class meeting, some 20 students showed up, three of them male. At the second meeting, I distributed copies of short pieces each of them had written, and we all got down to the serious business of writing for each other – much as the elementary children had done with reading each other's stories. Once I gave a brief lecture about the sound patterns of ancestral West African languages that influenced their own speech patterns and often their spelling. At another point, I asked aloud why no males from Sandy Island attended college. One of the young adult men wrote a piece about that, pointing me to the occupation-related factors underlying the gender differences in language use that I had been discovering in my recordings of both island and mainland adults (Nichols, 1983). I drew on his observation for my final report to the Georgetown County School Board. Although the board had not requested such a report, I submitted one in hopes that community members might better understand and address language challenges facing teachers in the newly integrated schools, especially those unfamiliar with local language and culture.

The final months of my fieldwork were intense, with the ongoing work in the elementary school, the weekly writing class on the island, and a heavy schedule of recording. On class nights, the island leader and his wife invited me to stay in their home, and on Sundays I alternated attending church services on the island and in the church I had attended as a child. With the help of county teachers, I also conducted a postcard survey of every elementary school in the district to get reported data on their repeated observations that a cluster of boys in every fifth grade could not read. With confirmation of this pattern, I argued in my report to the school board that students' lack of male models for whom reading was an important daily activity might be addressed by hiring male teachers at the lower levels (Nichols, 1977).

At the completion of my formal fieldwork on Gullah, Charles Joyner, a noted Southern historian, took me to visit Dell and Virginia Hymes at their home in Philadelphia. Then studying for a second degree in Folklore and Folklife, Joyner was also focusing on the Waccamaw Neck for his research as both historian and folklorist (Joyner, 1984). This conversation with Dell and Virginia Hymes, before I began my analysis of what I had learned, helped me think through the kinds of data I had gathered and sort through what might be presented quantitatively or qualitatively. Dell Hymes pointed out that the difference in time between October and January for the number of creole features in my recorded interviews could be significant because my status as "participant" had changed. He maintained that the fact that I had no documentation of increased creole features in natural conversation should not prevent me from reporting that fact. When he asked if I ever did not understand what was being said, I remembered that only toward the end of my fieldwork had I begun to hear the use of *duh* for repeated actions and *ee/um* for personal pronouns. He also pointed out that an early caution by an older to a younger child not to speak that "country talk" around me might reflect the desire of the older sibling to use language I would value but that it should not be taken as evidence that Gullah is devalued in all circumstances. Hymes recommended return visits to the community, if just for ethical reasons, and such visits did indeed help me understand more about language use over the life span of individuals who worked for a time "up North" and then returned to the island in retirement. Despite the lack of large amounts of data on the morphosyntactic features recorded and observed in my fieldwork, he pointed out that sometimes just raw numbers, using arithmetic, were enough to show what *could* be shown. Because of her own teaching at the middle school level in Philadelphia, Virginia Hymes had encountered children who frequently traveled back to homes of relatives in the South for extended periods, maintaining distinctive speech patterns on their return. She suggested looking at the strong verbs for evidence of absent tense marking in children's speech, which I later found useful in comparing verb usage among white and black adult speakers in the same geographic area (Nichols, 1991). I left with a new appreciation of how important social context is in describing discrete linguistic facts that both she and I had observed in what seemed like distant classrooms, eventually expanding my research to include European and indigenous communities as part of the sociolinguistic context when Gullah came into being (Nichols, 2009).

Had I not immersed myself in community life in as many ways as possible in my role as someone who worked with children in their classrooms, I would never have experienced the wide range of language use available to their families over the course of their daily experiences. My richest linguistic data came with my endeavor to "give back" in ways valued by the community that I came to know: helping their young children with reading and their young adults with a weekly writing class. My own contribution to helping the community value its own speech patterns was not the direct type of linguistic gratuity, as described by Wolfram (1993), in which the linguist helps the community appreciate its own speech patterns. It was, rather, one paid forward in my subsequent classes for prospective teachers on the other side of the continent, as I directed classroom activities that had California students from many cultures explore their own ancestors' linguistic heritage, analyze recorded conversations among their friends, and figure out the rules that govern dialects and genres beyond academic English. Blake's advice sounded simple. Executing it was not.

References

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. Cambridge: Cambridge University Press.

Jackson, J., Slaughter, S. C., & Blake, J. H. (1974). The Sea Islands as a cultural resource. *Black Scholar, 5*, 32–39.

Joyner, C. (1984). *Down by the riverside: A South Carolina slave community*. Champaign: University of Illinois Press.

Nichols, P. C. (1977). A sociolinguistic perspective on reading and black children. *Language Arts, 54*, 150–157.

Nichols, P. C. (1983). Linguistic options and choices for black women in the rural South. In B. Thorne, C. Kramarae, & N. Henley (Eds.), *Language, gender, and society* (pp. 54–68). Rowley, MA: Newbury House.

Nichols, P. C. (1991). Verbal patterns of black and white speakers of coastal South Carolina. In W. F. Edwards and D. Winford (Eds.), *Verb phrase patterns in Black English and Creole* (pp. 114–128). Detroit, MI: Wayne State University Press.

Nichols, P. C. (2009). *Voices of our ancestors: Language contact in early South Carolina*. Columbia, SC: University of South Carolina Press.

Wolfram, W. (2004). The sociolinguistic construction of remnant dialects. In C. Fought (Ed.), *Sociolinguistic variation: Critical reflections* (pp. 84–106). New York: Oxford University Press.

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

6 The Sociolinguistic Interview

Kara Becker

Sociolinguistic fieldworkers often apply a broad stroke when referring to their method of data collection as a sociolinguistic interview, allowing the term to stand for any face-to-face interaction that is recorded for use as sociolinguistic data. This chapter distinguishes between this broad use of "sociolinguistic interview" and what I refer to as "The Sociolinguistic Interview," defined more narrowly here as a methodology developed within the Labovian variationist paradigm with the goal of systematically eliciting variation across contextual styles for use as the primary evidence for sociolinguistic stratification and linguistic change. A strict definition allows for an emphasis on the specific utility of data gathered from The Sociolinguistic Interview in relation to other recordings of naturalistic speech and is meant to stress the importance of making informed methodological choices when gathering sociolinguistic interview data.

A Strict Definition

Despite the continuous expansion of sociolinguistic data collection techniques, The Sociolinguistic Interview as originally developed for Labov's (1966) study of New York City's Lower East Side remains ideologically central to the field. Broadly adopted is Labov's early statement about good data: "No matter what methods may be used to obtain samples of speech (group sessions, anonymous observation), the only way to obtain sufficient good data on the speech of any person is through an individual, tape-recorded interview" (1972, p. 209).

Yet there is far more to The Sociolinguistic Interview than the fact that it is an individual interview, so that referring to various kinds of face-to-face recordings using the term diminishes the theoretical import and specific goals of the original methodology. Further, fieldworkers often do not follow the strict methodology laid out in works like Labov (1984) (see the vignettes following this chapter). The consequence of this practice is more than practical, as The Sociolinguistic Interview methodology exists to serve certain foundational sociolinguistic principles advanced in the Labovian paradigm. It is the goal of this chapter to make clear what The Sociolinguistic Interview is and what kind of sociolinguistic questions it can answer.

What Is It?

The Sociolinguistic Interview is a controlled speech event designed to elicit a wide range of contextual styles from an individual speaker. During analysis, a linguistic variable (or variables) is quantified across these contextual styles to arrive at a range of that speaker's production.

The elicitation of contextual styles is the methodological goal of The Sociolinguistic Interview, and so these styles govern its structure. Often the interview's structure is conceptualized as moving temporally across five contextual styles: casual (A), careful (B), reading (C), word list (D), and minimal pair (D′) (Labov, 1972). A better way to describe its structure is to say that it has two parts that differ with respect to the topic of language. During the bulk of the interview, participants are not cued to the fieldworker's interest in language. The fieldworker gathers demographic information and uses topic modules (Labov, 1984) to elicit conversational speech, which is later divided into casual and careful styles. After a sizable amount of this speech has been elicited, topics and tasks are introduced where language is either explicitly or implicitly the focus. At a minimum, these are styles C–D′: reading tasks that target variables of interest, a word list with the target variable embedded, and a list of minimal pairs (words that differ only in the target variable). Additionally, many fieldworkers elicit meta-linguistic commentary, either conversationally or through other tasks such as surveys or subjective reaction tests. The most complete documentation of the methodology of The Sociolinguistic Interview is provided in Labov (1984); it is also outlined at length in Labov (1966).

What Kinds of Sociolinguistic Questions Can It Answer?

The Sociolinguistic Interview serves as the primary data in the investigation of sociolinguistic variation and change because the interview, and the individual speaker represented by that interview, never stands alone. Instead, it forms part of a set of comparable interviews gathered from a sampling of some speech community. The reason that an individual's production across contextual styles is the principal source of data in the Labovian paradigm has been defined well by Bell (1984) as the *Style Axiom*: "Variation in the style dimension within the speech of a single speaker derives from and echoes the variation which exists between speakers on the 'social' dimension" (p. 151). In short, an individual's stratified variation across contextual styles, in the Labovian paradigm, is considered to be a direct reflection of, or a point of access into, the socially stratified variation of a speech community; further, it is the community pattern that is ultimately of interest, not the speech patterns of the individual (Labov, 1972, p. 112).

In addition to demonstrating sociolinguistic patterning or the orderly heterogeneity (Weinreich, Labov, & Herzog, 1968) that was foundational to the sociolinguistic enterprise, intra- and inter-speaker variation work together to serve as the primary diagnostic tools for the identification of linguistic change in progress. The presence of intra-speaker variation can signal some degree of social awareness, a crucial component of change in progress, particularly change from above (Labov, 2001, p. 86). Moreover, when intra-speaker variation across contextual styles is regularly patterned for a socially stratified corpus of speakers, certain patterns elucidate change. A good example is the crossover pattern (Labov, 1966; 2001), wherein the second-highest-status group (usually the lower middle class) surpasses the highest-status group in production of a prestige variable in contextual styles D and D′, which provides evidence of a change from above.

Labov's (1966) study demonstrates the kind of findings that result from investigating data drawn from The Sociolinguistic Interview methodology for a speech community. The variable, non-rhoticity in the syllable coda shows the combination of social and stylistic stratification that characterizes variables undergoing change in progress. Speaker groups are finely stratified according to the social characteristic of socioeconomic status; in addition, the individuals who make up those groups pattern together in shifting along the continuum of contextual styles from casual to formal. The lower middle class shows hypercorrection in more formal styles D and D′, further confirmation for the change from above.

In short, The Sociolinguistic Interview methodology is constructed to provide evidence of sociolinguistic variation – data to distinguish "a casual salesman from a careful pipefitter" (Labov, 1972, p. 240) – and also of change in progress, the foundational principles of the Labovian variationist paradigm.

What Does It Assume?

To employ The Sociolinguistic Interview methodology is to adopt a set of assumptions. Many are part of the methodological axioms presented in Labov (1984, pp. 29–30); I focus on two here.

1. *Labovian contextual styles*. The first assumption is a set of related principles that underlie the Labovian contextual styles. One principle is that speakers, and the variables they use, shift along a continuum of formality (and if we further equate formality with standardness, this is a continuum with stigma and prestige at opposite poles). Another is the notion that this formality continuum can be accessed by regulating the amount of attention a speaker pays to their own speech – that is, attention to speech is the "cognitive mechanism" (Eckert & Rickford, 2001, pp. 2–3) from which emerges a speaker's navigation of the stigma–prestige continuum. Lastly, there is the notion that we can observe shifts along this continuum through the Labovian contextual styles.

Critiques of assumptions surrounding the Labovian contextual styles are both practical and theoretical. Practically, it has proven difficult to systematically distinguish contextual style in the body of the interview. For example, channel cues, or "changes in volume, pitch, tempo, breathing and laughter," were once thought to signal a shift from casual to careful speech (Labov, 2001, p. 89). An early rejection of this idea (see Wolfram, 1969) gave way to a more general concern with subjectivity in isolating styles, so much so that, according to Rickford and McNair-Knox (1994, p. 239), "most quantitative sociolinguists came to ignore the casual/careful distinction" (see also Singler, 2007, pp. 126–127). The introduction of the Decision Tree (Labov, 2001), while providing a framework for the division of interview speech into casual and careful styles, has not been widely adopted and is admittedly subjective (p. 91). Another practical issue concerns the more formal styles, all of which rely on reading tasks. Baugh (2001, p. 110) argues that the assumption that participants are literate, and comfortably so, represents a Western bias that restricts ethnographically informed community projects, while Milroy (1987, p. 173) notes that low literacy rates in her Belfast sample ruled out the analysis of reading style.

Other critiques center on the theoretical basis of both the attention to speech mechanism and the formality continuum, and the link between them (Eckert, 2001; Milroy, 1987; Milroy & Gordon, 2003; see also Rickford & McNair-Knox, 1994, p. 239). Milroy (1987), for instance, questions whether conversational and reading styles are comparable types of behavior. Schilling-Estes (1998) points out that many performative styles, in which often a great deal of attention is paid to speech, are highly vernacular, highlighting the potential for speakers to have a range of naturalistic styles that are not controlled by the attention to speech mechanism. In light of increasing interest in style in sociolinguistics, the style as Attention to Speech model has many competitors. It is important to note, however, that Labovian contextual styles concern intra-speaker variation and should not be confused with more holistic definitions of style (Eckert, 2001). Labov (2001, p. 87) notes that the Attention to Speech model was never intended to stand as the singular mechanism governing intra-speaker variation, nor should intra-speaker variation be seen as the singular expression of speaker style.

Accepting the Attention to Speech model does not preclude other conceptualizations of style, as they are not a priori mutually exclusive. The same speaker may shift certain variables along a continuum of formality while simultaneously attending to other stylistic practices, such as accommodating to audience members (Bell, 1984) or constructing some aspect of identity (Eckert, 2001), even within the strict methodological confines of The Sociolinguistic Interview. In Becker (2009), rates of non-rhoticity for a group of Lower East Siders in New York City shifted significantly across three contextual styles: interview style, reading style, and word list style. At the same time, speakers significantly shifted rates of coda /r/ production in the body of the interview according to whether local or non-local topics were being discussed, and not according to the casual–careful distinction, a shift that was tied to the agentic construction of a place identity by those speakers.

2. *The vernacular*. Not all Labovian contextual styles are created equal. It is the least formal style, where the least attention is paid to speech, that is most valuable; this is the vernacular. The definition of the vernacular utilized here is also a narrow one, moving beyond uses that place "local" or "not standard" in opposition to terms like *standard* or *formal*. In this narrow definition, rather, the vernacular is defined in terms of how it behaves; both the speakers producing the vernacular and the system itself are considered the most natural and regular of any style. Speakers produce the vernacular when behaving most naturally and comfortably, so that it is defined, according to Labov (1972), as "the everyday speech which the informant will use as soon as the door is closed behind us: the style in which he argues with his wife, scolds his children, or passes the time of

day with his friends" (p. 85). Second, we define the vernacular in terms of its own linguistic behavior: it is, or is claimed to be, the most systematic speech, ostensibly the variety first acquired in childhood. This systematicity drives our interest in privileging the vernacular in sociolinguistic investigation – what is known as the *Vernacular Principle* (Labov, 1972, p. 112).

Critiques of the vernacular as defined here do not generally reject our interest in it as an appropriate object of study (see Milroy, 1987, p. 60). Instead, concerns have been raised over the operationalization of the vernacular, as well as the value placed on it. The idea that a speaker has a way of talking that is most natural to her or him is generally not a part of the debate and is perhaps axiomatic. The problem here is that this natural variety is equated with the speech produced in the casual style of The Sociolinguistic Interview and, further, that casual speech has in almost all cases been equated with stretches of talk with the highest rates of non-standard speech. Vernacular speech is not by definition non-standard (Labov, 1984, p. 29, gives Received Pronunciation as an example of a vernacular that is a prestige variety), yet virtually all analyses of data from The Sociolinguistic Interview that seek the vernacular find it in casual speech and in non-standard forms. This three-way equating (vernacular = casual speech = non-standard usage) is highly problematic as a theoretical assumption. Schilling-Estes (2004, p. 188) presents data from situations in which low levels of non-standard variables are in fact linked to a casual style and, like Wolfson (1976) and others, urges the acknowledgment of multiple kinds of naturalistic speech.

Another critique concerns the argument that the vernacular is the most systematic of a speaker's varieties. Again, it seems logical from the perspective of acquisition that the variety first acquired would be the most systematic. Yet, ruling out exceptional cases (second language acquisition, extremely formal self-conscious speech), it remains an empirical question whether or not some stretch of naturalistic speech is more systematic than another stretch. Crucially, no work to date has found this result for speech from within the body of The Sociolinguistic Interview. And if it were to be shown, does this imply, as Singler (2007, p. 127) asks, that other types of naturalistic speech (careful speech, for instance) are a-systematic? In fact, given the observation above that most sociolinguists do not distinguish between careful and casual speech, it must be the case that we believe careful speech is systematic enough to quantify and analyze.

A third critique is aimed at how the notion of the vernacular is conceptualized as an entity or bounded variety (Schilling-Estes, 2007, p. 173). Researchers talk of "isolating" the vernacular, as if it were a thing that can be captured. Indeed, debates arise over the validity of sociolinguistic data where that validity relies on a claim to have obtained the vernacular when others haven't (cf. the discussion in Rickford, 2006). Intrinsic to these debates is the equation of a bounded vernacular to high rates of non-standard production, so that the evidence for "capturing" the vernacular is a higher rate of use for a non-standard feature than found by other researchers. The higher rates, conceptualized as a captured, bounded vernacular, are then validated as real, authentic data in opposition to other data that are inauthentic. One danger of this kind of practice is the sense of some kind of a chase (for higher and higher rates which would indicate the capture of the “true” vernacular), yet one lacking an identifiable finish line. As Milroy and Gordon (2003) put it: “The difficulty in pursuing the vernacular … lies with the impossibility of recognizing the quarry when it is caught” (p. 50). And as Labov has noted, it is not the search for the vernacular that is problematic, it is the danger of claiming to have accessed it. He cautions: “We are forced to recognize the limitations of our other methods of eliciting the vernacular … [We] have defined a direction but not the destination” (1972, p. 90).

There are other dangers involved with the search for a bounded vernacular. If we expect the vernacular to be found in stretches of talk with the highest rates of non-standard features, we risk using those rates as a diagnostic of vernacular speech. Labov (1972, p. 95) notes this danger, which motivated the early proposal to use channel cues as a guide. Having now as a field rejected the use of channel cues, we may utilize the Decision Tree (Labov, 2001) to distinguish careful speech from casual (vernacular) speech, yet, as noted above, the field has not adopted this tool. Whether our analyses are in fact circular or not, I would argue that an ideological circularity pervades the field: the vernacular is whatever style has the highest rates of non-standard speech, and the highest rates of non-standard speech indicate we have “captured” the vernacular. Even in Labov (2001), where the Decision Tree is presented specifically to avoid the subjectivity and potential circularity of the casual–careful chunking, a category of residual speech (speech not categorized by a node on the tree) is posited to contain some of the vernacular *based on rates of usage*: “The fact that the Residual category is not the lowest for (DH) or (ING) makes it seem likely that there are many other Casual speech categories that can be extracted from it” (p. 107).

Another danger in seeking the vernacular is the concomitant valorization of it. Bucholtz (2003) argues that a linguist researcher assumes the role of “arbiter of authenticity” (p. 407), so that as analysts we privilege certain kinds of speakers and certain kinds of speech. Eckert (2003) argues that while the selection of “authentic speakers” is perhaps not openly discussed, the valorization of “authentic speech” – the vernacular – is proclaimed. That is, we like to tell others when we capture it: “Sociolinguists boast special methods for getting at language in its natural state. If the Authentic Speaker is an elephant hovering in the corner, the vernacular is a moose sprawling in the middle of the table” (p. 394). Milroy and Gordon (2003, p. 50) note what they call a striking similarity between the ideologized vernacular and its polar opposite, the standard. Yet they argue that, theoretically, the abstraction of the vernacular is quite beneficial as long as the dangers as laid out here are acknowledged, and I would agree. With respect to The Sociolinguistic Interview, so much of its methodology centers on a pursuit of the vernacular. This pursuit, while fruitful, should not go unproblematized.

The Sociolinguistic Interview in Practice

Despite its specific theoretical ends, The Sociolinguistic Interview shares much in practice with the broader set of face-to-face interviews conducted by sociolinguists. First and foremost is the focus on the elicitation of naturalistic speech and the concomitant challenge of overcoming the well-known “observer’s paradox,” defined in Labov (1972): “To obtain the data most important for linguistic theory, we have to observe how people speak when they are not being observed” (p. 113). The focus in the field on naturalistic speech serves to heighten our interest in the vernacular, so that the methodology proves to be a useful one for many sociolinguistic pursuits. In particular, a number of questions and topics (e.g., childhood games) have been noted as successful in eliciting naturalistic speech, like personal narratives.

Other techniques used as part of The Sociolinguistic Interview also inform a broader range of face-to-face data collection. One is the practice of cataloging demographic information at the beginning of the interview, as a way of creating a speaker profile for later analysis of co-variation of social and linguistic phenomena, as well as to set the tone of the speech event as a casual conversation about the interviewee’s life, experiences, and opinions. This early section further serves as transitional material from the start of the interview, where interviewees may be nervous and/or hyper-aware of recording equipment, into later topics where we hope to overcome the observer’s paradox and elicit more naturalistic speech. The training of interviewers is another area where best practices are not limited to The Sociolinguistic Interview. Honing questioning techniques (for instance, learning to move past yes/no questions to those that elicit longer narratives), controlling topic (for instance, identifying and pursuing topics of interest to the interviewee), and monitoring the presentation of self (for instance, using the tactic of playing the role of “naïve” interviewer to promote conversation and to disrupt if possible any power asymmetries between interviewer and interviewee) are all utilized in sociolinguistic fieldwork more broadly and are not limited to The Sociolinguistic Interview.

The success of The Sociolinguistic Interview as a tool for eliciting naturalistic speech has, not surprisingly, been critiqued, most notably by Wolfson (1976). Wolfson argues that The Sociolinguistic Interview is a speech event (p. 190) in which participant roles and expectations cannot in some cases and some cultures be so easily overridden by interview techniques (see also Briggs, 1986; Milroy, 1987; Milroy & Gordon, 2003; Schilling-Estes, 2007). As Eckert (2001) puts it, the issue at hand is “how well the ‘constructed stylistic world’ of the interview maps onto the larger, natural stylistic world” (p. 119). These are challenges that any recording claiming to be naturalistic should and will face; perhaps the supremacy of The Sociolinguistic Interview as used in the field has led to a well-deserved and increased scrutiny of best practices. For instance, Feagin (2002) cites several sociolinguistic studies where the classic Danger of Death question (in which interviewees are asked to recall a time in which they thought their life was threatened) was unsuccessful in encouraging interviewees to produce a long narrative using naturalistic speech. This happened in Feagin’s own research and was unsuccessful in Becker (2010) as well. An early adoption of the Danger of Death question in the field led to a later move toward reflexivity in identifying topics more broadly that are of interest to interviewees from a range of backgrounds, as well as acknowledgment of the importance of asking pertinent questions that will elicit emotional, engaged telling of narratives within the specific fieldwork context.

Ultimately, in practice The Sociolinguistic Interview shares much with other face-to-face naturalistic recordings, both in its best practices for eliciting naturalistic speech and in the challenges inherent in doing so. It is perhaps because of this similarity across interviewing techniques that so many techniques are referred to with the term "sociolinguistic interview." Most face-to-face recordings may appear to "look" like The Sociolinguistic Interview (with a major exception being the presence or absence of more formal contextual styles) and will utilize many of the same methodological tools. In practice, then, there is much overlap. It is in analysis – in the goals for the data gathered – that The Sociolinguistic Interview differs greatly from the broader set of face-to-face sociolinguistic recordings.

Conclusion

Numerous scholars have noted that The Sociolinguistic Interview is not the appropriate methodological choice for many sociolinguistic studies. In some cases, this is due to the community of interest (Baugh, 2001; Briggs, 1986). In others, it is due to the variable of interest: many syntactic variables, for instance, as well as variables that hold covert prestige, can be difficult to elicit in interview situations (Milroy & Gordon, 2003; Wolfram, 2011). Yet The Sociolinguistic Interview continues to hold its place as the central methodological tool in sociolinguistics despite my argument here that, in fact, the majority of contemporary sociolinguistic studies do not utilize The Sociolinguistic Interview methodology. In Becker (2010), for instance, a three-year ethnography led to the recording of over 100 interviews with community residents on the Lower East Side of New York City. These interviews were conducted in collaboration with community partners and now reside in the community as The Seward Park Oral History Project. In designing the interviews, many methods from The Sociolinguistic Interview, including the use of topics and topic modules, were utilized, but the corpus of interviews does not abide by the strict definition presented in this chapter. Crucially, the contextual styles C–D′, as well as metalinguistic commentary, were not elicited. The implication of this for sociolinguistic analysis is that investigations of change in apparent time from the perspective of the Labovian paradigm are necessarily limited due to a lack of the full range of contextual styles. Thankfully, Labov's own (1966) Lower East Side sample exists for comparison, so that the Becker (2010) data can be seen as a quasi-trend study. Yet without the elicitation of contextual styles, the interviews from Becker (2010), however valuable they may be as sociolinguistic data, do not satisfy the criteria of The Sociolinguistic Interview.

I am not suggesting that the problem is simply in the use of terminology. The many studies that refer to data from sociolinguistic interviews – when in fact those interviews do not employ the strict methodology employed here – do not necessarily need to be re-termed. Rather, we need more clarity in the field with regard to our interview techniques. To continue centralizing this method in the field, either through habit or through an expanded use of the term, obscures the underlying assumptions and the goals of the methodology. It further diminishes the contributions made by those sociolinguists who adopt these assumptions and who have used the methodology to revolutionize our understanding of linguistic variation and change.

In short, sociolinguists need to be clear about what kind of data they want and what kind of questions they want to answer before adopting a methodology. In the case of The Sociolinguistic Interview, there may be ideological pressure to adopt this method when more general face-to-face interactions would suffice. As guides to sociolinguistic methods advise, designing a research study requires an answer to the question "What do I want to study?" If the answer falls within the Labovian variationist paradigm, then an appropriate and central methodology is The Sociolinguistic Interview.

References

Baugh, J. (2001). A dissection of style-shifting. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 109–118). Cambridge: Cambridge University Press.

Becker, K. (2009). /r/ and the construction of place identity on New York City's Lower East Side. *Journal of Sociolinguistics, 13*(5), 634–658.

Becker, K. (2010). Social conflict and social practice on the Lower East Side of Manhattan. (Unpublished doctoral dissertation). New York University.

Bell, A. (1984). Language styles as audience design. *Language in Society, 13*, 145–204.

Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge: Cambridge University Press.

Bucholtz, M. (2003). Sociolinguistic nostalgia and authentication of identity. *Journal of Sociolinguistics, 7*(3), 398–416.

Eckert, P. (2001). Style and social meaning. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 119–126). Cambridge: Cambridge University Press.

Eckert, P. (2003). Sociolinguistics and authenticity: An elephant in the room. *Journal of Sociolinguistics, 7*(3), 392–397.

Eckert, P., & Rickford, J. R. (2001). Introduction. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 1–18). Cambridge: Cambridge University Press.

Feagin, C. (2002). Entering the community: Fieldwork. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 20–39). Malden, MA: Blackwell.

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

Labov, W. (1972). *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press.

Labov, W. (1984). Field methods of the project on language change and variation. In J. Baugh & J. Sherzer (Eds.), *Language in use: Readings in sociolinguistics* (pp. 28–53). Englewood Cliffs, NJ: Prentice-Hall.

Labov, W. (2001). The anatomy of style-shifting. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 85–108). Cambridge: Cambridge University Press.

Milroy, L. (1987). *Observing and analysing natural language: A critical account of sociolinguistic method*. Oxford: Blackwell.

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

Rickford, J. R. (2006). Down for the count? The creole origins hypothesis of AAVE at the hands of the Ottawa Circle, and its supporters. *Journal of Pidgin and Creole Languages, 21*(1), 97–155.

Rickford, J. R., & McNair-Knox, F. (1994). Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 235–276). New York: Oxford University Press.

Schilling-Estes, N. (1998). Investigating "self-conscious" speech: The performance register in Ocracoke English. *Language in Society, 27*, 53–83.

Schilling-Estes, N. (2004). Constructing ethnicity in interaction. *Journal of Sociolinguistics, 8*(2), 163–195.

Schilling-Estes, N. (2007). Sociolinguistic fieldwork. In R. Bayley & C. Lucas (Eds.), *Sociolinguistic variation: Theories, methods, and applications* (pp. 165–189). Cambridge: Cambridge University Press.

Singler, J. V. (2007). Samaná and Sinoe, part 1: Stalking the vernacular. *Journal of Pidgin and Creole Languages, 22*(1), 123–148.

Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics: A symposium* (pp. 97–195). Austin: University of Texas Press.

Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Washington, DC: Center for Applied Linguistics.

Wolfram, W. (2011). Fieldwork methods in language variation. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The Sage handbook of sociolinguistics* (pp. 298–311). London: Sage.

Wolfson, N. (1976). Speech events and natural speech: Some implications for sociolinguistic methodology. *Language in Society, 5*, 189–209.

Vignette 6a
Cross-cultural Issues in Studying Endangered Indigenous Languages

D. Victoria Rau

My research on Yami, a Philippine Batanic language spoken by 4,000 speakers on Orchid Island, which lies off the southeastern coast of Taiwan, began in 1994 with a personal invitation from Maa-neu Dong, who was seeking an Austronesian linguist to compile a dictionary of her mother tongue. In response, I conducted a sociolinguistic survey of this indigenous language, following an SIL method (Blair, 1990; Grimes, 1995) to gather basic word lists, texts for intelligibility tests, information on bilingual ability, language use, and language attitudes.

In the process of analyzing the word lists, several linguistic variables emerged as potential candidates for Labovian-style sociolinguistic studies, including one similar to the centralization of diphthongs on Martha's Vineyard (Labov, 1972). The Yami diphthongs (ay) and (aw) were undergoing vowel raising on the island (e.g., *mangay* ~ *mangey* 'go', *araw* ~ *arow* 'day, sun'), with an isogloss separating the more progressive northeast from the more conservative southwest. However, even though most coding of the variants in the word list was completed within three years of the initial trip, I lacked adequate understanding of both the linguistic structure of Yami and the social structure of the speech community to be able to analyze the social stratification of the change. As it was an endangered minority language, no comprehensive grammatical sketch of Yami was available at the time. Not until our *Yami Texts with Reference Grammar and Dictionary* was published (Rau & Dong, 2006) did we feel we were ready to write up the centralization of diphthongs (Rau, Chang, & Dong, 2009). Furthermore, during 2005–2009 we were able to put our Yami materials online for language conservation, including documentation (http://yamiproject.cs.pu.edu.tw/yami), e-Learning (http://yamiproject.cs.pu.edu.tw/elearn), and an online dictionary (http://yamibow.cs.pu.edu.tw).

When a "cross-cultural" investigation involves a less commonly studied, endangered indigenous minority language, practical goals of language conservation should take precedence over theoretical sociolinguistic goals. Researchers may also need to accept the frustrating reality that it is never possible to interpret the limited data as quickly, accurately, and adequately as when studying dominant languages. Below, I describe some of the sociolinguistic issues raised during my work with the Yami speech community.

Urban Dialectology vs. Endangered Indigenous Language Studies

The goals of and data collection methods for a variationist sociolinguistic study of a minority language differ from those of urban dialectology. A typical Labovian-style study seeks to address the question of social motivation of linguistic change. A valid and reliable variationist study usually has several prerequisites. First, the researcher needs to have native or near-native command of the target language, whether the researcher personally conducts sociolinguistic interviews, hires a local interviewer to match the local speech style (Trudgill, 2010), or uses other supplemental techniques (Wolfram, 2011). Second, there are usually grammatical sketches, dialect studies, or records of historical linguistics in the target language to serve as a basis for comparison. Third, the linguistic variable to be investigated has to provide sufficient stratified data in the subsystem to meet Labov's (1972) principle of accountability and provide Tagliamonte's (2009) three lines of evidence for VARBRUL analysis. This makes a dominant language a perfect candidate for Labovian-style variation studies.

Endangered indigenous languages are a different story (Rau, 2011). It takes much longer to develop a basic understanding of the language before studies of variation can even be attempted. Data collection is usually restricted to word lists and narratives, as the researcher's proficiency in the language is limited. In addition, the range of linguistic variables is also compressed, as the consultants who assist in data transcription may edit out some "variations," both to make the language look more "standard" and because the transcriber naturally transcribes in their own dialect.

Methodological Differences

How did my methods of sociolinguistic data collection on Orchid Island differ from the Labovian method? My initial data were gathered in 1994 as part of a sociolinguistic survey to establish a relationship with the community. The word lists were transcribed phonetically, but the narrative data were transcribed phonemically by my Yami consultant, who came from the non-centralized /ay/ and /aw/ dialect area. To gather more narrative data, my consultant and I went back to the community when we were commissioned to do a Yami dictionary project in 1998–2000 and language documentation project in 2005–2007. We managed to glean enough data from the same speakers who had contributed word list and narrative data in 1994 for a variation study.

Unfortunately, as the language is not being transmitted to the younger generation (Lin, 2007), we cannot test our hypothesis of change in progress. Nor did we ever have a chance to conduct a trend study or panel study (Sankoff & Blondeau, 2007), since it took us over a decade to process and understand the data gathered in 1994. To study the two diphthongs, my Yami consultant had to go back to the previously transcribed texts to recode the pronunciation of /ay/ and /aw/. Lacking sociolinguistic interviews, we treated narratives as "informal style" in contrast with the "formal" word list reading style, as defined by the degree of attention paid to form. In our later study of word order variation, we found that narratives and conversations could be further distinguished by word order variation (Chang & Rau, 2011), but this insight came too late for our 2009 study.

Advice

On the basis of my experience with the Yami community, I recommend using a four-step approach to data collection for the purpose of producing useful sociolinguistic materials, following the principle of linguistic gratuity (Wolfram, Reaser, & Vaughn, 2008):

1. Conduct a sociolinguistic survey, gathering word lists and narratives.
2. Write a reference grammar and teaching materials as part of a language conservation effort to "give back" to the community under study.
3. Identify potential linguistic variables to contribute to both practical and theoretical issues. For example, our variation study of the two diphthongs (Rau et al., 2009) has provided the theoretical underpinning for orthography development in Yami. A study of Yami word order (Chang & Rau, 2011) has led us to understand how narrative and conversation styles can account for word-order variation between VS and SV. A recent study of the variation between path verbs and manner verbs in Yami (Rau, Wang, & Chang, 2012) has increased our understanding of motion events in cognitive linguistics.
4. Prepare to write a user-friendly socio-grammar (Nagy, 2009). A useful pedagogical grammar with a focus on language use suitable for indigenous language teacher training programs would be highly appreciated, as a linguist's grammar is usually perceived as incomprehensible or irrelevant to indigenous teachers who are teaching their language in a school setting for language conservation.

As studies on minority speech communities require a lifelong commitment, I hope this "been there, done that" account will give researchers a firm and practical foundation when collecting cross-cultural data.

References

Blair, F. (1990). *Survey on a shoestring: A manual for small-scale language surveys*. Dallas, TX: Summer Institute of Linguistics.

Chang, H.-H., & Rau, D. V. (2011). Word order variation in Yami. Paper presented at the New Ways of Analyzing Variation Asia-Pacific 1 conference. Delhi, India.

Grimes, J. E. (1995). *Language survey reference guide*. Dallas, TX: Summer Institute of Linguistics.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lin, Y.-H. (2007). A sociolinguistic study on Yami language vitality and maintenance. (Unpublished master's thesis). Providence University, Taichung, Taiwan.

Nagy, N. (2009). The challenges of less commonly studied languages: Writing a sociogrammar of Faetar. In J. N. Stanford & D. R. Preston (Eds.), *Variation in indigenous minority languages* (pp. 397–417). Amsterdam: John Benjamins.

Rau, D. V. (2011). Research designs for using VARBRUL to study interlanguage variation, grammaticalization, and word order variation. Workshop presented at the New Ways of Analyzing Variation Asia-Pacific 1 conference. Delhi, India.

Rau, D. V., Chang, H.-H., & Dong, M.-N. (2009). A tale of two diphthongs in an indigenous minority language. In J. N. Stanford & D. R. Preston (Eds.), *Variation in indigenous minority languages* (pp. 259–280). Amsterdam: John Benjamins.

Rau, D. V., & Dong, M.-N. (2006). *Yami texts with reference grammar and dictionary. Language and linguistics*. Monograph A-10. Taipei: Institute of Linguistics, Academia Sinica.

Rau, D. V., Wang, C.-C., & Chang, H.-H. A. (2012). Investigating motion events in Austronesian languages. *Oceanic Linguistics, 51*(1), 1–17.

Sankoff, G., & Blondeau, H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language, 83*, 560–588.

Tagliamonte, S. (2009). *Be like*: the new quotative in English. In N. Coupland & A. Jaworski (Eds.), *The new sociolinguistics reader* (pp. 75–91). New York: Palgrave Macmillan.

Trudgill, P. (2010). Norwich revisited: Recent linguistic changes in an English urban dialect. In M. Meyerhoff & E. Schleef (Eds.), *The Routledge sociolinguistics reader* (pp. 358–368). New York: Routledge.

Wolfram, W. (2011). Fieldwork methods in language variation. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The Sage handbook of sociolinguistics* (pp. 296–311). London: Sage.

Wolfram, W., Reaser, J., & Vaughn, C. (2008). Operationalizing linguistic gratuity: From principle to practice. *Language and Linguistics Compass, 2*(6), 1109–1134.

Vignette 6b
Conducting Sociolinguistic Interviews in Deaf Communities

Ceil Lucas

There are several issues to consider when conducting sociolinguistic interviews for sign language projects, including the selection of subjects, the use of contact people, the history of deaf education and how research has taken place in Deaf communities, and the anonymity (or lack thereof) of research subjects. A major component of data collection, of course, is the selection of the subjects. Sociolinguistic studies want to be able to determine the correlation between variation and speaker (in this case, signer) characteristics, including age, gender, ethnicity, region, and socioeconomic status. Although characteristics such as gender, age, and ethnicity are common in studies of linguistic variation, they often need to be articulated more fully when they are put into research practice in a given community. This is particularly true for studies of linguistic variation in Deaf communities. Notions of socioeconomic status or even age cannot be simply borrowed wholesale from studies of variation in spoken language communities. The differences in social characteristics when applied to Deaf communities are of two types. The first type includes characteristics, such as age and region, that may have a different meaning when the history of Deaf communities is taken into account. The second type includes characteristics such as language background that are unique to Deaf communities.

For deaf people, regional background or where they were born may be less important than where they attended school (especially if it was a residential school) or where their language models acquired American Sign Language (ASL). Age as a characteristic may have different effects on linguistic variation because of the differences in language policies in schools and programs for deaf children since 1817. Some differences in language use may result from changes in educational policies, like the shift from oralism to Total Communication or from Total Communication to a bilingual–bicultural approach. (Oralism is the approach which requires that only the spoken language be used to instruct deaf students, to the total exclusion of sign language, based on the assumption that, above all else, deaf students must learn to speak. Total Communication advocates the simultaneous use of speaking and signing, the latter strongly reflecting the structure of the spoken language, with the perspective that in this way, students can "see the spoken language on the hands." A bilingual–bicultural approach recognizes sign language as a full-fledged linguistic system structurally independent from the spoken language with which it coexists, and also the cultural context surrounding a sign language.) These language policies have affected not only what language is used in the classroom but also teacher hiring practices that have supported hiring deaf teachers who know the sign language in question or hearing teachers who cannot sign. These language policies have affected deaf children's access to appropriate language models, and this access may have varied across time to such an extent that it has affected the kind of variation that we see in sign languages today.

One strong example concerns the Black Deaf community in the United States. Following the Civil War, 17 states and the District of Columbia established separate schools for Black deaf children or opened "departments" – that is, separate buildings – on the campus of the school for White children. Even though deaf education started propitiously in 1817 at the American School for the Deaf in Hartford, Connecticut, with ASL as the medium of instruction, by 1880 oralism was firmly established in the schools for White deaf children, with many deaf teachers being fired. However, the policy of oralism was not extended evenly to the schools for Black deaf children, and the use of sign language as the medium of instruction was widely allowed. In addition, some schools for Black deaf children had White deaf ASL-signing teachers providing the children with ASL input. Then, following the *Brown vs. Board of Education* decision in 1954, Black and White deaf children slowly began to attend school together (even though some states, such as Louisiana, managed to delay integration until 1978!), and the practice of mainstreaming began to take over education at residential schools, so deaf children had increasingly more contact with their hearing peers. All of this context helps explain the variation in Black ASL that McCaskill, Lucas, Bayley, and Hill (2011) have found, such as noticeably less mouthing in older signers, since they had less direct exposure to oralism and hearing peers who spoke English.

The selection of subjects for sociolinguistic studies of sign languages must take into account the meaning of age, ethnicity, and region in Deaf communities, in order for the resulting analyses to be meaningful (see Hill, Vignette 6c). Furthermore, large studies of sociolinguistic variation in ASL (Lucas, Bayley, & Valli, 2001) and other sign languages such as Auslan (Schembri et al., 2009) have clearly shown the importance of whether a subject comes from a Deaf family in which the sign language is used or from a non-signing family, be it hearing or deaf. For example, Lucas et al. (2001) demonstrated that subjects from Deaf families were more likely to use the standard "citation" forms of signs, such as signs like KNOW produced at the forehead as opposed to lower locations.

Central to the selection of subjects are contact people. The approach to selecting participants in Lucas et al. (2001) and McCaskill et al. (2011), for example, was guided by the work of Labov (1972a; 1972b; 1982) and Milroy (1987). Groups were assembled in each area by a contact person, a Deaf individual living in the area with a good knowledge of the community. These contact people were similar to the "brokers" described by Milroy, individuals who "have contacts with large numbers of individuals" in the community (1987, p. 70). The contact people were responsible for identifying persons suitable for the study – in the case of the 2001 and 2011 studies, fluent lifelong users of ASL who had lived in the community for at least 10 years. Community members may be decidedly reluctant to participate in a study and may outright refuse. This is not at all unique to Deaf communities. As Wolfram (2013) explains,

> community members may have underlying questions and concerns about sociolinguists' motivations in working in their community. What are they really doing in their community? Why are they so obsessed with the minutia of language? Do they have an underlying sociopolitical agenda in terms of language?
>
> (p. 755)

He goes on to say, "We need to enter the community fully understanding and appreciating the legitimacy of the community's practical cautions and concerns about the motives of sociolinguistic researchers." As Feagin (2002) observes, "skin color, class affiliation, speech, or education may all set the investigator apart" (p. 26).

There are also particular concerns in Deaf communities, concerns directly tied to the history of deaf education and to how research on sign languages has taken place. Oralism, as mentioned earlier, played an important role in Deaf education. Even though Deaf education in the United States began in 1817 with sign language as the medium of instruction, by 1880 the oral method of instruction was well established in the White schools (Lane, Hoffmeister, & Bahan, 1996). As Burch and Joyner (2006) note, "the rise of oralism … motivated schools across the country to replace deaf teachers with hearing instructors who would speak to students rather than sign with them" (p. 21). In the mid-1970s, in light of low reading levels in deaf students, the transition was made to the simultaneous use of speaking and signing, based on the theory that if deaf students could see English being produced on the mouth and hands, it would help them learn English. Specific manual codes for English (MCEs) were devised, such as Signing Exact English (Gustason, Pfetzing, & Zawolkow, 1972), which purported to represent the syntax, morphology, and lexicon of spoken English. As Ramsey (1989) states, "The developers built the requisite MCE lexicon by borrowing ASL signs, modifying ASL signs with handshape features from the manual alphabet, and inventing signs specifically to represent English derivational and inflectional morphemes" (p. 123). She goes on to observe that "[t]he materials used to construct SEE 2 are highly valued linguistic resources in the deaf community: ASL lexical items and the medium of signing itself. These resources are being used to promote the linguistic values of another community" (p. 143) – that is, the teachers, parents and educational administrators who see MCEs as an answer for teaching deaf children.

At the same time that MCEs were being devised, research on the structure and use of sign languages was getting under way in many places, with many members of the Deaf community serving as participants and sign models for hearing researchers. It was not infrequent for this research to be published with only a brief mention, or no mention, of these informants and models, which naturally led to resentment. Singleton, Jones, and Hanumantha (2012) conducted a focus group study with members of the Deaf community and researchers. They report that two main issues emerge: lack of trust and confidentiality. The lack of trust has to do in part with feelings of tokenism on the part of Deaf researchers, "feelings of being exploited and that they had not received adequate credit for their contributions to the work." Resentment can also arise concerning the ownership of the research findings, and "[s]ome resented the academic superiority of English over American Sign Language in the publication world and the fact that published materials are predominantly in English." The fact that community members have frequently not been involved and empowered has led to caution and, often, reluctance by community members to cooperate with researchers, a reluctance that contact people have to mediate.

Finally, issues of anonymity need to be clearly and carefully handled during the consent process, so that subjects explicitly either provide or do not provide consent to having their images shown as part of conference presentations or in publications. Singleton et al. (2012) note the "need to translate informed consent documents into the native language of the host community (i.e., videos using ASL) to ensure that Deaf participants, who may have limited English proficiency, are offered accessible information regarding their rights as research participants" (p. 4). Of course, the actual tools used can include interviews, structured elicitation, and questionnaires, as well as free conversation sessions. The first three must be designed with the issues discussed here in mind: Who is doing the interviewing and the elicitation? Are questionnaires written in English entirely accessible to all deaf ASL users or does the researcher need to go over the questionnaire with the subject? and so forth.

References

Burch, S., & Joyner, H. (2007). *Unspeakable: The story of Junius Wilson*. Chapel Hill: University of North Carolina Press.

Feagin, C. (2002). Entering the community: Fieldwork. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 20–39). Malden, MA: Blackwell.

Gustason, J., Pfetzing, D., & Zawolkow, E. (Eds.). (1972). *Signing Exact English*. Los Alamitos, CA: Modern Signs Press.

Labov, W. (1972a). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1972b). *Language in the inner city*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1982). Objectivity and commitment in linguistic science. *Language in Society, 11*, 165–201.

Lane, H., Hoffmeister, R., & Bahan, B. (1996). *A journey into the deaf-world.* San Diego, CA: DawnSign Press.

Lucas, C., Bayley, R., & Valli, C. (2001). *Sociolinguistic variation in American Sign Language*. Washington, DC: Gallaudet University Press.

McCaskill, C., Lucas, C., Bayley, R., & Hill, J. (2011). *The hidden treasure of Black ASL: Its history and structure* [Book + DVD]. Washington, DC: Gallaudet University Press.

Milroy, L. (1987). *Observing and analysing natural language*. Oxford: Blackwell.

Ramsey, C. (1989). Language planning in Deaf education. In C. Lucas (Ed.), *The sociolinguistics of the Deaf community* (pp. 123–146). San Diego, CA: Academic Press.

Schembri, A., McKee, D., McKee, R., Pivac, S., Johnston, T., & Goswell, D. (2009). Phonological variation and change in Australian and New Zealand Sign Languages: The location variable. *Language Variation and Change, 21*, 193–231.

Singleton, J., Jones, G., & Hanumantha, S. (2012). Deaf friendly research? Toward ethical practice in research involving Deaf participants. *Deaf Studies Digital Journal*, *3*. Retrieved from http://dsdj.gallaudet.edu

Wolfram, W. (2013). Language awareness in community perspective: Obligation and opportunity. In R. Bayley, R. Cameron, & C. Lucas (Eds.), *The Oxford handbook of sociolinguistics*. New York: Oxford University Press.

Vignette 6c
Special Issues in Collecting Interview Data for Sign Language Projects

Joseph Hill

Since the emergence of sociolinguistics as a field, methods of data collection have been continually refined to capture natural language samples. For sign language projects, capturing targeted data in a natural form is a challenge because of the visual nature of sign languages and a set of social characteristics that are unique to Deaf communities. With these in mind, two issues need special consideration: minimizing the problem of the observer's paradox and being aware of the sensitivity of signers to the audiological status and ethnicity of the interlocutors.

The Observer's Paradox

Sociolinguists are interested in capturing utterances that are spontaneously produced in a specific context, but it is known that when language users are aware that they are being observed, they may exhibit self-consciousness in their language production and adjust their language to the perceived preference of a researcher. The presence of a recording device can make language users feel self-conscious. Even with the recording device concealed, the mere presence of a researcher influences language users' linguistic behavior. This phenomenon has been addressed by sociolinguists starting with Labov (1972), who discussed what he referred to as the "observer's paradox."

Researchers conducting sign language projects also face the issue of the observer's paradox. However, key differences between spoken language and sign language are the modalities that affect the use of a recording device in data collection. With spoken language in the oral-and-aural modality, researchers enjoy flexibility in the choice of recording format, which can be audio only or audio-visual. With advances in audio recording technology, powerful audio recording devices have become increasingly portable, affordable, and less distracting. With sign language in the visual-and-kinetic modality, however, a video recording device is an absolute necessity and the filming process is usually more overt. To ensure visual clarity in the filming of a signing production, signers must be in a well-lit setting and with their heads, hands, and torsos entirely visible to a camera. Also, the seating must be arranged to help with the clarity of the signing for the interlocutors to see each other and for the camera to record; for example, a pair of signers are seated next to each other with their fronts turned slightly toward one another and a group of signers seated in a half-circle. In some cases, a video camera must be placed close to the signers to capture a full view of the signing. With these arrangements and the use of video recording devices that are necessarily more obtrusive, the problem of the observer's paradox becomes much more acute. For example, in Lucas and Valli's study (1992) a few signers chose to use contact signing (a mixed system of ASL and Signed English's core features along with the continuous voiceless mouthing, which is the common feature) or Signed English (an invented manual code for English) instead of ASL, even though the interlocutors were Deaf ASL native signers. The signers' self-consciousness (which led them not to use ASL) was caused by the relative formality of the interview situation, which included the video camera's presence and the lack of familiarity with the interviewer and other interviewees (Lucas & Valli, 1992).

To address the problem of the observer's paradox and the fact that language users may be inhibited in their language production when they are aware of being observed, Labov (1972; 1984) developed the sociolinguistic interview to encourage speakers to use the vernacular or everyday language. Since the goal is to gather as much informal language production as possible, the sociolinguistic interview is designed to reduce the power differential between the interviewer and the interviewee(s) by avoiding a formal language variety, keeping questions brief, and including topics (such as childhood games, dating patterns, marriage and family, dreams) likely to encourage informal language production. Also, the chance of obtaining informal language production may improve if the interviewer shares similar social characteristics with the interviewee(s). The sociolinguistic interview technique has been shown to be effective in sign language projects (Lucas & Valli, 1992; Lucas, Bayley, & Valli, 2001; McCaskill, Lucas, Bayley, & Hill, 2011). Sign language projects also employ a technique that allows the participants to engage in a conversation without the interviewer's presence, which has been shown to be effective as well.

Sensitivity to Social Characteristics

Taking into account the social characteristics of interviewers and interlocutors is a second issue that researchers should consider when conducting interviews with members of Deaf communities. As Lucas and Valli (1992) show, social sensitivity is often manifested in switching between ASL, Signed English, and contact signing.

Sociolinguists have suggested that the production of informal language can be encouraged when interviewers share the same ethnicity as their interviewees (Rickford & McNair-Knox, 1994). Similarly, some Black Deaf participants in McCaskill et al.'s (2011) Black ASL study explain that they stylistically shift their signing when engaging in a conversation with a White signer. These instances of style shifting can be explained by Giles' Accommodation Theory (1973), which accounts for how language behavior may change according to the perceived language preference of an interlocutor.

In sign linguistics, ASL users are also sensitive to a signer's audiological status (e.g., Deaf or Hearing). The Deaf/Hearing dichotomy is a relevant criterion in defining in- and out-groups in the American Deaf community and is used as a guide in determining a signer's language preference or skills (Hill, 2012). The terms "Deaf" and "Hearing" have particular meanings in the Deaf community: "Deaf" is used to describe someone who is a skillful ASL signer who understands and observes the values, behavior, and customs of the Deaf community, while "Hearing" is used to describe someone who is not as skillful in their use of ASL and is less familiar with the Deaf community. Although a signer's audiological status is part of the Deaf and Hearing identities, the audiological status is not visible, so, instead, signing skills are used as an indicator of one's audiological status. Even though a number of identities are relevant in the Deaf community (e.g., hard-of-hearing, late-deafened, mainstreamed student, cochlear implant user, hearing children of deaf adults [CODA], hearing siblings of deaf people), Deaf and Hearing identities have a particularly powerful influence on language production (Hill, 2012).

The social considerations of racial/ethnic background and audiological status can also interact to affect interview situations. For example, at some point during data collection for the Black ASL project (McCaskill et al., 2011), a White Hearing researcher who was skillful in ASL was mindful of the influence of her racial identity and audiological status on the sociolinguistic interview between a Black Deaf researcher and a Black Deaf interviewee; she managed to lessen her influence by staying in the background during the interview. At the conclusion of the interview, the interviewee met with the White Hearing researcher and signed with her. When the interviewee asked about the researcher's audiological status, the interviewee made a dramatic shift across the modalities from signing to speaking, even though they had understood each other's signing perfectly prior to the discovery of the White researcher's audiological status as Hearing. This is a striking example of the influence of audiological status on one's language use, but it is in fact quite common for Deaf signers to switch to contact signing or Signed English when they learn the audiological status of a Hearing person (Lucas & Valli, 1992).

Conclusion

In summary, researchers who are conducting sign language projects must always be mindful of the audiological status and racial/ethnic identity of interviewers and interlocutors in relation to the researcher's goal of obtaining targeted language samples. It is always a challenge to make signers comfortable in a setting with the presence of a video camera, but researchers can overcome the problem of the observer's paradox by both following the design of the sociolinguistic interview and using an interviewer who shares the same audiological status and racial/ethnic background as the interviewees.

References

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics, 15*, 87–105.

Hill, J. C. (2012). *Language attitudes in the American Deaf community*. Washington, DC: Gaullaudet University Press.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1984). Field methods of the project on language variation and change. In J. Baugh and J. Sherzer (Eds.), *Language in use: Readings in sociolinguistics* (pp. 28–53). Englewood Cliffs, NJ: Prentice Hall.

Lucas, C., & Valli, C. (1992). *Language contact in the American Deaf community*. San Diego, CA: Academic Press.

Lucas, C., Bayley, R., & Valli, C. (2001). *Sociolinguistic variation in American Sign Language*. Washington, DC: Gallaudet University Press.

McCaskill, C., Lucas, C., Bayley, R., & Hill, J. (2011). *The hidden treasure of Black ASL: Its history and structure*. Washington, DC: Gallaudet University Press.

Rickford, J. R., & McNair-Knox, F. (1994). Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In D. Biber and E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 235–276). New York: Oxford University Press.

Vignette 6d
Other Interviewing Techniques in Sociolinguistics

Boyd Davis

The sociolinguistic interview has been critiqued for a number of years as being overly narrow or restrictive (Butters, 2000; Wolfson, 1976). This reaction may have contributed to the recent development of what Bucholtz and Hall (2008) call "new coalitions," the deliberate involvement of interactional and ethnographic approaches from conversation analysis and linguistic anthropology:

> [T]he use of interview methodologies, so widespread in sociolinguistics and linguistic anthropology, must be matched by the use of ethnographic and interactional methods of data analysis, in order to ensure that researchers approach interviews not as providing mere background information or as a medium from which to extract linguistic variables but as richly contextualized linguistic data in their own right.
>
> (p. 416)

Laihonen's (2008) study of language ideologies highlights how work on interaction and performance (e.g., Schilling-Estes, 1998) converges with the emphases of conversation analysis. Mendoza-Denton's complex *Homegirls* (2008) exemplifies how the sociolinguistic interview is enriched by a sociocultural, ethnographic context; as Fought (2009) comments, the text "argues convincingly that ethnographic studies focusing on situated practices and participants' own categories are crucial to sociolinguistic research" (p. 268).

Valuable and capable of adaptation as it is, the sociolinguistic interview is not, of course, the only way to elicit data that will support the study of such topics as identity, stylization, historical change, variation, dialect, and the vernacular, all of which are crucial components in sociolinguistics. For example, Feagin (2001) reminds us that postal, phone, and in-person rapid surveys can be highly useful. She cites both Labov's Telsur project (n.d.) and Bailey and Dyer (1992), who obtained random samples by working with the Texas Poll telephone survey, putting their own questions into the protocol for some polls, and obtaining tape recordings of one full poll. The expanded use of the internet for research and teaching not only supports data collection but also offers a way to involve students in relatively transparent analysis that can incorporate interviews: see Kiesling's (2003) study of teen slang, the *Dude* survey, and Van Herk's (2003) "Very Big Class Project," in which information gathered from the internet served as the data source. The internet now provides a rich source of data, particularly via blogs and video plus commentary on youtube.com.

The recent surge of emphasis on sociolinguistic features of big vs. "small" models for narratives (Wilson & Stapleton, 2010) and stance and interaction have led sociolinguists to explore these areas. For example, stance-shift analysis, a corpus-based, computer-mediated coding of 24 stance variables (taken from the burgeoning literature) as they shift in frequency across successive standardized interview segments can support the infusion of sociolinguistic analysis into studies of language and communication in other disciplines (Lord, Davis, & Mason, 2008). Englebretson (2007) claims that stance is "a personal belief or attitude" and indicates "a social value" (pp. 10–11); it is therefore both interactive and indexical of feelings and attitudes. Indeed, as Jaffe (2009) claims, "Social identity can thus be seen as the cumulation of stances taken over time" (p. 10). Stance-shift analysis quantifies both the frequency and the interconnections among the word patterns by which people indicate shifts in their stance. Each segment of a transcript is coded for the presence and frequency of the variables; scales created by multivariate statistical analysis (factor analysis and clustering) of two dozen language categories identify the key areas of a transcript in which a speaker has shifted stance. For example, in Lord et al. (2008), stance markers suggested how sex offenders deflect personal responsibility and justify their actions as "reasonable."

In a workshop for the New Ways of Analyzing Variation 39 meeting, Mallinson, Childs, and Van Herk (2010) identified approaches to interviews "that mitigate concerns surrounding the collection of language data and/or used other methodological approaches to adapt to social and technological changes," including:

- ethnographic interviews that obtain language data informed by relatively long-term participant observation within a community;
- interviews that differ by structure, as when community members are interviewers, participant-recorders, or otherwise co-participants;
- interviews that differ by topic, as when data are collected for non-language oriented purposes, whether research driven or not;
- speaker-generated online data produced in contexts where linguists have not been involved.

Another category is interviews that differ by speaker competence, as when participants have not been assumed to have the ability to present useful speech data. Sociolinguists could contribute much to psycholinguistic explorations of cognitive impairments such as Asperger syndrome (Niemi, Otsa, Evtyukova, Lehtoaro, & Niemi, 2010), drawing perhaps on research and experience since the 1990s with recording conversational interaction with speakers with Alzheimer's disease. Hamilton (1994), Ramanathan (1997), Davis (2005), and others have continued to identify aspects of style, ideology, and variation in discourse of cognitively impaired persons. Cognitive mapping, an ethnographic technique, is currently being used with persons with chronic disease, such as diabetes (Davis, Pope, Mason, Magwood, & Jenkins, 2011). In 1997, adolescents were asked to sketch maps containing important locations for talking (Davis, Smilowitz, & Neely, 1997); more recently, adults with diabetes have been asked to draw a map portraying the world they live in as people with diabetes, to help them speak more freely about places, people, and daily events that are important to them (Davis, Pope, & North-Lee, 2010). Drawing similar maps initiates the conversational interviews with older adults in the corpus called the Carolinas Conversation Collection, augmented by additional probes based on Kleinman's (1998) explanatory model of illness. If the speaker has advanced cognitive impairment, the interviewer or conversation partner can sketch the map and introduce or reinforce topics that could be associated with the "places" or nodes on the map. For example, an interviewee called Glory Mason offered details about food, family, or aspects of daily life in her youth whenever a farm was mentioned to her:

[BD: You lived on a farm when you were young.]
GM: I just lived in a regular farm home. Farmed cotton, corn, eh – everything you – grow on a farm. That's right. I had a big ol' cotton bag tied around me, pickin' a hundred pounds of cotton – uhhmm hmmm…
(Davis, 2010, p. 394)

The expanded development of language corpora – whether focused on material exhibiting sociolinguistic concerns, such as the Sociolinguistic Archive and Analysis Project (SLAAP) (Kendall, 2009; see also Kendall, Chapter 12), or combining medical and social science research approaches, such as the Carolinas Conversations Collection (Pope & Davis, 2011) – will encourage the archiving of new categories of data to support multiple analyses. As in the Teenage Health Freak corpus, part of the Nottingham Health Communication Corpus, email messages, instant messaging and texts from other wireless technologies can support the collection of data amenable to sociolinguistic analysis that can augment or stand alone beside the interview as a data resource.

References

Bailey, G., & Dyer, M. (1992). An approach to sampling in dialectology. *American Speech, 67*, 3–20.

Bucholtz, M., & Hall, K. (2008). All of the above: New coalitions in sociocultural linguistics. *Journal of Sociolinguistics, 12*, 401–431.

Butters, R. (2000). Conversational anomalies in eliciting danger-of-death narratives. *Southern Journal of Linguistics, 24*, 69–81.

Davis, B. H. (Ed.). (2005). *Alzheimer talk, text and context: Enhancing communication*. Basingstoke, UK: Palgrave Macmillan.

Davis, B. (2010). Interpersonal issues in health discourse. In M. A. Locher & S. L. Graham (Eds.), *Interpersonal pragmatics*. Berlin: de Gruyter.

Davis, B. H., Pope, C., Mason, P. R., Magwood, G., & Jenkins, C. M. (2011). "It's a wild thing, waiting to get me": Stance analysis of African Americans with diabetes. *Diabetes Educator, 37*, 409–418.

Davis, B., Pope, C., & North-Lee, B. (2010). Expanding explanatory models: The embodiment of agency and accountability in the talk of diabetics. Paper presented at the Conference on Communication, Medicine and Ethics 11 conference. Boston, MA.

Davis, B., Smilowitz, M., & Neely, L. (1997). Speaking maps and talking worlds: Adolescent language usage in a New South community. In C. Bernstein, T. Nunnally, & R. Sabino (Eds.), *Language variety in the South revisited*. Birmingham: University of Alabama.

Englebretson, R. (2007). Stancetaking in discourse: An introduction. In R. Englebretson (Ed.), *Stancetalking in discourse* (pp. 1–26). Amsterdam: John Benjamins.

Feagin, C. (2002). Entering the community: Fieldwork. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 20–39). Malden, MA: Blackwell.

Fought, C. (2009). [Review of the book *Homegirls*, by N. Mendoza-Denton]. *Language in Society, 38*, 266–269.

Hamilton, H. (1994). *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge: Cambridge University Press.

Jaffe, A. (2009). Introduction: The sociolinguistics of stance. In A. Jaffe (Ed.), *Stance: Sociolinguistic perspectives* (pp. 29–52). New York: Oxford University Press.

Kendall, T. (2009). The Sociolinguistic Archive and Analysis Project. North Carolina State University. Retrieved from http://ncslaap.lib.ncsu.edu

Kiesling, S. F. (2003). *Dude*: Supplementary materials. Retrieved from http://www.pitt.edu/~kiesling/dude/dude.html

Kiesling, S. F. (2004). Dude. *American Speech, 79*, 281–305.

Kleinman, A. (1998). *The illness narratives: Suffering, healing and the human condition*. New York: Basic Books.

Labov, W. (n.d.). Telsur Project. Retrieved from http://www.ling.upenn.edu/phono_atlas/home.html

Laihonen, P. (2008). Language ideologies in interviews: A conversation analysis approach. *Journal of Sociolinguistics, 12*, 668–693.

Lord, V. B., Davis, B., & Mason, P. (2008). Stance-shifting in language used by sex offenders: Five case studies. *Psychology, Crime and Law, 14*, 357–379.

Mallinson, C., Childs, B., & Van Herk, G. (2010). Participants, purpose, and process: An interactive workshop on data collection in variationist sociolinguistics. Workshop presented at the New Ways of Analyzing Variation 39 conference. San Antonio, TX.

Mendoza-Denton, N. (2008). *Homegirls: Language and cultural practice among Latina youth gangs*. Malden, MA: Blackwell.

Niemi, J., Otsa, L., Evtyukova, A., Lehtoaro, L., & Niemi, J. (2010). Linguistic reflections of social engagement in Asperger discourse and narratives: A quantitative analysis of two cases. *Clinical Linguistics and Phonetics*, *24*, 928–940.

Pope, C., & Davis, B. H. (2011). Finding a balance: The Carolinas Conversations Collection. *Corpus Linguistics and Linguistic Theory, 7*, 143–161.

Ramanathan, V. (1997). *Alzheimer discourse: Some sociolinguistic dimensions*. Mahwah, NJ: Lawrence Erlbaum.

Schilling-Estes, N. (1998). Investigating "self-conscious" speech: The performance register in Ocracoke English. *Language in Society, 27*, 53–83.

Van Herk, G. (2003). The very big class project: Collaborative language research in large undergraduate classes. *American Speech, 83*, 222–230.

Wilson, J., & Stapleton, K. (2010). The big story about small stories: Narratives of crime and terrorism. *Journal of Sociolinguistics, 14*, 287–312.

Wolfson, N. (1976). Speech events and natural speech: Some implications for sociolinguistic methodology. *Language in Society, 5*, 189–209.

7 The Technology of Conducting Sociolinguistic Interviews

Paul De Decker and Jennifer Nycz

The goals of a sociolinguistic interview, as articulated by Labov (1984), are twofold: first, to obtain "a large volume of recorded speech of high enough quality" for analysis; and second, to obtain "records of vernacular speech" of our participant(s) (p. 29). At times, these aims may seem to be in opposition to each other. While this chapter won't help resolve this conflict, it will help you attain goal number 1.

Setting Goals and Thinking Ahead

How do you determine whether a recording is of "high enough quality"? To some extent, it depends on what you want to do with your recording. If your aim is to carry out phonetic transcription or acoustic analysis, then you need a clear, richly sampled recording that allows you to decisively identify and/or measure the features of interest. If instead you want to count the occurrences of some morphosyntactic feature in the speech of your participants, analyze the way a conversation is structured, or explore the discursive strategies your interviewee uses to frame a topic, a "good enough" recording may simply be one that allows you to identify lexical items and the utterances that contain them. Not all recordings need to be of the highest quality modern technology can deliver; you don't need to preserve the component sound frequencies required for voice quality analysis if you are looking at the content and structure of the conversation between two informants. That said, we strongly recommend that all linguists, regardless of their current interest in phonetic questions, aim for the highest-quality recordings. The field of linguistics is slowly moving toward collaborative sharing of data, in the spirit of providing open-access data to researchers around the globe (see Childs, Van Herk, & Thorburn, 2011; Kendall, Chapter 12; Kretzschmar, Vignette 12a). We believe that progress toward open-access initiatives in linguistics will be aided by a unified set of recording practices. Therefore, we recommend digital recording technology under the assumption that you may eventually make your recordings available to other researchers for analyses even if you yourself are unlikely to perform them.

Operating under this assumption can benefit not just the larger research community but also your own future work. Consider the following scenario: You plan and carry out a sociolinguistic study with the initial goal of examining syntactic variation. During the analysis phase of your research, it becomes clear to you that this variation has interesting and important prosodic effects in your data, and you would like to explore these effects quantitatively by measuring vowel formants and pitch contours. Fortunately, because your recordings were recorded in a higher-quality format than required for your initial research goals, you are not confined to your first, planned analysis.

Equipment

Let's start with a basic outline of the equipment you'll need: recorder, power supply(ies), microphone, and storage media. For each item, we discuss the most important factors to consider as you decide what equipment to use.

1. *Recorders*. For the first few decades of sociolinguistic research, magnetic tape recorders were widely used to collect audio data. While these analog recorders are rarely used today, it is not uncommon to find data stored on cassette tape. We recommend converting old tapes to digital format before beginning any type of transcription or analysis; while foot pedals aid in transcribing language data from tape recorders, constant stopping and playing of the tape is destructive to the playback mechanism as well as the tape itself. If you are collecting fresh data, however, there is no reason to rely on tape, and every reason not to: recordings made on analog devices may be susceptible to motor noise from the recording mechanism, the tape itself is subject to physical wear and tear, and the imminent obsolescence of this technology means that it will soon become difficult, if not impossible, to access your data.

As a contemporary sociolinguist, then, you will need to decide on a digital recorder. There are many options on the market, but you can narrow them down by considering a few clear criteria. Minimally, you want a device that can record in uncompressed format, is capable of taking batteries as well as plug-in power, allows you to adjust the volume input level (and observe these levels as you record), can accept an external microphone, and records to memory cards.

2. *Digitization and compression*. Sociolinguistic interviews should be recorded at a sampling rate of 44,100 Hertz (Hz) and at a resolution of 16 bits, to an uncompressed audio format (such as a WAV file or an AIFF file).

What does this mean? Without getting into the nitty-gritty of digital signal acquisition (see Lebow, 1997, for a detailed introduction to this topic), modern digital technology works by recording sound in discrete pockets of data. Suppose your ears were fitted with earmuffs that impeded sound. A small door built into the earmuff opens for one second and then closes again for one second. This would allow you to catch conversations, if only for one second at a time. This is how digital signal capturing works: any digital recorder captures sound in brief repeated intervals. Recording at a rate of 44,100 Hz captures 44,100 snapshots every second, which is enough for a reasonably faithful reproduction of the original signal with no major audible or acoustic degradation. A general rule is that the sampling rate will record frequencies up to half its value (known as the Nyquist Frequency; see, for example, Johnson, 2002). So, a 44,100 Hz sampling rate will capture frequencies up to 22,500 Hz. This rate is appropriate for recording and analyzing human speech. Importantly, it will capture the higher-frequency sounds in the human voice, including those associated with sibilant noise (Ladefoged & Johnson, 2010). Of course, it is possible to find recorders that will sample at an even higher rate than 44,100 Hz. However, while their use will yield an even more acoustically detailed recording, it will also result in a larger file to store and manipulate, with no clear gains in terms of the phonetic results that can be wrung from the data.

Sampling rate indicates how often the signal is captured. It is also important to know how accurate this sampling is. This property is referred to as *bit depth*. Higher bit rates preserve more information by improving the signal to noise ratio (SNR). The standard bit rate used in sociophonetic recording is 16 bit. Again, many recorders will allow much higher bit rates, but is it not clear that there are sufficient acoustic gains to justify the increase in file size.

Finally, once you have sampled your signal at a particular rate and bit depth, you want to be sure that the file format in which you save your data retains all of this precious information. For this reason, you want a recorder that records to an uncompressed format, such as WAV or AIFF. Compressed formats may be either lossy or lossless; neither type is best for saving your data. Lossy formats such as MP3 selectively remove information from the digital signal to reduce file size – that is, some component of the signal is irretrievably *lost*. While this compression may not have significant audible effects, it will distort the acoustic signal, possibly hindering phonetic analysis. Lossless compressed formats, such as FLAC, reduce file size by eliminating only predictable data, allowing for a reconstruction of the original captured signal; while these formats are better than lossy formats, the processing time needed to rebuild these signals (and convert them to a format readable by popular phonetic analysis programs such as Praat; see Boersma & Weenink, 2011) makes these options less than optimal.

Now that we've looked at the basics of sound recording, which machines currently on the market will do the job? While we cannot recommend specific brands or models, we discuss the major types of digital recorders along with the advantages and disadvantages of each.

3. *Recorder types*. The current standard in sociophonetic research is to collect data using a solid-state recorder. These devices have no moving parts, noiselessly recording your data to compact reusable memory cards whose contents can easily be transferred to computer hard drive via USB cable. Moreover, these recorders tend to be more durable than recorders with moving parts, making them a good choice for use in the field.

Other digital options include DAT (Digital Audio Tape) and Mini-disc, though we do not recommend using either of these recorder types. Both devices require additional data conversion or digitization after recording, which is an unnecessary hassle. Both are subject to a certain amount of machine noise, which will make its way into your recording and your acoustic analysis. Finally, looming obsolescence means that data recorded to DAT or Mini-disc will need to be transferred to other formats in order to remain accessible.

It may be tempting to record directly onto your laptop computer, using one of the many available sound-recording programs and either the computer's built-in microphone or an attached microphone. While this option is convenient and economical, the noise in your recordings that will result from the laptop fan or the intermittently spinning hard drive (and perhaps the odd pop-up alert ping) rules it out as a choice for high-quality recordings.

Power Supply

Ideally, your recorder should allow batteries to be used during operation, as AC power adapters can contribute noise to your recordings. That said, always bring your AC adapter with you into the field, in case your batteries run out. Your choice of battery will mostly be determined by your choice of recorder, but opt for long-life batteries to reduce the chances of running out of juice mid-interview.

Microphones

The microphone is as important as the recorder, if not more so. The mic's job is to convert acoustic power into electric power, which is ultimately converted to digital code. It is not necessary to know every detail about how different kinds of microphones operate, but you should know what kind of microphone you are using and why.

1. *Key points in choosing a microphone*. Before we consider the types of microphones available on the market, we review key points to help you choose the most appropriate microphone. Microphones vary according to their frequency response and ability to manage signal levels in relation to ambient noise (an issue related to low amplitude sensitivity) and in terms of the directions from which they pick up sound.

First, the frequency response of a microphone refers to that mic's sensitivity in decibels (db) over a range of frequencies. Plichta (2002) notes that microphones with a "wide and flat frequency response curve" are optimal when making recordings for acoustic analysis (p. 2) (see also Hall-Lew & Plichta, Vignette 7a).

Second, some microphones are better than others at capturing sound. This property refers to the ability to focus on the signal among the noise. The SNR, measured in decibels, is a way to express how loud the desired sound (the participant's voice) comes across in a recording in relation to other unwanted noise from the recording environment. There are ways to reduce the noise component of the ratio and increase that of the signal (see below), though a good microphone with a high SNR will give you a head start in attenuating the loudness of the noise so that a higher-quality signal is captured.

Finally, the directionality of a microphone refers to which sounds are picked up in relation to the position of the microphone. Omnidirectional mics are, in theory, equally sensitive to sound coming from all directions. Directional mics pick up sound from only one direction: in front of the microphone head. Cardioid mics, a type of directional mic, pick up virtually no sound from behind the microphone.

Directional mics are good for recording single voices and reducing ambient noise; however, care must be taken to place these mics carefully, to ensure that the single voice you intend to capture will be picked up.

2. *Microphone types.* There are two types of microphones available to consumers. Condenser mics have the wide frequency response required to capture the range of frequencies that characterize speech, and they exhibit a high input sensitivity, recording a wide dynamic range of quiet to loud sounds. These microphones send a strong signal to the recorder, which minimizes the need to increase the input volume on the recorder (and thereby minimizes the small amount of unwanted noise that would have otherwise been generated). While they produce very high-quality recordings, these mics can also be fragile: if mishandled or dropped, they can easily break. Condenser microphones need power from an external source such as a battery pack or the phantom power supply built into some recorders (see below). They are most often found with an XLR connector, one of the most efficient ways to connect a microphone to a recording device (Plichta, 2002). XLR cables allow for the use of phantom power or battery packs to power the condenser mic. The end of an XLR cable has between three and seven pins or holes. The standard for audio connection is three.

Dynamic microphones, on the other hand, do not need batteries or phantom power and are often sturdier than condenser microphones. However, they produce a weaker signal and typically have an exaggerated frequency response, effectively boosting the amplitude of frequencies in the range of 3,000 Hz while downplaying the presence of other frequencies. They are attached to a recording device via the most commonly found ¼- or ⅛-inch TRS connection type. TRS connectors allow for plug-in power – that is, electricity derived from your recording device – to power the microphone.

Condenser microphones are most often used for recording speech in sociolinguistic interviews. While care must be taken in transporting and handling these mics in the field, their superior frequency response and input sensitivity make them the better choice.

3. *Placement of the microphone.* Of course, even the best microphone will not yield an excellent recording if it is not placed correctly. If the microphone is too close to the speaker (which can be the case in head-mounted or free-standing setups), some frequencies might get overrepresented in the resulting sound file, and "popping" sounds like labial stops may result in transients in the acoustic signal that will disrupt phonetic analysis. In many laboratories, great care is made to place the microphone at approximately 25 centimeters from the speaker's mouth. Interviewers who use a lavalier (a type of condenser mic) might attach it to the interviewee's lapel, which has the same effect as placing the mic about 25 centimeters from the speaker's mouth and ensures that this distance is maintained across interviews.

Other Related Equipment

All microphones require an electrical signal to power them. Some get their power from battery packs, others from the microphone input on the recorder, and yet others from a dedicated power unit called phantom power (which can also be housed inside the recording device). If you are using a condenser microphone, make sure that the recorder has phantom power; if not, you will need to purchase a dedicated battery pack. Dynamic microphones, on the other hand, do not require any additional power other than the electrical current from the input hole on the recorder.

If your recording device does not have an adjustable volume input, consider using a pre-amplifier. A pre-amp increases the strength of the signal without degrading the signal-to-noise ratio. However, it is easiest to avoid this issue by choosing a recorder that either allows you to adjust the volume input manually or does so automatically.

If you will be traveling with your equipment (as most sociolinguists do), you should also have a dedicated carrying case or bag with room for your recorder, mic, power source (and spares!), cables, and any other equipment. Ideally, the case will have separate compartments or sections for each item, to prevent cables from tangling and pieces from rattling against each other (see also Hall-Lew & Plichta, Vignette 7a).

Choosing a Suitable Environment

Although this is a chapter on the technology of conducting sociolinguistic interviews, we feel the need to comment on the technology/environment interface. Technology alone cannot produce the optimal recording setup; it must be a mix between your equipment and where you do the recording. Generally speaking, the quieter the environment the better, though many traditional sociolinguistic interviews took place (and still take place) in living rooms around the world. While environmental noises are considerably louder in a living room than in a soundproof booth, there are things you can do to reduce the ambient noise outside of laboratory conditions. Become familiar with things in various environments that make noise, such as computers, fans, air-conditioning units, fridges, and clocks; a good exercise might be to run your recorder in your apparently quiet living room or kitchen, and see just how many sounds it picks up. When you are about to record a speaker, make sure that such sound sources are turned off or moved to another room; if this is not possible, try to move your interview to a quieter place. Be aware also of animate sources of noise. While it might seem harmless and relaxing to allow the family dog or cat to hang around while you conduct your interview, noises such as purrs, pants, and meows will make their way onto your recording. (Removing pets from the room is also good for the health of your equipment: cats find mic wires in particular to be irresistible.) Finally, it is important to (delicately!) inspect the interviewees themselves for possible sources of noise: if your interviewee is wearing a jangly necklace, for instance, it may be prudent to ask him or her to remove it.

At the same time, while the presence of certain items may contribute unwanted noise to your recordings, so may the absence of things: if the room in which you want to record is too empty or large, then sounds in the room (including speech sounds) will echo back into your recording, seriously impeding your ability to make clear decisive phonetic measurements.

Recording Group Conversations

It is not uncommon for a sociolinguistic interview to involve more than two people. Often friends are included to put speakers at ease and thus minimize the effect of the interview situation. If you are interested in the linguistic behavior of all of those present, there are two major options: record to one device using a mixer, or record to multiple devices. Using a mixer to record to one device will allow you to manage the volume levels of each input microphone so that all speakers are relatively equally heard. A portable mixer, while not inexpensive, is a device suitable for managing the volume levels of multiple informants. The great disadvantage of mixers is that they ultimately record to a single audio file. This can be a significant problem when performing acoustic analysis of a conversation in which speakers talk over one another, as there is no way to recover the individual voices. Moreover, a mixer is yet another piece of equipment sitting in the room with you and your interviewee, drawing attention to the fact that you are recording their speech.

To avoid these problems, it is probably best to use multiple recorders, which allows you to isolate each voice to a separate audio file. Though each recorder will pick up the speech of other people in the room as ambient noise, each voice will be the most prominent on its particular recording. This will make later auditory (if not acoustic) analysis more feasible.

Storage

In this last section, we cover two aspects of the same topic, temporary and long-term data storage. What is the best way to store your data when first acquiring it, and how do you save your recordings from being destroyed by natural forces or becoming unrecoverable or incompatible with your next computer or software upgrade?

It is often said among digital photographers that the most important part of one's camera is the memory card; this is, of course, after the photo has been taken. The same applies to recording sociolinguistic interviews. Whether you are out in the field for months on end or making daily trips to your participants' homes, the best way to store your data is on multiple memory cards rather than one or two with monstrous memory capacities. A 32 GB card could fit over 70 hours of recordings, but if you store all of your acoustic eggs in this basket, you will be in serious trouble if it gets lost or damaged. We also recommend not maxing out the memory card. Always leave more room on the card than you need. For the purposes of determining how much memory you will need, the following website has a handy calculator: www.sounddevices.com/calculator/. For example, a one-hour interview recorded as a mono WAV file at a sampling frequency of 44,100 Hz and depth of 16 bits will require 303 MB of space. In this format, you could record for three hours on a 1 GB card. Store your cards in the holder they came in, away from extreme temperatures.

When faced with multiple brands of cards at various price points, err on the side of caution and always go with a known brand-name card. There are many cheap ones on the market, and the risk they present to your data is not worth the monetary savings. Once you have successfully transferred your recordings to a more permanent storage device, like your lab computer or an archival server, don't prepare the card for the next round by just deleting the files from it. Instead, format the card. Formatting resets the card's folder structure, maintains the performance of the card over time, and does not take much more time than merely deleting the contents.

We now want to devote space to the issue of long-term storage. All sociolinguistic interviews should be handled with a long-term storage plan in mind (see also Kendall, Chapter 12). It is not good enough simply to store your files on a lab computer or back them up to CD/DVD-ROM or a high-capacity external hard drive. These are reasonable and appropriate first steps, but all of these options are ultimately temporary storage: hard drives, optical storage media, and solid-state media cards all inevitably fail and are subject to the quality of their original construction, environmental factors, etc. Technologies also change (laser disc, anyone?), potentially trapping data forever in unreadable formats. If you think this is somewhat alarmist, consider the following: how many of the digital photos that you took eight years ago made it from your old computer to the one you are currently using? If the answer is none, then you get the point. If the answer is all of them, then you obviously made an effort to migrate your files, and you clearly get the point too. We therefore recommend a migration plan. It can be as simple as purchasing a new external hard drive every five years and copying the contents of your old drive over to the new one. Another option is to invest in an online storage service (there are many, and the prices vary) that seamlessly backs up the contents of your computer or selected folders.

Again, in planning for long-term storage, you will want to estimate the size of your corpus and determine how much space you will need to accommodate it. Knowing that an hour-long interview requires approximately 300 MB, 10 subjects (at the same rate) will take 3 GB. If you plan to interview 40 subjects, then you will need 120 GB. This is actually a drop in the bucket as far as memory storage options go. But let's also assume that you will collect data for projects in the future. If you do three more corpora like the first, then you're looking at 480 GB, which would only halfway fill a 1 TB hard drive. Whichever plan you decide on, continue to research your data storage options as newer, more efficient ones continue to appear.

Conclusion

From an equipment standpoint, here is a review of what you need to obtain recordings of *high enough quality* for acoustic analysis:

- a solid-state recorder with adjustable input volume (with optional phantom power);
- a condenser microphone with XLR cable (or battery pack);

- multiple storage cards;
- batteries for microphone and recorder;
- a dedicated carrying case or bag;
- a USB cable to transfer files;
- an external hard drive for longer-term storage.

You may be thinking to yourself: "Will all of these pieces of equipment exacerbate the observer's paradox?" While the very nature of knowing that one is being recorded produces this effect (Labov, 1972), training in interview design and plenty of practice with your equipment in advance can help dial down any anxieties on the part of the interviewer and interviewee. Before you set out to interview anyone, become familiar with your equipment, and do a few "rehearsal" recordings to ensure that everything works as you think it does. Know the functions on your recorder. Know how long the batteries will last. Know how to position the microphone. Know how many minutes you can get on your memory card. By the time you enter the field, you should be confident that you know the functions and limits of each device in your recorder bag.

References

Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer [Computer software]. Retrieved from http://www.praat.org/

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.

Johnson, K. (2002). *Acoustic and auditory phonetics* (2nd ed.). Malden, MA: Blackwell.

Labov, W. (1972). Some principles of linguistic methodology. *Language and Society, 1*(1), 97–120.

Labov, W. (1984). Field methods of the project on linguistic change and variation. In J. Baugh & J. Sherzer (Eds.), *Language in use: Readings in Sociolinguistics* (pp. 28–53). Englewood Cliffs, NJ: Prentice Hall.

Ladefoged, P., & Johnson, K. (2010). *A course in phonetics* (6th ed.). Boston: Thomson Wadsworth.

Lebow, I. (1997). *Understanding digital transmission and recording*. Piscataway, NJ: Wiley-IEEE Press.

Plichta, B. (2002). Best practices in the acquisition, processing, and analysis of acoustic speech signals. *University of Pennsylvania Working Papers in Linguistics, 8*(3), 209–222.

Vignette 7a
Technological Challenges in Sociolinguistic Data Collection

Lauren Hall-Lew and Bartlomiej Plichta

If there were a Murphy’s Law of sociolinguistic interviews, it would be that your most interesting interviews always seem to have the worst-quality recordings. Whether they take place in a particularly noisy room, or you forget to switch on a piece of your equipment, or a battery runs out midway through the interview, any number of technological challenges can come between you and your ability to collect and analyze your data. Saying that this seems to happen most often with the most interesting interviews may not be an empirically supported claim, but it is a sadly common tale. For sociophonetic work, in particular, the consequences can greatly impact your analysis. Field researchers must constantly balance the need for high-quality audio recordings with the need to minimize the level of social awkwardness with our participants. Problems arise when this balance tips too far to one side or the other. We begin by briefly discussing some of the consequences of paying so much attention to the pursuit of interactional “naturalness” that the usability of the audio data suffers as a result.

When Lauren began her fieldwork in San Francisco in 2008, she already had previous experience recording speech and so was more focused on making connections in the community than on the technical setup of the recording situation. Unfortunately, she paid the price. Despite checking the recording equipment briefly the night before, her very first San Francisco interview was recorded as one loud buzz – entirely useless, all because of a faulty microphone connection. It turned out later that that speaker represented a particularly interesting social demographic and would have potentially contributed significantly to the data sample. In retrospect, this situation could have been prevented or mitigated. Recording equipment must always be checked thoroughly prior to use. After that first interview, Lauren always had a secondary, battery-powered backup recording device. Even if the sound quality of the secondary recording is not as good as that of the primary source, you can at least retain a recording of the lexical content of the interview.

Lauren’s second-worst recording was again with a particularly interesting speaker who was recorded fairly early on during fieldwork. The recording was usable, but just barely; the quality of the recording could have been much improved with just a small dose of fieldworker courage. Every field linguist encounters the problem of the participant who invites them to make their recording in a noisy environment. Outdoor settings are famous for this, with animal noises, traffic noises, noises from other people, and the wind. But indoor settings can be just as damaging, with noises from appliances, clocks, computers, television or radio, air conditioning, squeaky furniture, other people passing through the room, or even the speakers themselves (clinking jewelry, tapping fingers on a tabletop, etc.). As a result, a fieldworker must never be afraid to ask an interviewee to move to a quieter setting. When you know your interviewee personally, this request is not difficult to make, but if you happen to interview someone you've just met, it can feel like an imposition – particularly if you're in that person's home. Yet despite the sense of awkwardness, it is imperative for the success of your data collection that you be able to request a move to another location. In Lauren's case, she had just completed an interview with a 40-year-old woman in her kitchen (already recorded with bad sound quality because of the hum of the refrigerator), when it suddenly became possible to have an additional interview with the woman's 16-year-old son. The boy was soft-spoken and aloof and hadn't met Lauren more than a minute before the interview. The right thing to do would have been to request a change of venue to a quieter part of the house, but the boy's mom wanted to listen in on the interview, and Lauren missed the opportunity to insist on privacy and confidentiality as the ideal excuse to relocate. The interview with the boy was therefore recorded against the backdrop of sounds of his mother making dinner: opening and closing the refrigerator and cupboards, running the faucet, and moving kitchenware. Although Lauren was aware of the damaging effect of these noises at the time of the interview, it wasn't until listening back that it became clear just how much of the data was unusable.

Many of the issues that Lauren confronted during her interviews were related to ambient noise. This type of noise can, potentially, be detected and evaluated by an experienced interviewer. However, there is another source of noise that is more difficult to detect, as it originates in the electronics of the recording chain. Bartek remembers interviewing a Polish American priest in Hamtramck, Michigan, using an analog cassette recorder (remember those?) that was plugged into an A/C outlet. The outlet happened to be improperly wired and caused the so-called 60-Hertz hum to contaminate the recording. The 60-Hertz hum (or 50-Hertz in some countries) can be caused either by ground loops or by electrostatic and electromagnetic induction from power lines. The most likely culprits are the recording device's A/C power supply or faulty cables used to connect the microphone with the recording device. Yes, it does sound complicated, but you do not need a degree in electrical engineering to try to eliminate the 60-Hertz hum.

Because the hum originates in the electrical circuitry, it is imperceptible unless you are monitoring your recording with headphones. You may be tempted to believe that wearing headphones during an interview would be awkward, but if you use small earbuds, they are less likely to be distracting to your interviewees. If you hear a hum in the recording channel, you should try to eliminate it before the session begins. The hum may be difficult to avoid, but the simplest solutions include using battery power, balanced cables with XLR connectors, or a hum-eliminating device, such as the Ebtech Hum X. By taking such simple precautions, you significantly increase your chances of obtaining hum-free recordings. (For more information about the technical aspects of noise reduction in field recording, visit Bartek's website at http://bartus.org/akustyk/noise.php.)

This brings us to another technology-related issue facing all sociolinguistic fieldworkers. Despite recent advances in miniaturization, recording equipment can be bulky and unwieldy. On the face of it, a small lavalier microphone and a pocket-sized digital recorder therefore might seem ideal. Unfortunately, with such a simple setup it is difficult to obtain the high level of spectral detail (the important acoustic information in the speech signal) and favorable signal-to-noise ratio (the difference in amplitude between speech and noise) needed for reliable acoustic analysis. Bartek frequently hears from linguists who are eager to try wireless microphones in order to further minimize equipment clutter. Unfortunately, such microphones can be problematic in that they often pick up radio-frequency (RF) interference from the surrounding area, including the interference known as GSM chatter that is caused by cell phones. Minimally, Bartek recommends that you use a battery-powered condenser microphone, an XLR cable, and a digital field recorder. Always have a pair of headphones and spare batteries, and it doesn't hurt to have a spare microphone and a couple of extra XLR cables with you as well. You can easily solve the portability issue by using a properly designed equipment bag, with a few separate chambers, cable routing holes, and easy access to the recorder's recording and playback controls. Several such bags are available on the market, or you can use a medium-size camera bag and make the necessary compartments with a little bit of foam (e.g., from an old yoga mat) and gaffer tape (which is similar to duct tape but is easier to tear with your fingers and leaves no sticky residue when peeled off). Once you assemble your kit, remember to practice (a lot!) before you start working.

It is difficult to recommend specific brands or models of equipment because of the constantly changing inventories, prices, and product availability in different markets. You should also consider warranty and technical support options to guide your purchasing decisions. Be sure to call the manufacturer (not the dealer) and explain your particular situation, to be sure that your needs are going to be met. To simplify the decision-making process, Bartek maintains a website with equipment reviews and recommendations. He recommends that you read the reviews and tutorials to help you decide which equipment to use and learn how to use it (http://bartus.org).

Finally, depending on your community of study it may be important to consider the attitudes toward technology that your speakers might hold. While most of Lauren's interviewees in both San Francisco, California, and Flagstaff, Arizona, were perfectly comfortable with the recording setup, one of Lauren's oldest rural Arizonan cowboys was, at best, distracted (and, at worst, concerned) about the pocket-sized silver-and-blue mini-disc recorder she used to record their interview (even making a disparaging comment about "technology these days.)" Although this kind of suspicion toward technology is waning, in certain cultural situations it still may be better to choose a recorder that is solid black and of a more recognizable size and shape, and to choose a lapel microphone rather than a headset microphone.

While sociolinguists working in field conditions want to maintain a certain level of consistency among recording sessions (such that one interview is comparable to another, for example), each participant and recording environment offers a unique set of characteristics and challenges. Being able to quickly assess the features of a given situation and how they might bear on the quality of an audio recording is a key skill that is just as important to fieldwork success as is the skill of asking the right question at the right time. We encourage you to participate in a field recording workshop or a class in field methods. You also might want to subscribe to a quality discussion forum, such as the H-OralHist Listserv hosted by the premiere humanities computing center, Matrix, at Michigan State University (www.h-net.org/~oralhist/).

Bartek is often asked what equipment to buy. The rule of thumb is that you should be prepared to spend as much on your recording gear as you would on a laptop computer. If you cannot afford such an expense, perhaps you can at least buy a good microphone and borrow a recorder from your department. There is a common misconception that digital technology makes recording easy and dramatically improves signal quality. It is not necessarily so. Proper recording technique is crucial to obtaining reliable data. Once you've mastered the technique, the technology will become transparent, getting out of the way of your interview, and you will be able to enjoy your interviews and focus on their content. You will be rewarded by the quality of your data.

References

H-Oralhist. (2005). HNet Humanities and Social Sciences Online. Retrieved from http://www.h-net.org/~oralhist/

Plichta, B. (2012). AKUSTYK. Retrieved from http://bartus.org

8 Surveys

The Use of Written Questionnaires in Sociolinguistics

Charles Boberg

Written surveys, or questionnaires, have long been an important means of gathering data on linguistic variation. The idea is simple enough: if you want to find out which words people use, or how they pronounce those words, or whether they find certain sentences grammatical, write up a list of the questions you want answers to, distribute it to members of the population whose speech you want to find out about, and ask them to fill it out and return it to you. This approach has been more commonly used in dialectology than in sociolinguistics, which reflects a belief that regional differences are less socially sensitive than social differences. Many respondents enjoy reflecting on how their speech differs from that of other regions but become self-conscious when asked about differences tied to levels of education or occupation, making responses to direct questioning more reliable and valuable in the former case than in the latter. While this belief is largely correct, I will argue that written surveys, which have advantages that complement their disadvantages, retain a useful role even in sociolinguistic methodology when they are deployed appropriately for certain purposes. The following discussion examines the advantages and disadvantages of surveys and the history of their use, with a special focus on the study of Canadian English, in which written questionnaires have played a major role. The chapter concludes with some methodological considerations that sociolinguists should bear in mind if they wish to maximize the benefits and minimize the drawbacks of surveys.

The Advantage of Surveys: Quantity

The principal advantage of surveys is quantity: they are capable of collecting a large amount of data in a relatively short space of time, using limited resources. Even in the days before computers and the internet, hundreds or thousands of copies of a written survey could be printed and distributed to respondents relatively cheaply. Institutional distribution and the use of intermediaries, or assistants, made this even easier: schoolteachers, for instance, could collect hundreds of survey responses from their students, which could then be forwarded to the investigator for analysis. In one of the examples to be discussed, Canadian investigators collected over 14,000 responses to the *Survey of Canadian English* from Canadian schoolchildren and their parents (Scargill & Warkentyne, 1972), an enormous dataset that allows regional, age, and sex differences to be identified with considerable confidence. Three times this quantity was collected by the first major written dialect survey, the *Deutscher Sprachatlas* (Wenker, Wrede, Mitzka, & Martin, 1927–1956). Carried out by Georg Wenker in the 19th century, this survey asked schoolteachers in 40,000 locations across Germany to translate a set of 40 sentences into the local dialect. By 1887, 44,000 responses had been collected from Germany, plus several thousand more from German-speaking areas in neighboring countries. Today, collecting dialect data with surveys has become still easier, thanks to electronic communication. For instance, a dialect survey can be posted on a website and responses to it solicited through personal contact, email, institutional memberships, advertisements, or social networking sites. By 2002, Bert Vaux had collected a nationwide sample of over 30,000 responses to his web-based Dialect Survey, which comprises 122 questions about lexical and phonological variants of American English (Vaux, n.d.).

By contrast, even the largest dialectological samples collected by means of in-person interviews have rarely amounted to more than 1,000 respondents. Edmond Edmont, interviewer for Jules Gilliéron's *Atlas linguistique de la France*, managed to interview about 700 individuals by riding around France on his bicycle (Gilliéron & Edmont, 1902–1920). A century later, Labov, Ash and Boberg (2006), substituting the telephone for a bicycle, interviewed a similar number for their *Atlas of North American English*. Its predecessor, the *Linguistic Atlas of New England*, was based on interviews with just over 400 people (Kurath, Hanley, Bloch, & Loman, 1939–1943). Since the development of modern sociolinguistics in the 1960s, studies of single speech communities have typically involved much smaller samples. For example, Labov (1972) reports on interviews with 69 islanders in his study of Martha's Vineyard, 264 employees in his study of New York City department stores, and 70 residents in his study of New York City's Lower East Side; other studies have recruited as few as two dozen participants.

The comparatively gigantic samples achieved with written questionnaires depend on the idea that thousands of respondents can be filling out the survey simultaneously, without the direct involvement of the investigator. By contrast, one-on-one, face-to-face interviews, the gold standard of sociolinguistic research, require a comparatively large investment of time and effort on the part of the investigator or a team of assistants: an hour or two for the interview alone, plus the time it takes to locate willing participants, schedule and travel to and from the interview, and process the resulting recording. By this method, achieving a sample of several hundred responses, let alone several thousand, is often impossible, or requires many years of continuous work. If we assume a team of three interviewers working 40-hour weeks for 50 weeks per year, and a conservative estimate of three hours per interview (including travel time), it would take over two years to interview the 14,000 participants in the *Survey of Canadian English*; if the interviewers were paid $20 per hour, the cost would be close to a million dollars, not including travel expenses. Few linguists have such resources at their disposal. At a more conservative estimate of one interview per day, given the time required for traveling between interview sites dispersed over a broad survey area, the project would take close to 20 years to complete: data from the first interviews would be obsolescent before the last interviews were completed and would not be strictly comparable. Using written questionnaires, by contrast, the *Survey of Canadian English* took less than a year, from circulation of the questionnaires to regional directors in the fall of 1971 to implementation, analysis, and even publication of the results in 1972 (Scargill & Warkentyne, 1972).

There is, of course, an important advantage to collecting a large quantity of data: to some extent, assuming data of reasonable quality, confidence in the results of an investigation increases in direct proportion to the size of the sample. A large sample size permits the investigator to draw quantitatively robust conclusions about regional or social patterns in the data, which can be subjected to rigorous statistical analysis and generalized to the population as a whole. The inability to draw equivalently reliable conclusions is often a serious drawback of smaller-scale studies based on in-person interviews. While many of the smaller sociolinguistic samples mentioned above were entirely adequate for their intended purposes, other small samples, such as those assembled for many student research projects, do not reach a similar level of adequacy, thereby limiting the conclusions their authors can draw.

Making possible the collection of a large quantity of data in a short span of time is not the only advantage of written questionnaires. Another is physical remoteness: the investigator can be thousands of miles away from the participant, cutting down on travel expenses, which makes surveys ideal for studies of regional variables over wide areas. Modern technology, from the telephone to the internet, has reduced this advantage, since samples of speech can now be remotely recorded as well, but if interviews need to be done face to face, the practical sample area is comparatively constrained.

Another advantage of surveys is inter-participant comparability: the use of a written questionnaire ensures that each participant responds to identical stimuli, thereby eliminating the possibility that inter-subject differences are influenced by the circumstances or techniques of data collection. Still another advantage is ease of analysis: responses to a written survey are generally easy to classify and enter into a spreadsheet for quantitative analysis, especially if the questions involve selection from a list of possible answers, rather than open-ended questions, as discussed below. Hundreds of response forms can be processed in a single day by a properly trained person, while multiple-choice questionnaires can be machine-read; some internet-based survey applications automatically tabulate and report results whenever they are needed. This makes the data from each respondent uniformly and immediately accessible. By contrast, a recording of natural speech can take hours or days to analyze, depending on the type of analysis, and may contain little or no data at all on the variable under study, if it occurs with varying frequency.

The Disadvantage of Surveys: Quality

Despite all of the advantages just discussed, many sociolinguists consider written surveys to be of little use in the investigation of sociolinguistic variables because of a methodological problem that Labov (1972) called the "observer's paradox": "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation" (p. 209). Labov asserts that "the most systematic data for [the] analysis of linguistic structure" occur in the "vernacular": the casual, everyday style of speech that people adopt when they are not aware of observation. More formal or self-conscious styles, in which people pay more attention to the way they are talking than to what they are talking about, show "irregular phonological and grammatical patterns, with a great deal of 'hypercorrection'" (p. 208); these data are of interest mainly for the information they provide on the social evaluation of language in the community. Consequently, Labov designed his sociolinguistic interviews around an experimental manipulation of attention to speech, ranging from maximal attention produced by direct questions about language to the minimal level apparent in narratives of personal experience, in which participants told stories about dramatic events in their lives. He did not use written questionnaires in his interviews, but these tend to produce a level of self-monitoring roughly equivalent to his most formal style, in which he asked participants to compare the sounds of individual words and report whether they sounded the same or different.

In Labov's view, direct questions about language, whether fill-in-the-blank exercises on dialect surveys or the grammaticality judgments elicited by syntacticians, tell us about people's opinions on language, not about language itself. They can reveal whether a particular linguistic form or feature is negatively or positively evaluated in a community, but not whether, or how often, it is used in particular social or linguistic contexts or by particular groups of people. On the contrary, people's linguistic intuitions are sometimes a very poor indication of how they actually speak. In his study of New York City's Lower East Side, Labov found that "speakers who use the highest degree of a stigmatized feature in their own natural speech show the greatest tendency to stigmatize others for their use of this form" (1972, p. 311). It is therefore possible not only that survey respondents will underreport their usage of socially disfavored forms, but that those participants who reject a form in their responses to direct questions may actually be the ones who use it, while those who do not use it themselves may find it more acceptable. To get around this problem and study language itself, we need to observe language in use, by ordinary people on ordinary occasions, as an instrument of oral communication in the community.

There is little doubt that Labov's reservations about the scientific value of metalinguistic conversations are largely valid. Yet the danger of distortion caused by self-conscious reference to community norms is surely correlated with the degree of social evaluation attached to a variable. Some variables, such as those involving non-standard grammatical forms or stereotyped, obsolescent words or pronunciations, are no doubt extremely socially sensitive, so that asking about them directly is of little use in establishing their real frequency. Answers will reflect the degree to which respondents wish to associate themselves with the groups that are perceived to use the features, rather than genuine levels of usage: depending on the social image they wish to project, some respondents may exaggerate their use of stigmatized forms, while others will exaggerate their use of standard forms. Not all variables, however, involve this level of social evaluation. Many, including Labov's "indicators," operate below the level of social awareness, noticeable only to linguistically trained observers. Others, such as many regional lexical variables and even some variant pronunciations, may be subject to public notice but not to social evaluation: rather than being perceived as "right" and "wrong," or more or less "educated" or "standard," their variants are associated with groups that are not related to one another in a clear social hierarchy, or they occur in apparently free variation without clear group associations. If it can be asserted with reasonable confidence that a survey is investigating this kind of variable, then the disadvantages of the observer's paradox may be diminished, in some cases to the point where they are balanced by the advantages of more data at lower cost.

Even where the observer's paradox remains an important concern, survey data are by no means completely useless. As long as the effect of direct observation is kept in mind, and survey data are not treated as equivalent to data extracted from actual speech, questionnaire responses can indicate social or regional patterns in the evaluation of variables, including evidence of changes in progress (e.g., Chambers, 1998a). A study that replicates an older survey, for example, cannot necessarily draw firm conclusions about changes in the real frequency of variants, but it can infer changes in the evaluation of those variants, which is also an important aspect of studying language in its social context: the attitudinal trend is itself an interesting subject of study. Moreover, the fact that language surveys are a kind of opinion poll can be exploited rather than compensated for: they can be used to investigate social evaluation of language in an overt way by asking respondents not which forms they would use themselves, but what they think of particular forms or of the people who use them.

Surveys and the Study of Canadian English

Surveys have played an unusually important role in the study of Canadian English. They were first used in the 1950s, a decade before Labov raised concerns about the observer's paradox, but have continued to be a standard method of Canadian English dialectology up to the present day. The Canadian dialect survey tradition began with Avis's study of speech differences along the Ontario–United States border (1954; 1955; 1956) and with a smaller pilot study of Alberta speech by Scargill (1954). Avis gives few methodological details in his reports but reveals in a footnote (1956, p. 56, fn. 1) that his data come principally from two questionnaires, circulated at Queen's University and the Royal Military College (both in Kingston, Ontario). His surveys examine questions of vocabulary, grammar, and pronunciation, as reflected in his sequence of published reports; the number of respondents varies by question but ranges between 85 and 165 (1954, p. 18).

Avis himself labeled his work a "pilot study" (1955, p. 14), a status it does indeed hold in relation to the vastly larger undertaking of Scargill and Warkentyne in 1972, mentioned earlier. The *Survey of Canadian English*, with 14,228 respondents from coast to coast and with data tabulated by province, age group, and sex, is the apex of the questionnaire approach to studying English in Canada. Developed by the Canadian Council of Teachers of English, it was directed by Scargill at the University of Victoria and implemented by regional directors at universities across Canada, responsible for liaisons with local schools. Three copies of the survey were given to students in grade 9 classes in each region: one for the student and two for the student's parents. The completed surveys were returned to Victoria for computerized analysis (by IBM Canada). The main thrust of the investigation was a comparison of grade 9 students with their parents; region and sex were secondary concerns, while other factors, such as education and urban–rural distinctions, were not addressed. The questionnaire included 104 linguistic items presented in the form of multiple choices, ranging across the categories of pronunciation, grammatical usage, vocabulary, and spelling. The resulting data are presented in full in Scargill and Warkentyne (1972) and in summary by Warkentyne (1973). Among many findings, Warkentyne highlights generational differences in responses to several phonological variables: the students were more likely than their parents to say they pronounced *new* without a palatal glide (as [nu]); to admit to pronouncing *butter* with a flapped /t/ (as *budder*); and to prefer American variants of words like *progress*, *lever*, *lieutenant*, and *either* (1973, p. 195). Though the main publication of the survey data does not divide the parents by education level, Warkentyne reports that where variables involved a choice between American and British forms, frequency of American responses was inversely correlated with formal education (1973, p. 198). These data form a valuable record that helps to document the social mechanism of a major shift in Canadian English over the late 20th century, in at least some respects, from a pro-British to a more pro-American orientation.

Canada was not unaffected by the development of sociolinguistics south of the border: during the later 1970s and 1980s, the attention of many Canadian researchers interested in language variation turned to sociolinguistic studies of urban communities, like those carried out in American and British cities, involving sociolinguistic interviews. Questionnaire research continued too, however, and by the 1990s was producing yet more results. Nylvek (1992; 1993) reports on a survey of Saskatchewan English, while Chambers (1994) took up the tradition of the *Survey of Canadian English* by initiating a new trans-Canada dialect questionnaire, which he labeled *dialect topography* (DT). Chambers himself undertook the survey of his own area, the "Golden Horseshoe" region around the western end of Lake Ontario, including Greater Toronto, and then, like Scargill, recruited collaborators at universities across Canada to disseminate his questionnaire in other regions. These regions have not yet extended across the country, but data for several areas, including the Golden Horseshoe, Montreal, the Ottawa Valley, Quebec City, Greater Vancouver, Quebec's Eastern Townships, and New Brunswick, as well as some adjacent regions of the United States, are now available on the web.

Like the *Survey of Canadian English*, some of whose questions it reprised, the DT questionnaire is a self-administered survey covering a wide range of phonological, morphosyntactic, lexical, and usage variables. Chambers (1994), the first of many publications presenting DT results, discusses the methodological principles that guided the project. It was designed to be more representative than traditional dialect geography, using random, sociolinguistic sampling rather than concentrating on the minority of the population (especially non-mobile, older rural men) that could be relied upon to produce the most conservative and consistent examples of traditional local speech. It also aimed to be more time-effective, decreasing the interval of time between data collection, analysis, and publication. To accomplish these objectives, Chambers exploited several contemporary conditions of late-20th-century Canadian society: mass literacy, making written questionnaires appropriate for almost all of the teenage and adult population; institutionalization, being the concentration of potential respondents in institutional settings such as schools, workplaces, or retirement homes; communication networks, through which he could communicate with local questionnaire distributors, and respondents could return their questionnaires directly to the project office in a postage-paid envelope, providing a degree of privacy; and computerization, by which means the data could be rapidly tabulated, analyzed, mapped, and reported. While few of these methodological approaches were truly new, most having been employed two decades earlier for the *Survey of Canadian English*, they were highly effective in producing a succession of articles presenting new research based on the DT data, such as Boberg (2004a; 2004b), Burnett (2006), Chambers (1995; 1998a; 1998b; 2000), and Chambers and Heisler (1999).

I initiated yet another national survey of Canadian English, focused entirely on lexical variables, at McGill University in 1999. Called the *North American Regional Vocabulary Survey* (NARVS), its sample area includes the entire United States as well as every region of Canada, with respondents from a wide range of ages and social backgrounds. NARVS began as a questionnaire circulated by students in an undergraduate sociolinguistics class at McGill to members of their own personal networks, but when the results of this initial phase were featured in radio, television, and newspaper reports, people across Canada contacted me asking if they could also participate, so that several thousand more responses were collected from people with no connection to McGill or its students. Finally, a web-based version of the questionnaire collected yet more responses from an even wider sample of the North American population. From a total of around 6,000 responses, a smaller set of 2,400, from respondents who could be confidently associated with particular regions of Canada or the United States (that is, who had spent their childhoods in a single region and still lived in that region at the time of the survey), was retained for regional and apparent-time analysis (Boberg, 2005; 2010). These data revealed the unique status of Quebec as the home of a highly distinctive variety of English, marked especially by the effects of its intensive contact with French, but also supported the division of eastern from western Canada using such variables as whether a multistory building for parking cars is called a *parkade* (West) or a *parking garage* (East); whether athletic shoes worn as casual footwear are called *runners* (West), *running shoes* (Ontario), or *sneakers* (Maritimes); or whether a small house in the countryside for weekend retreats during the summer months is called a *cabin* (West) or a *cottage* (East). It would have been comparatively inefficient to collect such data by means of face-to-face sociolinguistic interviews.

Methodological Considerations

The sociolinguist who wishes to collect data using a written questionnaire faces several choices in how to design the survey so as to make it maximally effective (see Cassidy, 1953, for a general discussion of questionnaire design and implementation). One basic issue concerns the types of variable that can be examined using written questionnaires. Perhaps the most important type of variation in terms of regional and social identity is phonetic, because of the relatively high frequency of many phonetic variables in spontaneous speech, yet phonetic variation cannot be investigated with written questionnaires since it requires a specialized set of symbols, unknown to the general public, for its transcription. Even if people without an education in linguistics can accurately perceive phonetic differences, they have no way of communicating this perception in writing, either by producing their own descriptions or by selecting from a list of alternative forms. Vowel shifts and variables like Canadian Raising, then, are beyond the reach of surveys, as Avis admits in his report on phonological variables (1956, p. 42). Publicly accessible variation begins at the phonemic level, which is represented in conventional orthography: written surveys can investigate mergers (whether two phonemes, or the words they occur in, sound the same) and phonemic incidence (which phonemes occur in which words). They are also, of course, appropriate instruments for asking people about morphological, syntactic, semantic, and lexical variants.

Phonemic inventory – that is, mergers and splits – can be investigated simply by presenting respondents with minimal pairs (e.g., *cot* and *caught*, or *shutter* and *shudder*) or potential rhymes (e.g., *father* and *bother*, or *hand* and *command*) and asking them to indicate whether the words sound the same or different, or rhyme, perhaps with a third, intermediate choice as well. The effect of spelling, which may exaggerate the frequency of "different" responses, is obviously a potential problem of this method, but respondents can be encouraged to say the words aloud or to imagine hearing someone say them on the telephone to get away from an overreliance on written forms. In a language like English, where the correspondence of spelling to sound is variable, phonemic incidence can be investigated by asking people to match the pronunciation of a variable word to a pair or set of words whose pronunciations are invariant, indicating which other word it sounds like or rhymes with. For example, Scargill and Warkentyne (1972, p. 57) report on whether *greasy* rhymes with *easy* or *fleecy*, while Chambers (1994, p. 46) discusses whether *shone*, the past tense of *shine*, rhymes with the man's name *John* or the woman's name *Joan*. Where rhymes are not practical, key words can be used: Scargill and Warkentyne (1972, p. 63) ask whether the <ei> of *either* is pronounced like the <i> of *bide* or the <ee> of *beet*, and one could investigate variation in the nativization of a large set of foreign (a) words, as reported in Boberg (2010, p. 140), by instructing respondents to classify words like *pasta*, *plaza*, *Iraq*, and *soprano* according to whether they contain the <a>

sound of *cat* or the <ah> sound of *father*. A final phonological variable that can be included on written surveys is stress, which can be indicated with capital or bold letters, as in *RE-search* or *FI-nance*, with initial stress, vs. *re-SEARCH* or *fi-NANCE*, with final stress.

While morphological and syntactic variants, such as alternative past-tense forms or relative clause markers, are not difficult to represent on written surveys, they are often subject to overt social evaluation, which makes them less appropriate topics for direct inquiry, as was discussed earlier. Lexical variation, by contrast, is often less affected by linguistic ideology and provides the most obvious application of written questionnaires: different words for the same thing, or different meanings of the same word. A general issue in this type of investigation is the suggestion of answers. Giving respondents a set of alternative forms to choose from and asking them to indicate which they would most often use in their daily speech by ticking a box or circling a word is a convenient way of limiting the range of responses to a set of pre-established variants. It also shortens both response time (compared to respondents having to write out responses) and analysis time (compared to analysts having to interpret free-form, handwritten answers). It does tend to discourage the discovery of new variants previously unknown to the investigator, but this effect can be mitigated, without losing the advantages of a response list, by encouraging respondents who do not find their preferred form on the list to write it in.

It is more difficult, however, to overcome other effects of suggestion on variant choice: unless instructed specifically to choose a single response, which runs the risk of obscuring real intra-speaker variation, respondents presented with a list may choose responses they would not normally produce simply because they see them in the list or because they have heard them from others in the community. Even the order in which the forms appear in the list may have an effect. This issue can be partly addressed by maintaining a constant order across questions, such as alphabetical order, and emphasizing this order to respondents, so that they deliberately disregard it in their choices. The opposite approach to providing a list of alternative forms for respondents to choose from is presenting them with a blank in which they are instructed to write the form they would normally use. This technique overcomes the problems of suggestion but presents its own difficulties. Answers may be illegible or confusing, as when it is not clear whether a respondent is indicating two choices of equal status or an order of precedence, or whether a minor deviation from a common response, such as a difference of one or two letters, represents carelessness, misspelling, or legitimate variation. Fill-in-the-blank questions may also elicit an overwhelming variety of minority responses, some chosen by only one or two respondents, as in the extraordinary variety of terms for a certain schoolyard prank reported by Chambers (1994, p. 53). Beyond the fact of its existence, this diversity of minor variants is often of little scientific interest and will normally end up being relegated to an "other" category that can be concisely reported along with the smaller set of responses that account for larger proportions of the sample. This takes time, however, and analysts may have difficulty determining what counts as a sub-variant of a larger response category for purposes of regional or social analysis and what should be considered a distinct response type.

Conclusion

Used responsibly, surveys can be a powerful tool in the investigation of language variation and change, especially at the phonological and lexical levels, provided that questionnaire responses are not treated uncritically as equivalent to data obtained from spontaneous speech. The convenience, low cost, and inter-subject uniformity of surveys, together with the large quantity of immediately accessible data they make it possible to collect, are all factors that help to balance the major disadvantage of asking people to report on their own linguistic behavior, rather than observing it directly. Ideally, surveys should be used as a complement to data from the observation of language in use, rather than a replacement for natural speech data, so that the weaknesses of one method are alleviated by the strengths of the other.

References

Avis, W. S. (1954). Speech differences along the Ontario–United States border I: Vocabulary. *Journal of the Canadian Linguistic Association, 1*(1), 13–18.

Avis, W. S. (1955). Speech differences along the Ontario–United States border II: Grammar and syntax. *Journal of the Canadian Linguistic Association, 1*(1), 14–19.

Avis, W. S. (1956). Speech differences along the Ontario–United States border III: Pronunciation. *Journal of the Canadian Linguistic Association, 2*(2), 41–59.

Boberg, C. (2004a). The dialect topography of Montreal. *English World-Wide, 25*(2), 171–198.

Boberg, C. (2004b). Real and apparent time in language change: Late adoption of changes in Montreal English. *American Speech, 79*(4), 250–269.

Boberg, C. (2005). The North American regional vocabulary survey: New variables and methods in the study of North American English. *American Speech, 80*(1), 22–60.

Boberg, C. (2010). *The English language in Canada: Status, history and comparative analysis*. Cambridge: Cambridge University Press.

Burnett, W. (2006). Linguistic resistance on the Maine–New Brunswick border. *Canadian Journal of Linguistics, 51*(2–3), 161–176.

Cassidy, F. G. (1953). *A method for collecting dialect.* Gainesville, FL: American Dialect Society.

Chambers, J. K. (1994). An introduction to dialect topography. *English World-Wide, 15*(1), 35–53.

Chambers, J. K. (1995). The Canada–US border as a vanishing isogloss: The evidence of Chesterfield. *Journal of English Linguistics, 23*(1–2), 155–166.

Chambers, J. K. (1998a). Social embedding of changes in progress. *Journal of English Linguistics, 26*(1), 5–36.

Chambers, J. K. (1998b). Inferring dialect from a postal questionnaire. *Journal of English Linguistics, 26*(3), 222–246.

Chambers, J. K. (2000). Region and language variation. *English World-Wide, 21*(2), 169–199.

Chambers, J. K., & Heisler, T. (1999). Dialect topography of Québec City English. *Canadian Journal of Linguistics, 44*(1), 23–48.

Gilliéron, J., & Edmont, E. (1902–1920). *Atlas linguistique de la France*. Paris: Champion.

Kurath, H., Hanley, M., Bloch, B., & Loman, G. S., Jr. (1939–1943). *Linguistic atlas of New England*. Providence, RI: Brown University Press.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.

Nylvek, J. A. (1992). Is Canadian English in Saskatchewan becoming more American? *American Speech, 67*(3), 268–278.

Nylvek, J. A. (1993). A sociolinguistic analysis of Canadian English in Saskatchewan: A look at urban versus rural speakers. In S. Clarke (Ed.), *Focus on Canada* (pp. 201–228). Amsterdam: John Benjamins.

Scargill, M. H. (1954). A pilot study of Alberta speech: Vocabulary. *Journal of the Canadian Linguistic Association, 1*(1), 21–22.

Scargill, M. H., & Warkentyne, H. J. (1972). The Survey of Canadian English: A report. *English Quarterly, 5*(3), 47–104.

Vaux, B. (n.d.). Dialect survey. Retrieved from http://dialect.redlog.net/index.html

Warkentyne, H. J. (1973). Contemporary Canadian English: A report of the survey of Canadian English. *American Speech, 46*(3–4), 193–199.

Wenker, G., Wrede, F., Mitzka, W., & Martin, B. (1927–1956). *Der Sprachatlas des deutschen Reichs*. Marburg: Elwert.

Vignette 8a
Language Attitude Surveys

Speaker Evaluation Studies

Kathryn Campbell-Kibler

In addition to understanding how people use language, sociolinguists often want to understand what people think about the language they use or that other people use. People's beliefs and feelings are related to their linguistic behavior, and feelings about language forms impact people who use those forms. Beliefs and feelings are also interesting in themselves and have practical implications, for example in language policy and planning.

When we study beliefs and feelings about language, we are studying language attitudes and language ideologies. We use a variety of methods to learn about them, including interviews, ethnographic observation, online data mining, surveys, and experiments. Many of these techniques are discussed elsewhere in this book; in this vignette, I focus on a particular technique for learning about language attitudes, called speaker evaluation studies (Giles & Billings, 2004).

In speaker evaluation studies, recordings are played to listeners, who then share opinions about the voices they heard. The recordings differ in specific ways, for example one in Spanish and one in English, or in two different regional accents. With this setup, listener reactions can (hopefully) be taken to indicate something about how the listeners view the language forms in question. Although speaker evaluation studies are not the only way to learn about people's language attitudes, they are a popular one, and there are a number of tricks to doing them well. Having run a few of these myself, I will walk through a few key issues to consider in planning an experiment of this kind.

Stimuli Are the Heart of Your Experiment

All of the conclusions you draw from your experiment are based on reactions to the actual stimuli to which your participants are exposed, not to whatever categories they are intended to represent. This means that the choices you make in selecting or creating your stimuli are the most important choices in your experimental design. Four important considerations are reviewed here.

1. *Creating guises*. In verbal guise studies (Cooper, 1974; 1975), reactions to people who speak different languages or speak with different accents are compared. This technique is handy if you are comparing a large number of varieties, or varieties that are not spoken in the same areas. Research using this technique has been used to find out that teens in Denmark think more highly of other teens

who use “Low Copenhagen” linguistic features, even though when asked directly they say that “High Copenhagen” is a better way to speak (Kristiansen & Jørgensen, 2005).

When using different speakers, you run the risk that the speakers are driving the results. If you choose more friendly or articulate speakers for one variety, that characteristic may come through in voice quality, prosody, or other characteristics irrelevant to your study. You may think that everyone in your study prefers that variety, when it is really the speaker(s) they like. To counteract that danger, Lambert, Hodgson, Gardner, and Fillenbaum (1960) developed the matched guise technique in which a single speaker produces both (or all) guises, unbeknownst to the listeners. This is not a perfect solution, since people may have different “personalities” across their different languages or accents, but it may reduce the variability.

Recently, technology such as the free software Praat (Boersma & Weenink, 2008) has allowed us to directly manipulate the acoustic stream, changing formant structures (Plichta & Preston, 2005) or splicing acoustic material (Campbell-Kibler, 2008; Labov et al., 2006). These techniques can yield stimuli that differ only in the precise linguistic characteristics we are interested in. While this control is very useful, such manipulation can risk creating unnatural stimuli that either strike the ear as audibly odd or, perhaps more dangerously, contain subtle anomalies that may lead listeners astray. Whether the benefits or drawbacks are greater depends on the project in question.

2. *Read, acted or spontaneous?* Most studies prefer speech that is read aloud, because of the control it gives over the word, the content, and the linguistic environments of variables. Some researchers (including me) think that using spontaneous speech where possible makes for a more natural evaluation task. As a compromise, sometimes it is possible to prompt your speakers to tell a well-known story (e.g., a fairy tale or legend) to control content while collecting spontaneous speech. It is also sometimes possible to give speakers short passages and encourage them to memorize the material, so that they are reciting rather than reading it.

3. *Message content*. Whether linguists like it or not, what we think of someone depends a lot on what they say, not just how they say it. Sometimes it’s a simple case of evaluating what someone says directly. In one of my stimuli, the speaker was often described as lazy because he’s talking about how much work it is to attend a movie. People thought he was lazy because he was saying things that sounded lazy! Sometimes it can be more complex, when content interacts with linguistic forms. Another speaker in that same study talks about other people in an ambiguous way; from the clips, you can’t tell whether she likes them or not. When she uses forms like *workin’* instead of *working*, some listeners think she is being condescending while others think she’s compassionate (Campbell-Kibler, 2008). The content is interacting with the language to produce different impressions. These kinds of interactions can be much harder to anticipate and prepare for.

How to solve this problem? One important thing to note is that there’s no such thing as socially neutral content. Even when content doesn’t obviously require a particular interpretation or identity, small cues are always going to affect how listeners perceive a speaker. Instead of trying to find neutral content, it is more useful to think about the potential effect of the particular content you have, pilot your stimuli, and, as noted below, have more than one example.

4. *Have more than one example*. The best way to avoid problems associated with message content and irrelevant (to you) speaker characteristics is to make sure that your study design doesn't rest entirely on the quirks of a single recording. Have more than one sample representing each category you're studying. For a study that compared *-in* and *-ing* guises in speakers of different genders and regions, I made sure I had two speakers of each type (e.g., male Californians) and four example utterances from each speaker. If possible, more than two speakers would be even better.

Tasks and Context

Once you have stimuli that represent the language forms of interest, you need a set of participants to respond to them and something for those participants to do that will reflect their impressions of the stimuli. This section touches on three important decisions to be made in this part of the process.

1. *Choosing tasks*. The most common task for speaker evaluation studies is rating the speakers on a set of qualities (Zahn & Hopper, 1985), but this is not the only thing you can have people do. Some of the most interesting studies have used real tasks such as choosing to fill out a questionnaire at a theater (Bourhis & Giles, 1976; Kristiansen & Giles, 1992) or choosing whether to return a letter or email that has been misaddressed (Bushman & Bonacci, 2004; Milgram, 1977). In these highly realistic studies, participants may not even be aware that they have participated in a study, so little does the task impinge on their lives. More explicit evaluation tasks may be made more realistic by providing a framework of evaluation, as when asking a teacher to evaluate schoolchildren (Williams et al., 1976).

2. *Pilot testing*. One handy technique for improving data quality is to pilot-test the tasks and stimuli with participants sampled from the same population as the main study. Pilot testing can involve open-ended questioning – I used focus groups – to get a sense of how your participants perceive the stimuli and what concepts and terms appear naturally as they discuss them. Pilot testing can also be used to check the quality of your stimuli; for example, whether manipulated stimuli sound natural or whether pictures covey the social qualities intended. A related technique called cognitive interviewing (Willis, 2005) involves asking participants to answer the questions then walk you through their reasoning, to make sure your questions are clearly worded.

3. *Choosing words*. Even within the traditional approach of rating along personal dimensions, a task may be more or less natural for a participant. It is often a good idea to conduct pilot studies to ascertain what qualities your intended population are most concerned with and what terms they use to discuss them. It is possible that I may find results when asking pre-teens about the perceived "loquaciousness" of a set of speakers, but it is almost certain those results will be

more difficult to interpret than if I'd piloted and discovered that (1) they didn't know what the word meant, and (2) they don't really care about the quality referred to, so they tend not to evaluate people on it. Such piloting can even drive research questions, as when Giles, Smith, Browne, Whiteman, and Williams (1980) talked to people in the late 1970s and discovered that both men and women reported that one of the first things they assessed about a woman whose acquaintance they were making was whether she was affiliated with the feminist movement.

Speaker evaluation experiments can be a great tool for assessing language attitudes by providing a task that feels natural to participants and, if done well, conceals the specifics of your research question. Careful selection of your stimuli and your tasks make all the difference in building a study that will yield interesting and understandable results.

References

Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer [Computer software]. Available from http://www.praat.org

Bourhis, R. Y., & Giles, H. (1976). The language of cooperation in Wales: A field study. *Language Sciences, 42*, 13–16.

Bushman, B. J., & Bonacci, A. M. (2004). You've got mail: Using e-mail to examine the effect of prejudiced attitudes on discrimination against Arabs. *Journal of Experimental Social Psychology, 40*, 753–759.

Campbell-Kibler, K. (2008). I'll be the judge of that: Diversity in social perceptions of (ING). *Language in Society, 37*(5), 637–659.

Cooper, R. L. (1974). Language attitudes I. *International Journal of the Sociology of Language, 3*.

Cooper, R. L. (1975). Introduction to language attitudes II. *International Journal of the Sociology of Language, 6*, 5–9.

Giles, H., & Billings, A. C. (2004). Assessing language attitudes: Speaker evaluation studies. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 187–209). Malden, MA: Blackwell.

Giles, H., Smith, P., Browne, C., Whiteman, S., & Williams, J. (1980). Women's speech: The voice of feminism. In S. McConnell-Ginet, R. Borker, & N. Furman (Eds.), *Women and language in literature and society* (pp. 150–156). New York: Praeger.

Kristiansen, T., & Giles, H. (1992). Compliance gaining as a function of accent: Public requests in varieties of Danish. *International Journal of Applied Linguistics, 2*(1), 17–35.

Kristiansen, T., & Jørgensen, J. N. (2005). Subjective factors in dialect convergence and divergence. In P. Auer, F. Hinskens, & P. Kerswill (Eds.), *Dialect change: Convergence and divergence in European languages* (pp. 287–302). Cambridge: Cambridge University Press.

Labov, W., Ash, S., Baranowski, M., Nagy, N., Ravindranath, M., & Weldon, T. (2006). Listeners' sensitivity to the frequency of sociolinguistic variables. *Penn Working Papers in Linguistics, 12*(2), 105–129.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology, 60*(1), 44–51.

Milgram, S. (1977). *The individual in a social world*. New York: McGraw-Hill.

Plichta, B., & Preston, D. R. (2005). The /ay/s have it: The perception of /ay/ as a North–South stereotype in US English. *Acta Linguistica Hafniensia, 37*, 243–285.

Williams, F., Hewett, N., Hopper, R., Miller, L. M., Naremore, R. C., & Whitehead, J. L. (1976). *Explorations of the linguistic attitudes of teachers*. Rowley, MA: Newbury House.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Zahn, C. J., & Hopper, R. (1985). Measuring language attitudes: The speech evaluation instrument. *Journal of Language and Social Psychology, 4*(2), 113–123.

Vignette 8b
Cultural Challenges in Online Survey Data Collection

Naomi S. Baron

Online survey data collection brings enormous advantages for doing sociolinguistic research. A well-designed online survey instrument can help ensure that all questions are answered and even provide a first pass at data analysis. Moreover, online survey tools, be they stand-alone products such as SurveyMonkey or surveys embedded in Facebook, facilitate collection of larger and more diverse samples than is often possible working face to face. They also enable researchers to collect data from sites where they are not physically present.

However, the comparative ease of working in cyberspace can lull researchers into lowering their sensitivity to the importance of controlling for variables (particularly when doing cross-cultural research) that might be more obvious were the researcher working *in situ*. In this vignette, I recount personal experiences – and lessons learned – when using an online survey regarding use of and attitudes toward mobile phones by university students in Sweden, the United States, Italy, Japan, and Korea (see Baron, 2010; 2011; Baron & Hård af Segerstad, 2010). In particular, I focus on assumptions I made about one deceptively simple variable, age, and another I already knew to be complex, culture. (It turned out that the age question was itself culturally embedded.)

The Age Question

My initial age-related challenge came in identifying enough subjects in Sweden. Through ads in student newspapers, signs on bulletin boards, and postings on course home pages, I sought students between the ages of 18 and 24 (a typical age range for undergraduates in the United States), but I received relatively few responses. Though Swedish shyness (or reserve) was probably a contributing factor, I eventually learned there were relatively few undergraduates in this age range to be had. While students in much of the world begin university studies immediately upon completing high school, most Swedes do not. Instead, they typically work for several years before resuming their studies. In fact, I learned that the average age of undergraduates at the University of Gothenburg (where I began my research) was 27.

A different sort of obstacle arose with my Korean sample. I had been physically present in Sweden, the United States, Italy, and Japan (to make arrangements for the online research and to conduct focus groups), but not present in Korea.

Through the generosity of a colleague in Korea, the survey was administered on my behalf. The data collection itself went smoothly. The challenge came when I examined the subjects' ages. There was no one aged 18 or 19. Rather, subjects reported ages between 20 and 26 – despite the fact that Koreans (unlike their Swedish counterparts) typically proceed directly from high school to college. Only when I discussed my bewilderment with a Korean student of mine did I learn that there was a different system used in Korea for calculating age.

In the West, when a child is born, he or she is aged zero. Twelve months later, the baby is aged one. In Korea, however, you are one year old at birth. Then on January 1, regardless of the month and day you were born, you become one year older. So, for example, Americans born in October 1988 would report on a survey administered on December 31, 2010, that they were 22 years old. However, Koreans would be 23. If the survey were administered on January 1, 2011, Koreans would report being 24, while Americans would still be 22. (Japan also has a "traditional" Asian system for calculating age, though today Japanese conventionally use the Western system.)

Both my colleague in Korea and I had assumed we knew how to gather data on subjects' age, but we were operating under different cultural frameworks. To salvage the Korean data, I subtracted one year from each student's reported age, while recognizing that comparison with subjects from other countries was not precise. Fortunately, age did not turn out to be a relevant variable in the analysis.

The Cultural Question

In the same study, I came to further appreciate (in one case, too late) the challenges in comparing subjects from different countries and/or cultures. Building upon pioneering work by Edward Hall (1959; 1976) and Geerd Hofstede (1980), social scientists continue to note how difficult it is, methodologically, to conduct sound comparative research. For example, a single culture may span multiple countries (e.g., the Sami in Sweden, Norway, and Finland), and a single country may contain multiple cultures (e.g., Muslims [Shiite and Sunni], Christians, and Kurds in Iraq). Accurate translation of survey instruments and responses is another hurdle. Moreover, the "same" design variable (such as privacy) may have very different meanings in diverse cultures (Livingstone, 2003). Internet researchers have noted the challenges of gathering cross-national data (Guo, Tan, Turner, & Xu, 2007; Kayan, Fussell, & Setlock, 2006; Massey, Hung, Montoya-Weiss, & Ramesh, 2001), as have researchers studying mobile communication (Haddon, 1998; 2005).

Before undertaking my study, I had read some of the relevant literature on doing cross-cultural research. I knew to have fluent bilinguals translate the survey into Swedish, Italian, Japanese, and Korean (and then translate the open-ended responses back into English). I also took colleagues' advice to gather data from universities in two different cities in each country. But then, like many investigators operating on a restricted time schedule (and budget), I chose my research sites largely on the basis of where I had colleagues who could help me locate subjects.

In the case of Japan, I ended up working in Kyoto and Tokyo. As I learned while in Japan, the Kyoto and Tokyo areas belong to different subcultures (Kansai and Kanto, respectively). Kansai culture (Kyoto) tends to be more informal and Kanto (Tokyo) more formal. While etiquette is important across Japan, it is more so in the Kanto region. This cultural distinction helped explain some of my research findings (for example, that subjects in Tokyo complained more often than their Kyoto counterparts about the bad manners of some mobile phone users).

In Italy, I was fortunate to have colleagues teaching in the towns of Pordenone, Udine, and Modena, who graciously encouraged their students to complete the survey. When I compared the Italian data with those from the other four countries, I found a number of distinctions that I dubbed "Italian" (for example, very strong reluctance to use mobile phones when at dinner with family and, compared with Sweden and the United States, some reluctance to use mobiles while walking in public space). But was I justified in making these generalizations?

When the study had been completed, I presented my findings in a number of venues, including at a conference in Seattle. Following my talk there, a professor from Rome approached me to discuss the project. While he found the data interesting, he wondered whether my findings were indeed generalizable to "Italy." Where, he asked, had I collected the Italian data? As I opened my mouth to respond, I suddenly realized his unspoken point: All the data were from northern Italy – nothing from the south. As any Italian (or student of Italy) knows, northern and southern Italy (particularly the farther south you go) have markedly different cultures. Would data from Rome or Naples have yielded different results? I'll never know. But next time, I will know to ask.

In designing surveys of any sort, two vital steps are identifying relevant variables to control for and ensuring that subjects respond to questions with the same assumption structures as the investigator. The importance of both these issues is magnified when subjects are from divergent cultural backgrounds and when the researcher does not see subjects face to face. In my own work, I managed to salvage the integrity of the study by inserting the necessary caveats when publishing my findings. I also learned at least as much about the challenges of online survey design – and about the richness of human culture – as I did about my original research question.

References

Baron, N. S. (2010). Control freaks: How online and mobile communication is reshaping social contact. *Language at Work, 7*. Retrieved from http://www.languageatwork.eu/readarticle.php?article_id=32

Baron, N. S. (2011). Attitudes towards mobile phones: A cross-cultural comparison. In H. Greif, L. Hjorth, A. Lasen, & C. Lobet (Eds.), *Cultures of participation* (pp. 77–94). Frankfurt am Main: Peter Lang.

Baron, N. S., & Hård af Segerstad, Y. (2010). Cross-cultural patterns in mobile-phone use: Public space and reachability in Sweden, the USA and Japan. *New Media and Society, 12*(1), 13–34.

Guo, Z., Tan, F. B., Turner, T. & Xu, H. (2007, September). Messaging media perceptions and preferences: A pilot study in two distinct cultures. Paper presented at International Conference on Wireless Communications, Networking and Mobile Computing (WiCom) (pp. 6725–6728). IEEE.

Haddon, L. (Ed.). (1998). *Communications on the move: The experience of mobile telephony in the 1990s*. (COST248 Report). Farsta: Telia.

Haddon, L. (Ed.). (2005). International collaborative research: Cross-cultural differences and cultures of research. *COST Action 269*. Luxembourg: EU Publications Office.

Hall, E. (1959). *The silent language*. New York: Doubleday.

Hall, E. (1976). *Beyond culture*. New York: Doubleday.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Thousand Oaks, CA: Sage.

Kayan, S., Fussell, S. R., & Setlock, L. D. (2006). Cultural differences in the use of instant messaging in Asia and North America. In *Proceedings of CSCW 2006* (pp. 525–528). New York: ACM Press.

Livingstone, S. (2003). On the challenges of cross-national comparative media research. *European Journal of Communication, 18*(4), 477–500.

Massey, A. P., Hung, Y.-T. C., Montoya-Weiss, M., & Ramesh, V. (2001). When culture and style aren't about clothes: Perceptions of task–technology "fit" in global virtual teams. *Proceedings of the 2001 ACM SIGGROUP Conference on Supporting Group Work, Boulder, CO* (pp. 207–213). New York: ACM Press.

9 Experiments

Cynthia G. Clopper

Experiments, alone or in combination with other methods of data gathering, are growing in popularity among sociolinguists because they can provide different kinds of data that contribute to our understanding of social variation in language use. Carefully designed production experiments allow us to efficiently collect large quantities of speech that directly bear on our research questions. Complementary data from perception experiments provide concrete evidence for how social variation in speech is perceived and interpreted by non-linguists. Although many experiments rely on highly constrained forms of speech, such as read words or sentences, more natural production and perception data can be obtained through the use of interactive tasks that both constrain linguistic content and allow participants to converse more naturally.

Production Experiments in Sociolinguistics

A primary strength of production experiments is their efficiency. In an experiment, the target list of utterances (words, sentences, or constructions) is designed to elicit an answer to the research question and is established by the experimenter before any data are collected. As a result, we can ensure that the phenomenon that we are interested in will be elicited a sufficient number of times from each participant over the course of the experiment. In ethnographic observation or sociolinguistic interviews, however, we may need to record many hours of speech from a single participant before the phenomenon of interest is produced a sufficient number of times for analysis.

This experimental strength is particularly notable when we are studying a relatively rare phenomenon, such as the vowel /oj/ or modal constructions. For example, in two randomly selected interviews from the Buckeye Speech Corpus (Pitt et al., 2007), the vowel /oj/ occurred only 12 times in 43 minutes in one interview and only six times in 69 minutes in the other. Similarly, modals (including *can*, *might*, *would*) occurred only 43 times in the first interview and 46 in the second interview. Whereas these sociolinguistic interviews did not elicit many examples of these phenomena, we can specifically target the linguistic categories that we are interested in when we design an experiment. If we wanted to explore the monophthongization of /oj/ in regional varieties of American English, for example, we could supplement a sociolinguistic interview, in which we might elicit only a small number of /oj/ tokens, with a short word list reading task to ensure that we have enough examples of /oj/ to confidently answer our question. Similarly, if we wanted to study the double modal construction (e.g., *might could*) in American English, we could design a story completion task to elicit a sufficiently large sample of modal constructions from each of our participants.

This data-gathering efficiency applies to virtually all well-designed production experiments and allows us to explore variation in linguistic categories at many levels of linguistic structure, including segments, prosody, and morphosyntax. Segmental variation can be explored through word list, sentence list, and paragraph reading tasks, in which participants are recorded reading aloud a carefully constructed set of materials that manipulates the variables of interest to the researcher. For example, regional vowel variation has been examined in American English (Clopper, Pisoni, & de Jong, 2005), Dutch (Adank, van Hout, & van de Velde, 2007), and Portuguese (Escudero, Boersma, Rauber, & Bion, 2009) using word list reading tasks, either with the words embedded in a simple carrier sentence (Dutch, Portuguese) or not (English). Prosodic variation can also be examined using reading tasks, although longer passages and scripted dialogues are often used to ensure that the intended information structure is conveyed to the participants. For example, Clopper and Smiljanic (2011) examined regional prosodic variation in American English using read paragraphs, and Arvaniti and Garding (2007) and Elordieta and Calleja (2005) used scripted dialogues between the experimenter and each participant to explore prosodic variation in American English and Spanish, respectively. More interactive tasks, such as the map task (Anderson et al., 1991), in which two participants work together to negotiate a route along a map, can also be used to elicit prosodic variation (e.g., Barry, 2007). Morphosyntactic variation cannot easily be examined using reading tasks, because participants' ability to read a given construction is independent of the extent to which they use that construction in non-scripted social interactions. However, with a little creativity on the part of the researcher, partially scripted tasks (e.g., sentence completion, story completion, and responses to questions by the experimenter) and interactive tasks (e.g., the map task) have the potential to provide us with multiple examples of the variable(s) of interest from each participant. For example, Balcetis and Dale (2005) used a picture description task to explore syntactic priming as a function of the social relationship between the participants. Participants were more likely to describe pictures using the same syntactic structure (e.g., passive vs. active voice) as an unrelated prime sentence produced by the experimenter if the experimenter was perceived as "nice" rather than "mean."

A second primary strength of production experiments is the apples-to-apples comparisons that they permit. Many factors contribute to the phonetic, prosodic, and morphosyntactic realization of an utterance, and experiments allow us to control the factors that we are not interested in so that we can focus on the factors that we are interested in. The effects of preceding and following consonantal context on vowel variation have long been recognized in the sociolinguistic community (e.g., Labov, 1972). More recent research has demonstrated that the vowel in the following syllable can also affect the realization of a target vowel (Cole, Linebaugh, Munson, & McMurray, 2010), suggesting that coarticulatory effects extend over multiple segments. Similarly, morphological context is a well-known factor affecting variable processes such as consonant cluster reduction in African American English (Guy, 1980). Finally, semantic and discourse contexts also affect the realization of individual segments. Words in semantically predictable contexts tend to be reduced relative to words in less predictable semantic contexts (Lieberman, 1963), and words that are repeated in a discourse tend to be reduced relative to the first time they are produced in the discourse (Fowler & Housum, 1987). These discourse-level effects of predictability and repetition also interact with prosodic structure. Words referring to new referents in a discourse tend to be prosodically prominent, and prosodically prominent words are hyperarticulated relative to less prominent words (de Jong, 1995). Segmental strengthening is also observed at prosodic boundaries (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996). Thus, contextual effects emerge not just from immediately neighboring segments but also from more distant segments and higher-level linguistic structure, including prosody, morphology, and semantics.

Independent of the discourse context, properties of words themselves also have a substantial effect on their phonetic and syntactic realization. For example, high-frequency words tend to be reduced relative to low-frequency words (Munson & Solomon, 2004), words that are phonologically similar to many other words are hyperarticulated relative to words that are phonologically similar to few other words (Wright, 2004), and content words are more likely to be prosodically prominent than function words (Calhoun, 2010). Similarly, some variation in syntactic constructions, such as the dative alternation in English, can be attributed to individual word biases. For example, whereas the recipient of *bring* is very likely to have been previously mentioned in the discourse context, the recipient of *take* is much more likely to be new in the discourse context (Bresnan, Cueni, Nikitina, & Baayen, 2007). As a result, *take* is more likely to be produced in a dative PP construction (e.g., *take the bag to school*) than *bring*, which is more likely to be produced in a dative NP construction (e.g., *bring me the book*). Finally, the use of specific words or phrases also contributes to syntactic variation. For example, polarity and the selection of a pronoun vs. a full noun phrase affect the realization of *was*/*were* variation in British English (Cheshire & Fox, 2009; Tagliamonte, 1998). Thus, the specific words that a talker uses to convey his or her message will also have a substantial impact on how the variables of interest are realized.

In an experiment, these contextual and lexical effects can be controlled within and across participants, which increases the likelihood that true effects will be observed and decreases the likelihood that spurious results will be interpreted as significant. For example, in an experiment exploring degrees of /u/ fronting by male and female speakers of Southern American English, each participant could be recorded producing the same set of 20 target words in isolation or embedded in a carrier phrase or paragraph. By controlling for consonantal context, lexical variability, and discourse context, we would be more likely to observe a significant difference between genders if one actually existed. If instead we were to extract words containing /u/ from interviews with the same set of participants, we would not be able to control as well for phonological, lexical, and discourse properties in our selection of target words, and the true effects of gender could be obscured. For example, if the tokens produced by the female talkers happened to be in low-frequency words, and the tokens produced by the male talkers happened to be in high-frequency words, the /u/ productions might appear to be equally fronted across genders. That is, a true gender effect could be masked by the frequency effect: the male and female /u/s would appear to be overlapping because the female talkers produced relatively more peripheral (i.e., backed) /u/s in their low-frequency words, and the male talkers produced relatively less peripheral (i.e., fronted) /u/s in their high-frequency words. In addition, a spurious gender effect might be obtained for /u/ raising: the female /u/s would appear to be higher than the male /u/s because the female talkers produced relatively less peripheral (i.e., raised) /u/s in their low-frequency words, and the male talkers produced relatively less peripheral (i.e., lowered) /u/s in their high-frequency words.

In spontaneous and interview speech, the researcher typically has very little control over the many factors that contribute to variation in speech production, and the resulting data are often very noisy. Although our statistical models are improving to allow us to include many different contributing factors in our analyses of production data, we must still identify all of the potentially relevant factors, quantify those factors in an appropriate and meaningful way, and avoid overfitting our data by including too many variables in our analysis. In addition, noisy datasets typically require more data points than clean datasets to observe true results and to avoid spurious results. In production experiments, the researcher has much more control over the materials, and carefully designed experiments can yield valid and highly reliable data. Having comparable data from each participant allows us to be more confident that any differences that we observe across participant groups are due to social factors rather than accidental linguistic factors.

Perception Experiments in Sociolinguistics

The primary strength of perception experiments is that they allow us to explore how the variation that we observe in production is used by non-linguists to interpret the intended message and to identify social characteristics of the talker. These kinds of perceptual judgments are essential for a complete understanding of sociolinguistic variation. One classic example of the important contribution of perceptual judgments to sociolinguistic research is the phenomenon of near-mergers. In a near-merger, language users maintain a phonetic distinction between two phonemes in production but report that minimal pairs containing those two phonemes are the same in a perception task (Labov, Karen, & Miller, 1991). Thus, a near-merger is a perceptual phenomenon that we cannot observe directly from production data.

Direct linguistic judgment tasks, such as grammaticality judgment tasks or the minimal pair test used by Labov et al. (1991), tap participants' explicit knowledge of linguistic structure and allow for conscious reflection by the participants. When combined with an interview in which the participants are encouraged to explain or discuss their linguistic judgments, these tasks can provide very rich data on segmental and morphosyntactic variation. However, because direct tasks may allow for conscious reflection by the participants, the data may be colored by the participants' prescriptive notions of grammaticality or knowledge of orthography. Indirect tasks that focus on processing or interpreting the linguistic content of an utterance may therefore be preferable for exploring what participants actually do, rather than what they believe to be true about their language. For example, tasks that require participants to respond as quickly as possible, such as lexical decision tasks (Floccia, Girard, Goslin, & Konopczynski, 2006) or speeded classification tasks (Clopper & Pate, 2008), can be used to examine the processing difficulties associated with an unfamiliar variant. Slower response times are associated with difficult processing, whereas faster response times are associated with easy processing. Similarly, familiarity or exposure to a particular variant can be examined using semantic priming (Sumner & Samuel, 2009; Warren & Hay, 2006). Familiar phonological variants will effectively prime semantic associates (e.g., *chair* primes *sit*), but less familiar variants will not. This priming effect is realized in perception tasks as faster response times to primed targets than to unprimed targets. These tasks that involve response time data are very sensitive to subtle differences in linguistic processing and therefore have the potential to uncover aspects of sociolinguistic variation that are more difficult to observe in production experiments, such as passive competence in a second dialect, or in linguistic judgment tasks, such as familiarity with a particular variant.

Perception experiments can also be used to determine how the social categories that we identify in our production data map onto the social categories of non-linguists. Although production data inform us about the variation that exists in the world, the researcher is ultimately responsible for carving the data up into relevant social categories for interpretation. Perception experiments can provide essential complementary evidence for the relevance of the social categories for the local community. For example, if neighborhoods within an urban area partially overlap with social class divisions, a perception experiment could help determine which of the two variables is more central to social identity in the local community. We can imagine a study in which participants are asked to categorize talkers with varying social and geographic backgrounds into groups first based on neighborhood and then based on social class. If neighborhood is a more important category for the participants than social class, performance across participants should be more consistent in the neighborhood classification task than in the social class task. Alternatively, if social class is more central for the participants than neighborhood, greater consistency should be observed in the social class task than the neighborhood task. This experimental design would also have the potential to capture the relevant neighborhood or social class distinctions for the participants. If the neighborhoods identified by the researcher were either too broadly or too narrowly defined, the mismatch between the researcher's categories and the participants' categories would be revealed by the pattern of responses.

The perceptual dialectology tasks that Preston (1989) has developed, including map drawing, correctness ratings, and pleasantness ratings, are all motivated by this question of the relationship between sociolinguists' maps of linguistic variation and the cognitive maps of non-linguists. Perceptual dialectology is primarily concerned with participants' beliefs about linguistic variation and therefore does not typically involve explicit perception of a particular stimulus. However, perception experiments can also be used to obtain perceptual judgments in response to variable linguistic stimulus materials. For example, Clopper and Pisoni (2004; 2007) examined how participants identify the regional dialect of unfamiliar talkers using forced-choice categorization tasks (in which listeners assign each talker to one of a researcher-defined set of categories) and free classification tasks (in which listeners group talkers together without any predefined categories). Finally, perception experiments can be used to elicit attitude judgments about talkers based on their speech (see Campbell-Kibler, Vignette 8a). One of the most prevalent paradigms in attitude research is the matched-guise technique (Lambert, Hodgson, Gardner, & Fillenbaum, 1960), which has been used to explore attitudes toward both phonological variation, such as regional and foreign accents (e.g., Giles, 1970), and grammatical variation, such as copula absence in African American English (e.g., Bender, 2005).

Perception experiments also provide opportunities to explore the relationships between linguistic and social categories in speech processing. For example, Clopper and Pisoni (2004) used regression techniques to determine the phonetic properties that contribute to the identification of the regional dialects of unfamiliar talkers, and Strand (1999) observed differences in /s-ʃ/ identification as a function of the gender and gender typicality of the voices. Thus, perception experiments can be used to examine how linguistic variation affects social categorization as well as how social variation affects linguistic categorization.

Potential Limitations of Experiments in Sociolinguistics

The primary limitation that is often identified for experiments for sociolinguistic research is the lack of naturalness that these tasks may involve. To capitalize on the strengths of experimental design and maximize the control over the materials, read speech is often elicited in production experiments and presented to listeners in perception experiments. However, read speech is known to differ from spontaneous speech at all levels of linguistic structure and is therefore not representative of all kinds of speech. In addition, the use of read speech in production experiments requires participants to be literate, which may limit the population of potential participants in an undesirable way. Production experiments involving read speech are also unfeasible for varieties that do not have a relatively standardized written form. The problems with read speech have received particular attention in the prosody literature, and many researchers are developing novel interactive tasks that constrain linguistic content but allow for the collection of relatively natural speech. These tasks include the map task mentioned earlier (Anderson et al., 1991), as well as the diapix picture description task (Van Engen et al., 2010), partially scripted games (Speer, Warren, & Schafer, 2011), and the holiday tree decorating task (Ito & Speer, 2006). Each of these tasks was developed with a somewhat different research goal and is therefore structured somewhat differently. However, these tasks have several properties in common that are central to developing any kind of successful interactive task: two participants interact with each other to achieve a shared goal, and the materials are carefully designed to elicit the target utterances. These approaches have the potential to provide us with controlled production materials in spontaneous, interactive speech.

A second potential limitation of experiments as a method of data gathering in sociolinguistics is that phonetics and psycholinguistics experiments are traditionally run with university students using high-quality sound equipment in quiet locations. However, university students are not representative of the adult population, particularly with respect to literacy, computer skills, and experience with formal testing. Thus, experimental paradigms that are effective with university student populations may not work with other populations that are of interest to sociolinguists. In addition, bringing participants into a laboratory setting at a university is a particular social context that may have substantial effects on how participants perform. Now that high-quality microphones and digital recording equipment are more portable and affordable than ever, it is becoming possible to conduct both production and perception experiments in the field. Another option is a web-based perception experiment, such as Campbell-Kibler's (2007) matched-guise study of the (ING) variable in American English. However, the quality of the audio output that individual participants experience is difficult to control in these studies. Thus, web-based experiments are often more suitable for studies of lexical or syntactic variation, in which fine phonetic detail is not central to the research question, and for experiments involving perceptual dialectology tasks or survey methods that do not require perceptual responses to auditory stimulus materials.

Methodological Considerations in Experimental Design

Regardless of the experimental paradigm that we adopt, we need to carefully select our stimulus materials. In particular, we need to ensure that we have the right kind of linguistic materials (e.g., words, sentences, passages, or dialogues) to address our research question and that we have controlled for as many as possible of the potentially relevant factors (segmental and semantic contexts, frequency, etc.) that may affect the variable(s) of interest. Longer stimulus materials will necessarily require greater care in controlling for these other factors, but there are many ways to control for these sources of variability without creating an unmanageably large experiment. Variables that we are interested in should be balanced (i.e., appear equally often) across experimental conditions, whereas variables that are not central to our research question can be controlled by selecting a single level of the variable for all of our materials (e.g., choosing only high-frequency words) or allowing the materials to vary freely with respect to that variable across experimental conditions (e.g., choosing words of varying frequencies for all conditions).

In addition, we typically do not want participants to know (or be able to figure out) our precise research question, because their behavior may change if they are trying to help us get good data. For example, if we are interested in a potential vowel merger, we may not want to ask our participants to read a list of minimal pairs containing the relevant contrast, because they may try to produce the targets differently simply because they know they are supposed to be different words. Similarly, if we are interested in the prosody of compound nouns, we may not want to label all of the landmarks in a map task with compound nouns, because participants may notice the pattern and produce particularly marked prosody on the targets. One effective way of reducing the possibility that participants will figure out what the experiment is about is to use fillers. Fillers are extra stimulus materials that are unrelated to the target materials and serve to maintain some apparent randomness in the complete set of materials that the participant is exposed to. In our experiment about vowel mergers, the fillers could be words containing vowels that are not involved in the merger. In our experiment about compound nouns, the fillers could be proper nouns or adjective–noun sequences.

Even if participants are not aware of the purpose of the experiment, their experience with some of the stimulus materials may affect how they perform on other materials. For example, if the minimal pairs in our vowel merger experiment are presented one after the other, participants may produce a larger difference between the pairs than if they are separated by several fillers. Presenting materials in a random order can reduce these effects, known as order effects. Randomization can either be performed once, so that each participant is exposed to the materials in the same random order, or separately for each participant. Different randomizations for each participant allow the effects of order to vary randomly across items and participants, and are effective for larger experiments with many participants. For smaller experiments with fewer participants, a single random order may be more appropriate so that order effects are constant across participants and can be included as a factor in the statistical analysis. A second method for controlling for order effects is counterbalancing, in which the order of experimental conditions is balanced across participants. For example, in our compound noun experiment we might have one map with a “white house” landmark and one map with a “White House” landmark so that we can compare the prosody of the adjective–noun sequence (“white house”) to the prosody of the compound noun (“White House”). To counterbalance for order effects, half of our participants should complete the map with “white house” first, and the other half of our participants should complete the map with “White House” first, so that any effects of having already encountered the other form will be balanced across the two participant groups.

Finally, the “observer’s paradox” is a well-known problem for many kinds of data-gathering methods in sociolinguistics (Labov, 1972), including experiments. Accommodation in speech production is well-documented among interlocutors (e.g., Giles, 1973) and has been observed in sociolinguistic interviews (e.g., Rickford & McNair-Knox, 1994) and interactive laboratory tasks (e.g., Pardo, 2006). Hay, Drager, and Warren (2010) have also found effects of experimenter dialect on performance in a speech perception task that was independent of the variation in the stimulus materials. Similarly, Goldinger and Azuma (2003) found that the experimenters' expectations about the outcome of the experiment biased participants to perform in a particular way. Thus, the experimenter plays a crucial role in how participants perform, and it is important to ensure that all experimenters are well trained. If more than one experimenter is involved in a particular experiment, randomization and counterbalancing techniques should be used to ensure that any effects of experimenter bias or social characteristics are evenly distributed across the data.

Conclusion

Experiments allow us to efficiently obtain large amounts of relevant data. Although many experimental paradigms involve less natural speech than other methods of data gathering in sociolinguistics, more interactive tasks can be used to elicit comparable target utterances containing the variable(s) of interest from all participants. Thus, experiments have the potential to complement ethnographic, sociolinguistic interview, and survey methods to provide converging evidence for both descriptive observations and theoretical claims.

References

Adank, P., van Hout, R., & van de Velde, H. (2007). An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America, 121*, 1130–1141.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., & Weinert, R. (1991). The HCRC map task corpus. *Language and Speech, 34*, 351–366.

Arvaniti, A., & Garding, G. (2007). Dialectal variation in the rising accents of American English. In J. Cole & J. I. Hualde (Eds.), *Laboratory phonology 9* (pp. 547–575). Berlin: Mouton de Gruyter.

Balcetis, E. E., & Dale, R. (2005). An exploration of social modulation of syntactic priming. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 184–189). Mahwah, NJ: Lawrence Erlbaum.

Barry, A. S. (2007). The form, function, and distribution of high rising intonation in Southern Californian and Southern British English. (Unpublished doctoral dissertation). University of Sheffield.

Bender, E. M. (2005). On the boundaries of linguistic competence: Matched-guise experiments as evidence of knowledge of grammar. *Lingua, 115*, 1579–1598.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Calhoun, S. (2010). How does informativeness affect prosodic prominence? *Language and Cognitive Processes, 25*, 1099–1140.

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

Cheshire, J., & Fox, S. (2009). *Was/were* variation: A perspective from London. *Language Variation and Change, 21*, 1–38.

Clopper, C. G., & Pate, J. K. (2008). Effects of talker and token variability on perceptual learning of dialect categories. *Proceedings of Meetings on Acoustics, 5*(060002).

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics, 32*, 111–140.

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics, 35*, 421–438.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America, 118*, 1661–1676.

Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics, 39*, 237–245.

Cole, J., Linebaugh, G., Munson, C. M., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics, 38*, 167–184.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America, 97*, 491–504.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics, 24*, 423–444.

Elordieta, G., & Calleja, N. (2005). Microvariation in accentual alignment in Basque Spanish. *Language and Speech, 48*, 397–439.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *Journal of the Acoustical Society of America, 126*, 1379–1393.

Floccia, C., Girard, F., Goslin, J., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1276–1293.

Fowler, C. A., & Housum, J. (1987). Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*, 489–504.

Giles, H. (1970). Evaluative reactions to accents. *Educational Review, 22*, 211–227.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics, 15*, 87–105.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics, 31*, 305–320.

Guy, G. R. (1980). Variation in the group and the individual: The case of final stop deletion. In W. Labov (Ed.), *Locating language in time and space* (pp. 1–36). New York: Academic Press.

Hay, J., Drager, K., & Warren, P. (2010). Short-term exposure to one dialect affects processing of another. *Language and Speech, 53*, 447–471.

Ito, K., & Speer, S. R. (2006). Using interactive tasks to elicit natural dialogue. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurtzky, I. Mleinek, .. J. Schliesser (Eds.), *Methods in empirical prosody research* (pp. 229–257). Berlin: Mouton de Gruyter.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W., Karen, M., & Miller, C. (1991). Near-mergers and the suspension of phonemic contrast. *Language Variation and Change, 3*, 33–74.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology, 60*, 44–51.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*, 172–187.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research, 47*, 1048–1058.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382–2393.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech* (2nd release). Retrieved from http://www.buckeyecorpus.osu.edu

Preston, D. R. (1989). *Perceptual dialectology*. Providence, RI: Foris Publications.

Rickford, J. R., & McNair-Knox, F. C. (1994). Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 235–276). New York: Oxford University Press.

Speer, S. R., Warren, P., & Schafer, A. J. (2011). Situationally independent prosodic phrasing. *Laboratory Phonology, 2*, 35–98.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology, 18*, 86–99.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60*, 487–501.

Tagliamonte, S. (1998). *Was/were* variation across generations: View from the city of York. *Language Variation and Change, 10*, 153–191.

Van Engen, K., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech, 53*, 510–540.

Warren, P., & Hay, J. (2006). Using sound change to explore the mental lexicon. In C. Fletcher-Flinn & G. Haberman (Eds.), *Cognition and language: Perspectives from New Zealand* (pp. 105–125). Bowen Hills, Queensland: Australian Academic Press.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology VI* (pp. 75–86). Cambridge: Cambridge University Press.

This page intentionally left blank

Part III

Working with and Preserving Existing Data

This page intentionally left blank

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.

11 Written Data Sources

Edgar W. Schneider

Investigating sociolinguistic correlations and indexicality essentially builds upon the observation of oral performance, i.e., speech – ideally, as unmonitored as possible. In contrast, writing counts as a cultural artifact; it represents a secondary encoding of speech via letters and transliteration, and it is conventionally constrained by its proximity to standard norms, "proper English." Acquiring spelling conventions happens in formal schooling and requires effort. Knowing how to spell complicated words correctly and being able to express oneself fluently and stylistically adequately in writing count as indicators of education, and these abilities thus constrain the range of sociolinguistic strata in focus here. So at first sight, written sources seem a far cry from sociolinguistic concerns.

Still, written data sources have been appropriate bases for sociolinguistic investigations and will continue to be, for good reasons. This chapter addresses relevant methodological issues and concerns that need to be considered in such an investigation. In surveying these issues, I choose a hands-on approach, starting from simple, practical questions that a researcher may ask her- or himself and moving along subsequent stages in carrying out such a project:

- Why would we want to, or have to, consider written data in the first place?
- What kinds of text sources are available and where may we find them?
- What criteria will have to be applied in selecting specific texts from a universe of sources one may have identified?
- What kinds of procedures are applicable in investigating written data, and in what ways do these differ from and relate to the default situation of studying speech?
- What are the limitations imposed by the nature of the data, and how can they be considered and, hopefully, overcome?

Why? Motivations for Investigating Written Data Sources

Life is rich and variable, and there may be all kinds of reasons for why all kinds of people do things, but, disregarding marginal possibilities, there are essentially two types of motivations for sociolinguists to investigate written data: for want of "anything better" (i.e., unavailability of oral records) or as a goal in its own right.

There are speech communities and sociolinguistically interesting communicative contexts from which oral recordings, assuming that these would be an “ideal” data source, are simply not available and cannot be obtained. However, it is not infrequently the case that written sources, indirect representations of speech, are available from such ecologies and may offer a researcher a “second-best solution” – or, even more, a window into such orally unrecorded linguistic ecologies. It is possible to conceive of such a constellation in a present-day context as well (though this is probably rare; if we really want to know what some sort of speech situation today is like, we can usually pick a recording device and go there); but the typical situation relates to the past, to speech produced at a time from which we have no recordings simply because recording technology was not available then and there. Written data sources are all we have as representations of speech until the 19th and, in some cultures and contexts, deep into the 20th century. In a general sense, this situation applies to the entire discipline of historical linguistics, which has traced language change since the earliest records of a culture on the basis of textual representations that have come down to us. But given that many of these texts, because of the circumstances of their preservation and production, represent “high” styles, and given the traditional esteem for literature, this discipline has tended to disregard or slight a sociolinguistic perspective. The position has changed during the past two decades, however: there is now a subdiscipline of “historical sociolinguistics” that attempts to recover socially conditioned variation of speech in earlier centuries (see Milroy, 1992). The best-known and most sophisticated approach along these lines is the compilation and analysis of a Corpus of Early English Correspondence (CEEC: Nevalainen & Raumolin-Brunberg, 1996). In addition to such approaches, which, despite a sociolinguistic orientation, focus largely on standard (or near-standard) speech, there have also been projects and attempts at unearthing the history of vernacular varieties – which, for obvious reasons, have been less commonly recorded and preserved (see Bailey, 1997; Schneider, 2012). Classic cases in point and the best-known examples are attempts at reconstructing the history of African American Vernacular English (Poplack, 2000; Schneider, 1989; Van Herk & Poplack, 2003).

In addition, writing represents a special form of linguistic performance that is unavoidably shaped by sociolinguistic conditions of its production, so texts of whatever kind also constitute a type of linguistic source that may be of interest in its own right from a sociolinguistic perspective. Linguists may be interested in the properties of specific business text types (e.g., Kretzschmar, Darwin, Brown, Rubin, & Biber, 2004); they may want to study literary dialect and the representation of variability in it (Schneider & Wagner, 2006); or a forensic linguist may be confronted with a piece of written text that plays a role in a lawsuit (see Angermeyer, Vignette 11b). These linguistic questions, obviously, are also sociolinguistic questions, requiring a proper analysis and contextualization of written sources.

What? Text Types

As was implied earlier, many written standard texts reveal little that is of sociolinguistic interest, as most of them represent the standard linguistic norm and thus are not, or hardly at all, indexical of social or contextual parameters. Usually, texts of greatest interest to sociolinguists contain colloquial and vernacular speech forms, and then, in order to understand the significance of the language forms in context, we want to know something about the social parameters that characterize the text producer and the contextual setting of the writing process. A wide range of pertinent contexts is possible and of interest. Written texts may be direct and close records of a speech event, of what a specific individual said at a specific place and point in time, with this utterance having been written down by somebody else, more or less verbatim, for whatever reason. Alternatively, texts of interest may not represent real speech but have been produced as such, representing potential speech, as it were. It is helpful to assess the usefulness of written sources for sociolinguistic purposes by categorizing them into text types that share characteristic contexts and recording conditions. Table 11.1, reproduced from Schneider (2002, p. 73), categorizes and characterizes text types into five broader types that allow us to assess the proximity of a written representation to an "underlying" speech act. Accordingly, it is possible to list a number of text types that a sociolinguist working with written data may want to look at and to consider their characteristics and resulting constraints.

Transcripts of a specific speech event or recording are clearly most suitable for a sociolinguist's purpose – and in fact this is the kind of representation that most sociolinguists who work on oral recordings also use and rely on at certain stages of their work. In real life, however, there are not too many occasions on which such transcripts are produced. Parliamentary debates are transcribed in many countries, but this example illustrates additional constraints that need to be considered, two in particular: First, do the speakers produce vernacular language? Second, is the scribe able and willing to fully record all the non-standard

Table 11.1 Categorization of Text Types According to Their Proximity to Speech

Category	*Reality of Speech Event*	*Speaker Writer Identity*	*Temporal Distance Speech – Record*	*Characteristic Text Types*
Recorded	Real, unique	Different	Immediate	Interview transcripts, trial records
Recalled	Real, unique	Different	Later	Ex-slave narratives
Imagined	Hypothetic, unique	Identical	Immediate	Letters, diaries
Observed	Usu. real, unique	Different	Later	Commentaries
Invented	Hypothetic, unspecified	n.a.	Unspecified	Literary dialect

forms that may occur? Political debates are usually conducted in the standard form of a language – which is why, for example, Hansards (transcripts of parliamentary debates in Britain and many of its former colonies) have not yet attracted sociolinguistic attention. A different situation may occur, and did occur, in court cases, however: defendants were often from the lower classes, and the judicial procedure often required their statements to be recorded verbatim, to be used as evidence in future considerations, and so trial records are among the best historical vernacular sources we may be able to come by. Cases in point are the Salem witchcraft trial records (Rissanen, 1997) or the official records of London's Old Bailey court from earlier centuries (Huber, 2007).

Sometimes speech is recalled and written down some time after the utterance itself, a procedure that may entail lapses of memory and introduce errors. Personal narratives of one's memories fall into this category but tend to be rather standard. A special case, however, are the Works Progress Administration (WPA) ex-slave narratives analyzed in Schneider (1989). These are transcripts from notes, made on location or from memory, of interviews conducted in the 1930s and 1940s with very old African Americans who had seen the days of slavery. Most of them are first-person narratives, and the interviewers were instructed to record the interviewee's words, including non-standard language, as closely as possible. Whether or to what extent they really did is difficult to assess (and it probably varied), and some of the texts may have been edited, so the validity of these texts has been questioned and reassessed – but still, this is the most comprehensive record of earlier African American speech we have to date.

Another source of interest, recording not actual speech but one's imagined performance, as it were, is personal letters and diaries. They are not produced for a public audience, so considerations of linguistic decorum matter to a lesser extent, if at all. Still, in earlier days people who wrote such texts were often from an upper-stratum minority, so the question remains as to what their linguistic output is typical of. Under very special circumstances, however, people who were barely literate were forced to bring statements that were important to them onto paper. For lack of familiarity with writing conventions, such semi-literate writers produced texts that were remarkably vernacular and that may thus be a goldmine for a sociolinguist. Montgomery (1997, p. 229) identifies, categorizes, and describes such contexts, labeling the writers in question "lonelyhearts" (people separated from their loved ones), "desperadoes" (people who were in urgent need of something), and "functionaries" (people who were obliged to report on some state of affairs). Clearly, such texts are relevant for sociolinguistic investigations, so a few projects were set up with the explicit aim of systematically compiling such sources. These include the Southern Plantation Overseers' Corpus (SPOC: Schneider & Montgomery, 2001), the Ottawa Repository of Early African American Correspondence (OREAAC: Van Herk & Poplack, 2003), and the Corpus of Older African American Correspondence (COAAL: Schneider, 2012).

Travelers' reports, a popular genre in earlier centuries, are another possible source: after returning home from faraway lands, travelers wrote books or articles about what they had experienced and observed, including, sometimes, samples and quotations of the "strange" speech forms they had encountered. These sources may provide unique documentation about contexts about which otherwise very little is known, but the reliability of such reports is somewhat questionable (as the travelers may have misunderstood, half-forgotten, or even deliberately distorted relevant details).

Dialect writing, commonly called literary dialect, constitutes a genre in its own right, and again, such writing may be of interest for sociolinguistic purposes. In earlier times, stage characters sometimes used vernacular speech forms for humorous purposes even in serious plays, and in novels, short stories, or other texts with a regional setting the direct speech of characters often attempts to represent a local dialect reliably. It may be possible to correlate social parameters of fictitious characters with their (equally fictitious) speech forms (Schneider & Wagner, 2006). Ellis (1994) is a successful example of an investigation of a regional (in this case, Southern US) dialect on the basis of literary sources, and there are other such studies (e.g., Mille, 1997; Minnick, 2004).

Where From? Locating Possible Sources

Where will a sociolinguist wishing to embark on a project with the goal of investigating written data sources find such data? It varies, but it may be difficult or cumbersome. For example, finding attestations of interesting speech forms in travelers' reports is likely to be time-consuming; just a few islands of interesting forms will be scattered in a sea of standard texts.

Historians may be helpful, as they too are interested in reconstructing things of the past on the basis of textual recordings; their interests are different, but they frequently know where texts of linguistic interest can be found. In fact, many collections edited by historians may contain or even largely consist of material of sociolinguistic interest. For example, there is a set of well-known books on the fate of African Americans after emancipation, or repatriation to Africa, expressed in their own words in the form of (often semi-literate) letters; Kautzsch (2000), for example, analyzed some of them linguistically, and others have been integrated into COAAL. In such cases, it is important to make sure that the texts were reproduced in their original form and not edited (historians, unlike linguists, are interested in content, not linguistic form and detail, and may be tempted to "cleanse," i.e., standardize, texts for easier readability). In order to assess the editorial policy, a linguist will have to consult either an introductory section that contains such methodological information or locate the corresponding manuscript originals, which may be available as facsimile reproductions published in a more recent book, or may be stored in the original archives.

This, of course, is the alternative possibility: to carry out archival searches. There are numerous regional and national archives in the United States and in many other countries, and their voluminous holdings may be a treasure trove for a sociolinguist who is willing to spend the time and energy needed to search them. Catalogs are helpful (and frequently available online), and so are professional archivists (though it may be difficult to explain what a sociolinguist is interested in: being interested not in content but in documents characterized by non-standard linguistic usages is an unusual perspective for most people, to say the least). Another methodological issue in archival work, which can be solved but needs to be considered, is the question of how to get the data and take them home for further analysis. Some archives offer reproduction services but most will not allow users to photocopy their rare and valuable holdings themselves; digital photography may be a solution. A related issue is the purposes for which the researcher is allowed to use the source material. Building a corpus for one's own linguistic analysis is usually not a problem, but obtaining copyright for publishing significant selections from the source data may be difficult to obtain.

Similarly, dialect writing may or may not be easy to find, and a researcher interested in it needs to know where to look – that is, needs some familiarity with the cultural context of a given community or region in which dialect writing is likely to have been produced, preserved, and possibly made publicly accessible. Certainly there are specific domains and publication outlets where searches are promising: products by publishers that specialize in regional literature, certain newspaper columns or cartoons, works by authors who are known for reproducing their characters' speech accurately, and increasingly websites and possibly also blogs employing down-to-earth usage. In other works of literature, reproductions of vernacular speech may be rare and scattered.

Which Ones? Selecting a Working Corpus

Given the difficulty of getting hold of reliable written data, in many cases a sociolinguist will be happy to use as a source of investigation whatever is available from the sociohistorical or cultural context one is interested in; this is simply a consequence of the "bad data" or "insufficient data" problem outlined earlier in the chapter, which tends to characterize this approach. Hence, most of the time there is no need to consider principles of further selection, say, of a sample from a "universe" of available source texts. This problem seems a luxury that many sociolinguists will not be able to afford; the procedure will simply be to "compile and analyze whatever you can get hold of." But there are of course limitations to the corpus size a sociolinguist can investigate (given real-life constraints of time and funding), and it does happen that so much written material is available that there is a need to select. (The WPA ex-slave narratives edited by George Rawick [1972–1979] constitute an example. There are dozens of volumes of published narratives available, including thousands of individual tales, too many to consider in full.) If that is the case, two main principles apply: securing the quality of data in the corpus, and proceeding analogously to a "regular" sociolinguistic project.

"Quality of the data" basically means the accuracy of rendering colloquial speech – that is, the degree of vernacularity of a given source. It is not unusual for a rather large-scale database to consist of texts that are different in character, for example in terms of their proximity to the standard. The WPA ex-slave narrative collection and letter corpora such as COAAL contain some texts that are very close to the standard and show no or very few vernacular forms, for a variety of reasons. It may make sense to exclude such texts from the sample for investigation, as they obviously fail to reflect the style level and type of variability a sociolinguist is interested in (or, conversely, to identify and select the interesting, vernacular ones). One principle to be adopted here is to avoid circularity: it is not acceptable to select texts on the basis of their containing the forms the researcher wishes to investigate. In practice, texts that are fully standard will have to be excluded, but there are likely to be borderline cases with a very small number of non-standard forms. It is recommendable to look for occurrences of specific non-standard forms that tend to occur at a reasonable frequency but that will not be a later target of investigation (see Schneider, 1989, pp. 54–57). In letters, high-frequency phenomena that count as reliable indicators of vernacularity and lower levels of literacy are erratic spelling, capitalization, and punctuation.

Having thus circumscribed the set of acceptable texts from a text collection, the researcher can proceed analogously to the sociolinguist who needs to select informants from a community without skewing the sample. If there are identifiable social characteristics associated with the speakers or writers to be investigated, then it may make sense to decide on a predetermined quasi-quota sample in which these categories will be represented in reasonable proportions. Categories may have to be adjusted – so, for instance, in the CEEC the social class dimension is realized as "lower gentry," "upper clergy," "merchants," and so on (Nevalainen & Raumolin-Brunberg, 1996), categories that may sound alien to a modern sociolinguist but nevertheless capture the same phenomenon. Selections from within any such category should then, all other things being equal, follow the established principle of random sampling, in order not to introduce any subconscious bias – for instance, by selecting every *n*th text from a collection, where *n* is a number chosen to secure the distribution of texts selected for analysis from the entire range of sources available.

How? Analysis Procedures

In principle, the methodology to be applied in analyzing the data should adopt the same procedures and methodological toolkit as in the study of spoken data once the data collection has been concluded. In other words, the usual quantitative methodology can be applied: determine the envelope of variation (that is, the different variants that realize a variable); count token frequencies by category; represent the results in tables or graphs and interpret them properly; possibly apply multivariate techniques and test for statistical significance.

It is probably safe to say that for projects based on written data, sample sizes and token numbers tend to be smaller (though not necessarily so) – a fact that may impose limitations such as the non-availability of certain statistical tests which require minimum token numbers. This is a vague assessment out of experience, however, not a matter of principle. One important consequence that may result from smaller sample sizes and token bases may be that increasing importance has to be assigned to a purely qualitative perspective – that is, simply determining the presence of certain phenomena or forms in their respective contexts, irrespective of their frequency. Occurrences of individual forms as such, the presence of even a single token, may be of interest if they occur in an unexpected context or variety. More broadly, especially when looking at historical data, a researcher interested in evolutionary patterns may be interested in finding out which variants can be observed.

How Far? Assessing and Validating the Quality of Data

There are certain things that simply cannot be done on the basis of written data, and other limitations that the nature of such investigations imposes. In many cases, we will have to live with such limitations and do the best that can be done, but consider the circumstances in interpreting the results with care and reluctance. I distinguish three types of problems to be considered: the unavailability of sound; limited sample sizes; and issues of representativeness and validity.

Writing renders pronunciation only highly indirectly, via conventional grapheme–phoneme correspondences, which are hardly ever one to one and leave a lot of room for interpretation and ambiguity. Some forms seem justifiably interpretable – for example, spellings like *heah* 'hear' and *dat* 'that' (found in the character Jim in *Huckleberry Finn*; Minnick, 2004, p. 66) suggest a lack of rhoticity word-finally and stopping of the word-initial fricative, respectively. Others are impossible to interpret with certainty: do the graphs <ea> in *beaker* signal /iː/, /eɪ/, or possibly even /aɪ/ or /e/? And some phonologically relevant distinctions simply cannot be rendered by means of conventional orthography, such as the presence or absence of voicing in <th> or <s>. So, investigating pronunciation and sound changes on the basis of written records is possible but always implies uncertainty and requires reluctant interpretation.

Earlier, I mentioned and addressed the "bad data" or "insufficient data" problem, or the fact that, owing to coincidences of text recording and transmission, text sizes representing an interesting variety may be small and, more importantly, cannot be increased, and consequently the same applies to token numbers of some phenomena of interest. Strategies for the researcher to remedy this limitation can include conducting further searches for more data (which may or may not be successful), adopting a qualitative rather than a quantitative methodology, and generally being careful in drawing conclusions.

The issues of representativeness and validity concern the quality of the match between the variety we are interested in and the nature of the records we have of it – in the case of written records, by necessity an indirect one and thus one that needs qualification. Representativeness is defined as the relationship between a sample and the population it stands for. It may be affected by a biased selection, given that the skill of reading and writing in certain cultures was available only to higher social ranks, and written texts may thus not have been produced by the vernacular speakers a sociolinguist is interested in. This problem seems almost insurmountable, if it applies and seriously affects the record. In a sense, the procedure then reverses a modern sociolinguist's approach: the issue is not how to best construct a sample but how to make the best (and recognize the limitations) of the available sample (sociolinguists working with existing oral data, not collected for the investigation in question, may be faced with the same problem).

Validity, on the other hand, concerns the quality of a record, in this case the faithfulness of a written representation to a specific speech act that it records and represents. How accurately a written record at hand matches the spoken original we are interested in depends on a number of factors: the temporal or physical distance between the speech act and the recording process, the varying individuals involved (e.g., speaker and writer), and the extent to which the writer desired to provide an accurate, verbatim transcript. Typically, these qualities can be assessed by considering the nature of text types, as worked out above. In addition, there are various criteria and techniques for validating the quality of a record to some extent, such as (1) internal consistency within a document: if a certain phenomenon is consistently rendered in a specific way that seems meaningful, and if the ways of rendering related linguistic phenomena yield a seemingly systematic paradigmatic system, this will inspire confidence in the quality of the record; and (2) external fit: if the results observed seem to largely match descriptions of comparable projects, based on related varieties or sources, this backs the belief in the value of one's data.

Conclusion

Investigating written data sources offers a rich potential to sociolinguistics and allows us to ask interesting questions, despite (or perhaps even because of) certain inherent limitations in the nature of the data. The value of this approach is greatest for situations in which we have no other original information on record, most notably in the study of the history of vernacular varieties.

Note, however, that even if at first glance sources and procedures to be adopted seem somewhat unusual and possibly unique, in fact this methodology is not fundamentally different from the one applied in synchronic, oral-based sociolinguistics. We should avoid a naïve belief in the apparently immediate authenticity of oral data: material that a "modern" sociolinguist investigates also needs to be collected, selected, transcribed, and so on – all processes that involve human (researchers') agency and are thus error-prone, possibly guided by one's subconscious goals. Methodological awareness and care are important in any kind of project, irrespective of whether the data are oral or written: we have to step back for a moment, assess what we have, and what is being done, as distantly as is reasonably possible; we have to be careful in interpreting our results and be considerate about what the nature of our data, oral or written, reasonably allow us to do with them.

References

Bailey, G. (1997). When did Southern English begin? In E. W. Schneider (Ed.), *Englishes around the world 1: General studies, British Isles, North America* (pp. 255–275). Amsterdam: John Benjamins.

Ellis, M. (1994). Literary dialect as linguistic evidence: Subject–verb concord in nineteenth-century Southern literature. *American Speech, 69*, 128–144.

Huber, M. (2007). The *Old Bailey proceedings*, 1674–1834: Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In A. Meurman-Solin & A. Nurmi (Eds.), *Annotating variation and change*. E-series "VARIENG: studies in variation, contact and change in English." Retrieved from http://www.helsinki.fi/varieng/journal/volumes/01/index.html

Kautzsch, A. (2000). Liberian letters and Virginian narratives: Negation patterns in two new sources of earlier AAE. *American Speech, 75*, 34–53.

Kretzschmar, W. A., Jr., Darwin, C., Brown, C., Rubin, D. L., & Biber, D. (2004). Looking for the smoking gun: Principled sampling in creating the Tobacco Industry Documents Corpus. *Journal of English Linguistics, 32*(1), 31–47.

Mille, K. W. (1997). Ambrose Gonzales's Gullah: What it may tell us about variation. In C. Bernstein, T. Nunnally, & R. Sabino (Eds.), *Language variety in the South revisited* (pp. 98–112). Tuscaloosa: University of Alabama Press.

Milroy, J. (1992). *Linguistic variation and change: On the historical sociolinguistics of English*. Malden, MA: Blackwell.

Minnick, L. C. (2004). *Dialect and dichotomy: Literary representations of African American speech*. Tuscaloosa: University of Alabama Press.

Montgomery, M. (1997). A tale of two Georges: The language of Irish Indian traders in colonial North America. In J. L. Kallen (Ed.), *Focus on Ireland* (pp. 227–254). Amsterdam: John Benjamins.

Nevalainen, T., & Raumolin-Brunberg, H. (Eds.). (1996). *Sociolinguistics and language history: Studies based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.

Poplack, S. (Ed.). (2000). *The English history of African American English*. Malden, MA: Blackwell.

Rawick, G. P. (Ed.). (1972–1979). *The American slave: A composite autobiography*. Westport, CT: Greenwood Press.

Rissanen, M. (1997). "Candy no witch, Barbados": Salem witchcraft trials as evidence of Early American English. In H. Ramisch & K. Wynne (Eds.), *Language in time and space* (pp. 183–193). Stuttgart: Steiner.

Schneider, E. W. (1989). *American Earlier Black English: Morphological and syntactic variables*. Tuscaloosa: University of Alabama Press.

Schneider, E. W. (2002). Investigating variation and change in written documents. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 67–96). Malden, MA: Blackwell.

Schneider, E. W. (2012). Tracking the evolution of vernaculars: Corpus linguistics and earlier Southern US Englishes. In J. Mukherjee & M. Huber (Eds.), *Corpus linguistics and variation in English: Theory and description* (pp. 185–212). Amsterdam: Rodopi.

Schneider, E. W., & Wagner, C. (2006). The variability of literary dialect as a reflection of pan-lectal competence: Jamaican Creole in Thelwell's *The harder they come*. *Journal of Pidgin and Creole Languages, 21*, 45–95.

Van Herk, G., & Poplack, S. (2003). Rewriting the past: Bare verbs in the Ottawa Repository of Early African American English Correspondence. *Journal of Pidgin and Creole Languages, 18*, 231–266.

Vignette 11a
Accessing the Vernacular in Written Documents

France Martineau

While the intrinsically immediate nature of spontaneous oral exchanges is difficult enough to reconstitute in modern sociolinguistics, it is quite impossible to reconstitute fully in the context of ancient times, as written documents are the only traces left of this exchange. It is a mistake, however, to transfer the dichotomy between the oral code and the written code to the continuum between the poles of "language of proximity" and "language of distance" (Koch & Oesterreicher, 2001), as many researchers have done. The oral and the vernacular have so monopolized sociolinguistic research as to push into the background a broad segment of the range of variation: written material. I have shown, through my research on written vernacular documents, that oral features are not completely absent from some types of written documents, such as family correspondence or plays. While it is true that the ultimate resource – the actual sound of speech – is gone forever, written traces may nevertheless restore its immediacy in part through pronunciation features, morphosyntactic variation, or stylistic idiosyncrasies. The value of written documents goes beyond the evidence they can provide on the contemporary vernacular; these documents can also reveal the linguistic attitudes, norms, and standards that constitute a community. My experience using written documents such as family correspondence for sociolinguistic ends shows that doing this involves long searching in archives, a slow reconstruction of the relationship between written language and traces of the oral, and a patient contextualization of the profile of the writers and their community. This work requires the researcher to step out of her or his comfort zone and develop abilities in more than one domain of the social sciences and humanities: linguistics, literature, sociology, social history, genealogy, paleography, and/or anthropology.

Digging into Data

Not all written documents have the same linguistic interest. Some that are dear to historical researchers are of little use to linguists; trial transcripts, for example, reproduce what was said by accused persons or witnesses drawn from various social classes but must be handled very carefully by a linguistic researcher because of the filtering applied by the transcriber. A second type of document – promissory notes or bills of sale – gives access to a larger social palette, since more people may have been able to write such short documents, but it is almost impossible to explore statistical linguistic trends on the basis of such sparse material. And accounting ledgers, while very useful for lexicographical purposes, offer little information on the bookkeeper's grammar.

Moving on to long-term diaries or regular exchanges of correspondence, we may hope to come upon language that is closer to contemporary usage. In a similar way to exchanges between friends, family letters reflect a fairly close relationship between the writer and recipient, despite the use of the written medium. The topics they address are generally of interest to both parties, who subscribe to a kind of letter writers' pact not found in such oral material as folktales or plays featuring popular characters, or even some modern sociolinguistic interviews.

Collections of correspondence constitute an important proportion of the documents in archives. They have been widely used in the humanities and social sciences, particularly social history, in the investigation of micro-societies. Unfortunately, because documents of linguistic interest often offer little of immediate value in tracking major historical events, they are not easily found. Archivists' taxonomies generally organize collections and holdings in a way that makes it difficult to locate documents of linguistic interest – for instance, through the names of the great families that made history, such as the Baby, Papineau, and Mackenzie families – or collections linked to major historical events, such as the War of Secession or the conquest of New France. Even more importantly, the descriptions of documents in finding aids are in many cases more detailed if the items are associated with known writers, or events that made history. In many cases, even the language of correspondence is not specified, not to mention the fact that some linguistic groups have not left many documents of their own (for example, in research on colonial languages, Amerindian languages are not well represented in documents, except through colonial documents written by missionaries or administrative officers). The job of the linguist, faced with this mass of documents organized primarily to benefit research on the sociopolitical history of communities, is to find the thread that permits access to the language of individuals.

Archival searches for documents of linguistic interest must take into account what prompted the writer to take up his or her pen and, equally importantly, what motivates people to preserve documents over the centuries. Where there were documents created by persons with little education, few have survived to our day, given the scant interest generated in their preservation. If they did survive, it was usually because they were preserved as part of larger collections that held historical interest for their contemporaries or for collectors. Large family holdings are therefore of particular interest because, in addition to correspondence between family members, they often contain documents from less-well-off members of the extended family, as well as from small storekeepers or employees.

Some Features of Written Vernacular

Even in spontaneous and informal exchanges, people with more education tend to eliminate traces of everyday conversation from their letters, either by avoiding vernacular characteristics or by observing letter-writing conventions. This makes such letters valuable sources of information on the relationship of the elite with oral or written standards. On the other hand, those unskilled in the use of language may realize that there is a model to be followed but be unaware of some of the rules. They may try to give their prose a written polish, using remembered openings or closings: *Je vous écris pour faire assavoir de mes nouvelles qui sont bonnes Dieu marci!* (I am writing to acquaint you with my news, which is good, thank God!) (see also Schlieben-Lange, 1998). Moving on from such clichés, a writer with little education – particularly when addressing a family member – may also seek to reassure the family by describing her or his circumstances, which necessitates assuming a more personal tone that often contains features of the spoken language.

In written material, spelling features – and their variation from the standard spelling – are the most salient features. I have shown that, for French, before the 19th century and the increased prestige of written language following the progression of literacy in the population, it was not unusual for a writer from a higher social class to deviate from standard usage, while complying with grammatical norms (Martineau, 2007). Yet spelling deviations that give us clues to vernacular pronunciation are often found in the misspellings of those of lower social class. For instance, in the exceptional 1765 diary of an ordinary merchant that I found (Martineau & Bénéteau, 2010), we see well-known consonantal reduction features, including:

- final consonant [d]: *Le ſûe* [<sud} *de Lariviere* "south of the river";
- final [s]: *jetue ſejour La un nourre* [<*un ours*] "that day, I killed a bear";
- internal [r]: *Mecredie* [<*mercredi*] *Le 6 je pertie* "Wednesday the 6th, I left";
- [l] in an obstruent + liquid group: *Lartique* [<*l'article*] *du Cha* "the thing about the raccoon."

Those with little education may have a poorer grasp of written conventions but nevertheless sufficient knowledge to avoid merely writing what they say. Reconstructing pronunciation from written material is an exercise that requires taking into account the phonological environment, written consistency, indications from other material from the same region or period, and traces in the contemporary spoken language. As with spelling, only a comparative approach makes it possible to place the regional vocabulary or vernacular that sometimes surfaces in private correspondence.

Written documents must always be contextualized relative to other texts in the same genre and to texts from other genres. Historical linguistic researchers often refer to comedies (plays), arguing that their language is more representative of the oral than the language of letters. However, plays are also subject to the conventions of a genre and a community. For instance, when vernacular features acquire social saliency, they generally disappear from correspondence, while being widely used to represent popular speech in play material. On the other hand, when they are not salient, they are less common in plays because of their weak symbolic value. They may, however, appear in the writings of those with little education, thus signaling incipient linguistic change. One example is the omission of the negative particle *ne* in French (e.g., *Je veux pas* [<*Je ne veux pas*] "I don't want"), a feature shared nowadays by speakers of every social class. As I have shown through comparison of plays and letters from poorly educated writers, this feature shows up as early as the 18th century in the writings of the uneducated but is not used to suggest popular speech in theatrical material until a century later (Martineau, 2011).

"Bad" and "Best" Data

Historical sociolinguistics accesses the oral code through the written medium for lack of an alternative. Contemporary situations are different, as both oral and written media are accessible but the oral receives the most attention. Written documents from the past are not "bad data"; they become so if contrasted strictly with contemporary oral material, gathered by methods that stress certain types of register. What is lacking in many cases is not historical documents but a shared perspective between historical linguistics and modern sociolinguistics and bases for comparison.

Historical sociolinguistics forces us to work with writing and, at the same time, to reexamine the relationship between the written and the oral, as well as to consider more usages in various situations. The sociolinguistics of contemporary speech would also profit by considering the influences of written language, whose distinctions from spoken language are sometimes less than clear. In online chat, for example, writing becomes the medium for exchanges that appear to be much like oral exchanges between friends, to the point where the boundaries are sometimes completely blurred, thanks to tools like Skype, where participants can maintain an oral conversation at the same time as they exchange written comments. Even in interviews, although the medium is oral, the speaker's relationship with the written language may interfere with his or her spoken production, through reminiscences of written ads, literary citations, explicit grammar rules, and so on. The characteristics of oral and written codes are not independent of each other, although the strong writing tradition in most contemporary cultures, particularly since the 19th century, tends to place them on parallel tracks.

References

Koch, P., & Oesterreicher, W. (2001). Langage parlé et langage écrit. In G. Holtus, M. Metzeltin and C. Schmitt (Eds.), *Lexikon der romanistischen Linguistik* (Vol. 1, pp. 584–627). Tübingen: Max Niemeyer Verlag.

Martineau, F. (2007). Variation in Canadian French usage from the 18th to the 19th century. *Multilingua, 26*(3), 203–227.

Martineau, F. (2011). *Ne*-absence in declarative and *yes/no* interrogative contexts: Some patterns of change. In P. Larrivée & R. P. Ingham (Eds.), *The evolution of negation: Beyond the Jespersen cycle* (pp. 179–208). Berlin: Walter de Gruyter.

Martineau, F., & Bénéteau, M. (2010). *Incursion dans le Détroit. Journaille Commansé le 29 octobre 1765 pour Le voiage que je fais au Mis a Mis.* Quebec City: Les Presses de l'Université Laval.

Schlieben-Lange, B. (1998). Les hypercorrectismes de la scripturalité. *Cahiers de Linguistique Française, 20*, 255–273.

Vignette 11b
Adapting Existing Data Sources

Language and the Law

Philipp Sebastian Angermeyer

Language and law is a growing interdisciplinary field that comprises research on a wide range of topics in law, linguistics, anthropology, and sociology, focusing on the investigation of language use in legal settings or for legal purposes. Much of this research can be broadly qualified as sociolinguistic (see, for example, Eades, 2010), and many leading sociolinguists have become involved in the field at some point in their careers. This research draws on various types of spoken and written data. In contrast to other fields, data in language and law generally exist independently of their linguistic analysis and often are not collected by researchers at all but rather are produced by institutional actors for institutional purposes before becoming the subject of linguistic investigation. This has significant ethical and methodological implications, and it raises questions about the nature of linguistic data more generally – that is, about sociolinguistic practices of transcription and annotation and their relationship to practices outside of academia (Bucholtz, 2000).

Sociolinguistic research on language and law can be divided into two broad categories depending on its goals, namely whether scholars are drawing on data from legal contexts to address broader sociolinguistic research questions or whether they are applying sociolinguistic theories and methodologies to address questions of legal significance (often referred to as forensic linguistics). These lines of research often differ in the types of data they rely on and in the conditions for obtaining data, so I shall discuss them separately. At the same time, some scholars have engaged in both types of research, often beginning with forensic linguistic analysis and then turning to questions of interest for the field, while drawing on the same data.

Data Sources in Forensic Linguistics

Forensic linguists provide expertise on a wide range of legal issues, including, for example, comprehensibility of written texts, trademark disputes, or speaker or author identification (Tiersma & Solan, 2002). Sociolinguists have become involved in forensic linguistics especially in situations where expertise on a person's vernacular is of legal relevance (for example, as evidence for a person's place of origin, particularly in asylum interviews; Singler, 2004) or to interpret the meaning of utterances made in a language or variety other than the standard variety used in court (Bucholtz, 2009). Sociolinguists have also provided expertise in discourse analysis, for example drawing on speech act theory to address disputes over the meaning of utterances in interaction (Shuy, 1996). Such analyses are generally based on prerecorded audio recordings that are made available to linguists by prosecutors or defense attorneys; types of such recordings include distress calls, threatening phone calls (Labov, 1988), wiretap surveillance recordings (Bucholtz, 2009; Shuy, 1996), and official recordings of police interrogations (Berk-Seligson, 2009; Bucholtz, 2000).

Sociolinguists may also rely on data collected through interviews or reading tasks (Labov, 1988). This is particularly common for language analysis in the determination of origins (LADO), where asylum seekers are interviewed in order to identify their vernacular variety, which is then taken as potential evidence for their place of origin. Singler (2004) compares this type of "linguistic asylum interview" to the sociolinguistic interview, noting that the circumstances of the asylum interview systematically discourage the use of the vernacular by the interviewee. The comparison shows that there may be systematic obstacles for the application of sociolinguistic methods to forensic data. Bucholtz (2009) notes that ethical and methodological concerns arise from the facts that linguistic experts work at the request of a particular client and that analyses are often conducted without the recorded speakers' awareness or consent.

Data Sources in Sociolinguistic Studies of Language and Law

Sociolinguists have also turned to legal settings as a source of data for studies that address issues of broader sociolinguistic significance in fields such as conversation analysis, language and gender, intercultural communication, and language ideology. Such research generally investigates power and inequality in spoken interaction in settings such as courtrooms, police interrogations, or asylum interviews, and focuses on narratives and on question and answer sequences, as well as on the consequences of linguistic variation, diversity, and interpreting.

In contrast to forensic linguistic analysis, this type of research often encounters significant obstacles in the process of data collection. Data collection in the form of audio or video recordings, accompanied by participant observation, is not possible in many jurisdictions and for various types of legal interactions (but see Conley & O'Barr, 1990; Komter, 1998; Maryns, 2006; McElhinny, 1995). Where linguists are barred from making their own recordings or from using institutionally produced ones, they have turned to various alternative data sources. Several studies have taken advantage of televised court proceedings, investigating trials or inquiries for which audio and video data are publicly available (e.g., Cotterill, 2003; Ehrlich & Sidnell, 2006; Matoesian, 2001). Similarly, Heritage and Clayman (2010) draw on a documentary film to investigate talk in a jury deliberation, a setting that is normally inaccessible to researchers.

Finally, some linguists have used official, institutionally produced transcripts of court proceedings, arguing that they represent a suitable source of data for some types of analysis, particularly when accompanied by ethnographic research (Atkinson & Drew, 1979, p. xviii; Heffer, 2005, p. 58). Official transcripts have also been used as a source of data in studies of institutional entextualization – that is, the process by which spoken language use is transformed into written documents such as witness statements, confession statements, or court records (Rock, 2001). As Bucholtz (2009, p. 507) notes, such texts are not "neutral records of what was said"; rather, they have often been found to be biased in favor of institutional actors. To make matters worse, such entextualizations effectively replace recordings in the evidentiary record, being viewed "as reflections rather than representations of prior speech events" (p. 516).

Besides official transcripts of spoken interaction, other types of legal texts have also been used as data sources for sociolinguistic research. In particular, studies that investigate language ideologies have used published judicial opinions as a further written source of data (Berk-Seligson, 2009; Haviland, 2003; Lippi-Green, 1994). For my own research on interpreter-mediated court proceedings, I obtained permission from court administrators and from my university's Institutional Review Board (IRB) to make my own audio recordings (Angermeyer, 2009). My negotiations with court administrators were aided by the fact that they viewed my research as having a potential real-world application (evaluating and improving the "quality" of interpreting services), even if this was not my primary focus. Securing approval from the IRB was facilitated by the fact that the court proceedings were open to the public.

Conclusion

Whether serving forensic or sociolinguistic goals, sociolinguistic research in legal settings is often dependent on institutional data sources and thus relies on the cooperation of legal institutions in the research process. Consequently, access to such data may be jeopardized by a researcher's critical stance toward institutional practices. Furthermore, confidentiality often prevents linguists from sharing such data with other scholars and may limit the extent to which findings can be disseminated. At the same time, language and law continues to be a growing area of interest in sociolinguistics, particularly as experience with forensic data has motivated sociolinguists to advocate for changes to language practices in the legal system.

References

Angermeyer, P. S. (2009). Translation style and participant roles in court interpreting. *Journal of Sociolinguistics, 13*(1), 3–28.

Atkinson, J. M., & Drew, P. (1979). *Order in court: The organisation of verbal interaction in judicial settings*. London: Macmillan.

Berk-Seligson, S. (2009). *Coerced confessions: The discourse of bilingual police interrogations*. Berlin: Mouton de Gruyter.

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

Bucholtz, M. (2009). Captured on tape: Professional hearing and competing entextualizations in the criminal justice system. *Text and Talk, 29*(5), 503–523.

Conley, J. M., & O'Barr, W. M. (1990). *Rules versus relationships: The ethnography of legal discourse*. Chicago: University of Chicago Press.

Cotterill, J. (2003). *Language and power in court: A linguistic analysis of the O. J. Simpson trial*. New York: Palgrave Macmillan.

Eades, D. (2010). *Sociolinguistics and the legal process*. Bristol: Multilingual Matters.

Ehrlich, S., & Sidnell, J. (2006). "I think that's not an assumption you ought to make": Challenging presuppositions in inquiry testimony. *Language in Society, 35*(5), 655–676.

Haviland, J. B. (2003). Ideologies of language: Reflections on language and U.S. law. *American Anthropologist, 105*, 764–774.

Heffer, C. (2005). *The language of jury trial: A corpus-aided analysis of legal-lay discourse*. Basingstoke, UK: Palgrave Macmillan.

Heritage, J., & Clayman, S. (2010). *Talk in action: Interactions, identities and institutions*. Malden, MA: Blackwell.

Komter, M. L. (1998). *Dilemmas in the courtroom: A study of trials of violent crime in the Netherlands*. Mahwah, NJ: Lawrence Erlbaum.

Labov, W. (1988). The judicial testing of linguistic theory. In D. Tannen (Ed.), *Linguistics in context: Connecting observations and understanding* (pp. 159–182). Norwood, NJ: Ablex.

Lippi-Green, R. (1994). Accent, standard language ideology and discriminatory pretext in the courts. *Language in Society, 18*, 213–234.

Maryns, K. (2006). *The asylum speaker: Language in the Belgian asylum procedure*. Manchester: St. Jerome Publishing.

Matoesian, G. M. (2001). *Law and the language of identity: Discourse in the William Kennedy Smith rape trial*. New York: Oxford University Press.

McElhinny, B. S. (1995). Challenging hegemonic masculinities: Female and male police officers handling domestic violence. In K. Hall & M. Bucholtz (Eds.), *Gender articulated: Language and the socially constructed self* (pp. 217–243). New York: Routledge.

Rock, F. (2001). The genesis of a witness statement. *Forensic Linguistics, 8*(2), 44–72.

Shuy, R. (1996). *Language crimes: The use and abuse of language evidence in the courtroom*. Malden, MA: Blackwell.

Singler, J. V. (2004). The linguistic asylum interview and the linguist's evaluation of it: Liberian applicants for political asylum in Switzerland. *International Journal of Speech, Language and the Law, 11*(2), 222–239.

Tiersma, P., & Solan, L. (2002). The linguist on the witness stand: Forensic linguistics in American courts. *Language, 78*(2), 221–239.

Vignette 11c
Advances in Sociolinguistic Transcription Methods

Alexandra D'Arcy

I am an avid proponent of transcription. Admittedly, transcription entails a huge time investment (and, if you are not doing it yourself, a huge financial investment). How much time exactly? The standard baseline is six to ten hours of transcription time for every hour of speech, but the dividends pay off. The result is a permanent record that is electronically searchable, reproducible, and concordance-able. Depending on the software you select (e.g., CLAN, ELAN, Transcriber), the orthographic files can be time-aligned with the audio files, which vastly simplifies the task of data management. Of course, basic .doc(x) or .txt files are no less valuable; they just perform fewer "tricks."

The first time I had to transcribe data, I initially felt overwhelmed. How to begin? Someone gave me a transcription manual, but it was written for conversation analysts and followed their conventions. Without an understanding of why those conventions (which mark vowel and pause length, voice quality, types of laughter, and the like, and which may or may not also be appropriate for quantitative sociolinguistic work), I was unsure how to apply what the manual recommended. Finally, I simply sat down and started typing out the conversations, and while I still use those original files, baptism by fire is not the ideal introduction to a foundational instrument of the sociolinguistic toolkit. My goal in this vignette is to lay the groundwork for you to make your own informed choices about transcription in your studies.

Transcription enables analysis of spoken language. Its primary goal is to reproduce speech faithfully and consistently, creating an authentic representation of language in use (Poplack, 1989, p. 434; Tagliamonte, 2006, p. 55). The details of transcription tend to receive more attention from qualitative sociolinguists, for whom what appears in a transcript both influences and constrains the generalizations that can be drawn (i.e., transcriptions tend to be theoretically informed) (see Ochs, 1979). For quantitative sociolinguists, the procedure tends to be pre-theoretical. Regardless of your orientation, the transcript forms the basis for analysis, and, as such, all the "messy" phenomena that characterize unscripted, unprompted, casual speech (overlaps, hesitations, false starts, malapropisms, etc.) cannot be left out or glossed over. Similarly, the language cannot be "cleaned up": non-standard forms should be retained, not replaced with their standard counterparts.

But there is more to transcription than verbatim reproduction. Before the first word is typed, some decisions have to be made. As the literature suggests, you need a protocol. It may not need to be as thorough as a transcription manual, but an account of your transcription choices makes your transcription process transparent and verifiable and helps you make consistent decisions. However, there is no standard protocol for sociolinguistic transcription. Every corpus is different, built for different purposes, to answer different questions, in different locales, with different demographics. A defining element of sociolinguistic corpora is their specialized nature, in that they are designed with a particular research question in mind (D'Arcy, 2011; Poplack, 2007). Therefore, no single decision can hold for all projects. Most decisions revolve around four themes: orthography, punctuation, phonetic detail, and spontaneous speech phenomena.

Orthography

Most researchers stress the need for standard orthographic conventions (to simplify concordances and automatic searches), but sometimes there is good reason to use non-standard spellings. Consider speech containing dialect forms (e.g., *nae* for *no*, *tiv* for *to*). Since dialect words are fundamental features of local speech, standardizing their spelling would distort the authenticity of the data. Hyphenation is something to keep in mind as well, as it affects word counts and concordances: contracted, fused, and hyphenated sequences count as a single word or entry. This issue may or may not matter to you, but what will matter, ultimately, is consistency. Consistency is the reason I advocate that frequent homophones get one spelling. Consider a form such as *like*. Setting aside its established "grammatical" functions (verb, noun, conjunction, suffix), *like* is also (increasingly frequently) used to perform a number of discourse functions: quotative, adverb, discourse marker, discourse particle (D'Arcy, 2007). I have seen protocols that stipulate using *like* for grammatical functions and *lyke* for discourse functions (Poplack et al., 2006). Crucially, it is not always easy, or even possible, to disambiguate what *like* is doing in a particular instance. In such cases, consistency is not only lost but also unachievable, compromising the integrity of the transcripts because errors of interpretation are inevitably introduced.

Punctuation

We may take standard punctuations for granted in an academic setting, but in the context of sociolinguistic transcription it is an open question. Some researchers feel that using full stops, commas, and questions is critical (Preston, 1985; 2000; Tagliamonte, 2006; 2007), as they increase readability and searchability. Others reserve features of standard punctuation for special cases (Maclagan & Hay, 2011). For example, the Origins of New Zealand English Project (ONZE) does not allow commas, uses a full stop strictly to indicate hesitation, and permits question marks only for intonational questions. One thing punctuation does affect is syntactic parsing (i.e., tagging the data with structural information). Most parsers depend on full stops to disambiguate embedded clauses from independent sentences. If parsing is an option you would like to allow for, then it helps to build at least a minimal level of punctuation into your protocol from the outset.

Phonetic Detail

Regardless of intent, a transcription can never be more than an interpretation of speech. No orthographic rendering can be "so detailed and precise as to provide for the recreation of the full sound" (Macaulay, 1991, p. 282). In the end, you need to decide how much interpretation to impose on your data. Selectivity is encouraged (cf. Ochs, 1979, p. 44). Many features of spoken language are predictable from general processes (phonetic or otherwise), and so whether or not to mark consonant cluster simplification, assimilation, vowel reduction, and/or ellipsis is a matter of weighing the risk vs. the reward. Consistency is critical, and the more things there are to remember, hear, and do while transcribing, the greater the likelihood of making errors. We can also ask what a particular decision buys us, analytically speaking. For example, if you decide to mark (ing) variation orthographically (e.g., *running*, *runnin'*), then you are forcing yourself to perform auditory analysis while also attempting to faithfully reproduce the whole of the spoken text itself. And the question is, at what cost? The transcriptions will take longer to produce, they will require greater attention to detail, and – since your attention was divided – you will have to go back and listen to the data again to check your original "coding" of the variants.

Finally, it helps to have a plan for overlapping or incomprehensible speech, interruptions, backchanneling cues, and all the sundry discourse phenomena that happen when people talk (for a good introduction, see Ochs, 1979; Maclagan & Hay, 2011). Also think about colloquialisms, which are frequent in speech (e.g., the ONZE protocol stipulates that *gonna*, *gotta*, and *wanna* are acceptable renderings but does not permit *hafta*, *woulda*, or *mighta* to appear in transcriptions).

After all this, what is the best way to start? Slowly. And with a clear understanding of what you will do and why. Plan to punctuate? Hyphenate? Capitalize? Use a concordance program to search and/or extract data? You do not need a 20-page protocol before you start, nor do you have to have all the answers, but you should at least have thought about why you are making these transcriptions. Ultimately, what you plan to use them for and how (or whether) you intend to share them factor into the decisions you make. If you have never transcribed data before, then proceed with the following general caveat in mind: increased detail does not necessarily lead to increased quality. At the end of the day, you cannot work from either the transcriptions or the audio alone. You will need both, so make them work for you. After all, you worked hard for them.

References

D'Arcy, A. (2007). *Like* and language ideology: Disentangling fact from fiction. *American Speech, 82*, 386–419.

D'Arcy, A. (2011). Corpora: Capturing language in use. In W. Maguire & A. McMahon (Eds.), *Analysing variation in English* (pp. 49–72). Cambridge: Cambridge University Press.

Macaulay, R. K. S. (1991). Coz it izny spelt when they say it: Displaying dialect in writing. *American Speech, 66*, 280–291.

Maclagan, M., & Hay, J. (2011). Transcription. In M. Di Paolo & M. Yaeger Dror (Eds.), *Sociophonetics: A student's guide* (pp. 36–45). New York: Routledge.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York: Academic Press.

Poplack, S. (1989). The care and handling of a mega-corpus: The Ottawa–Hull French Project. In R. W. Fasold & D. Schiffrin (Eds.), *Language change and variation* (pp. 411–451). Amsterdam: John Benjamins.

Poplack, S. (2007). Foreword. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora* (Vol. 1, pp. ix–xiii). Basingstoke, UK: Palgrave Macmillan.

Poplack, S., Walker, J. A., & Malcolmson, R. (2006). An English "like no other"? Language contact and change in Quebec. *Canadian Journal of Linguistics, 51*, 185–213.

Preston, D. R. (1985). The Li'l Abner syndrome: Written representations of speech. *American Speech, 60*, 328–336.

Preston, D. R. (2000). Mowr and mowr bayud spellin': Confessions of a sociolinguist. *Journal of Sociolinguistics, 4*, 614–621.

Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Tagliamonte, S. A. (2007). Representing real language: Consistency, trade-offs, and thinking ahead! In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora* (Vol. 1, pp. 205–240). Basingstoke, UK: Palgrave Macmillan.

Vignette 11d
Transcribing Video Data

Cécile B. Vigouroux

Learning how to transcribe is like learning how to ride a bicycle: one can only grasp how to do it by practicing, although some theoretical guidelines may prevent missteps. The main difference between bicycling and transcribing is that whatever bicycle one rides, it is always the same way of riding. Transcription must be adjusted with every kind of linguistic material one transcribes, which requires its customized "protocol." The reason is that the how-to-transcribe question mainly boils down to what one has to transcribe.

Transcription has certainly been one of the most discussed epistemological topics in sociolinguistic research since Ochs (1979), which draws attention to the theoretical implications of the activity. Nonetheless, little thinking has been devoted to transcription of video material, for at least two main reasons: First, video recordings have been used only recently in sociolinguistics, although the practice has had a long tradition in the social sciences (for a brief historical perspective, see Erickson, 2011). Second, theories of language still rest heavily on speech, despite linguists' increased awareness of its multimodal aspects (Duranti, 1997).

Many theoretical and technical issues regarding the transcription of video data are similar to those pertaining to audio recordings. For instance, transcription must be approached in connection with other research activities involved in data construction, namely the fieldwork before it and, afterward, writing and data analysis. At the same time, as is argued by many, a clear distinction between transcribing and analyzing is hardly tenable. Before getting to the point at which video materials must be transcribed, the researcher must have asked her- or himself why a given linguistic event should be video- rather than audio-recorded in the first place. The belief that video recordings may provide more information on the language practices and social actions under study is not a sufficient reason if the researcher doesn't know what to do with such (often overwhelming) information.

In addition, without really knowing what she or he is looking for, a researcher cannot frame her or his video recording adequately. For example, in my work on language practice in a Congolese Pentecostal church in South Africa, I was particularly interested in the pastor's and his English interpreter's joint performance. Prior observations in the church had helped me position my camera where it would not distract from the church service while capturing both performers' facial expressions and body postures. Because, prior to the video recordings, I hypothesized that this joint performance should not be analyzed in isolation from concurrent staged semiotic events (such as music playing and singing), I chose a wide frame in order to record simultaneous social actions. My choices in the field shaped my transcription and therefore my analysis, especially in two ways: (1) my account of the joint performance, and (2) my downplaying of the audience's role in co-constructing the staged performance. In order to emphasize the joint performance, I chose to display the transcription in a three-column format in which each performer is given equal "visual weight" (Table 11c.1).

Table 11c.1 Transcription with Performers Given Equal "Visual Weight"

	Pastor	*Interpreter*	*Audience*
1	est-ce que vous croyez en		*((drum playing and*
2	la parole de Dieu/		*audience claps hands and*
3	do you believe in God's		*shouts))*
4	word/		↓
5		do you believe in *the	
6		word of God/*	
7		**((right index pointed to*	
8		*audience))**	
9	vous croyez en la parole		
10	de Dieu/		
11	you believe in God's		
12	word/		
13		do you believe in	
14	vous croyez en la parole	[XXXXXX/	
15	de Dieu/		
16	you believe in God's		
17	word/	do you believe in the word	*((drum stops))*
18		of God/	
19			
20	vous croyez en la parole		
21	de Dieu		
22	you believe in God's	do you believe in the word	
23	word/	of God/	amen
24			
25	(2:45) alleluia hallelujah		
26		amen	
27	*et si*		
28			
29	* and if*		
30	**((pointing left index*	(1:82) *if*	
31	*toward audience**))*	**((pointing right index*	
32		*toward audience))**	
33	*((gaze at the interpreter))*		
34			
35	vous êtes avec Dieu		
36	you are with God		
37			
38		you would be with God	

This display helps me, as an analyst and reader, visualize at a glance a crucial aspect of the joint performance: its temporality (for example, the arrow indicates the drum's overlap with the pastor–interpreter interaction). It also highlights the interpreter's gestural mimicking of the pastor (e.g., lines 31–32), leading me to raise questions about his communicative and symbolic role in the shaping of the pastor's sermon. My choice of a three-column format was also informed by my experience as a reader of scientific journals who often finds it strenuous to read their interactional data. As was pointed out by Ochs (1979), the visual display of data shapes the way readers assess the relevance of the researcher's hypotheses and analyses. While no transcriptions are better than others, some lend themselves to more accessible reading than others. Although all transcriptions are context based (as are the choices made in relation to the analyst's research questions), transcriptions are also embedded in a history of conventions and codifications of linguistic materials, in which the reader has been socialized (e.g., Jefferson's 1984 transcription conventions). Innovative idiosyncratic transcriptions may well serve a researcher's analysis, but they may also impede a reader's understanding of data (Eisner, 1997). In other words, ways of transcribing are also shaped by expected ways of reading.

Because video recording and the transcription activity are selective processes (e.g., deciding what to film, what to edit, what to transcribe), they both determine and bear on the formulation of hypotheses. For instance, because I filmed the pastor–interpreter joint performance with one camera, I lost a crucial aspect of the ongoing communicative activity at church: the participation of the congregation. As a transcriber and analyst, my access is limited to the congregants' vocal response (line 24, *amen*), which is a truncated "picture" of the interactional dynamics between the three parties. My three-column display of the interactions would have been enriched by data from a second camera turned on the audience. Of course, this point raises an important issue that has not yet been touched upon here: video-framing constraints. Filming the congregation would have meant getting formal consent from each of the congregants, which would have proved difficult. Indeed, the wealth of multimodal information provided by video recordings comes with sensitive ethical issues regarding informants' privacy and anonymity. Such issues must be well thought out in the preparation of fieldwork and the subsequent stage of transcription.

Choosing to use video data not only implies that we, analysts, conceive of language as intrinsically multimodal but also raises methodological and theoretical issues of workable representations of this multimodality. Scholars have explored various ways of representing multimodality, such as the use of still video footage, drawings, diagrams, and pictures from digitized still footage, to name a few. (See Plowman & Stephen, 2008, and Bezemer & Mavers, 2011, for critical overviews of different multimodal transcriptions, and Dicks, Soyinka, & Coffey, 2006, for a case study.) Today, there are also multiple alignment software programs that enable transcribers to establish a link between the recorded source and the transcription (e.g., Transana, CLAN, ELAN; Mondada, 2007, provides an overview of the different programs). However, useful as these tools are, they do not spare the transcriber the upstream theoretical work needed to construct a workable and, to a certain extent, readable transcript.

Out of context, no transcription method is better than others; a transcript should be approached as the outcome of multilayered decisions made according to specific research questions and informed by a specific research design. A comparison of Vigouroux (2007; 2009) in which I used the same video data illustrates this point: in my 2007 transcription, the data were displayed in line, whereas in 2009 I chose a column organization. This change was motivated by my closer attention to the synchronicity of speakers' gestures and gazes in the shaping of the interactional activity under scrutiny. Therefore, a transcript, be it video or audio, is never a definitive product; it is always shaped by the researcher's prospective analysis. In addition, any decisions made should be transparent not only to the researcher but also to her or his readers. Although any transcription is labor-intensive, it is a crucial process that determines working hypotheses and is one that is worth the strenuous ride.

References

Bezemer, J., & Mavers, D. (2011). Multimodal transcription as academic practice: A social semiotic perspective. *International Journal of Social Research Methodology, 14*(3), 191–206.

Dicks, B., Soyinka, B., & Coffey, A. (2006). Multimodal ethnography. *Qualitative Research, 6*(1), 77–96.

Duranti, A. (1997). *Linguistic anthropology*. Cambridge: Cambridge University Press.

Eisner, E. W. (1997). The promise and perils of alternative forms of data representations. *Educational Researcher, 26*(4), 4–10.

Erickson, F. (2011). Uses of video in social research: A brief history. *International Journal of Social Research Methodology, 14*(3), 179–189.

Jefferson, G. (1984). Transcript notations. In J. M. Atkinson & J. Heritage (eds.), *The structures of social action: Studies in conversation analysis* (pp. ix–xvi). Cambridge: Cambridge University Press.

Mondada, L. (2007). Commentary: Transcript variations and the indexicality of transcribing practices. *Discourse Studies, 9*, 809–821.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York: Academic Press.

Plowman, L., & Stephen, C. (2008). The big picture? Video and the representation of interaction. *British Educational Research Journal, 34*(4), 541–565.

Vigouroux, C. B. (2007). Trans-scription as a social activity: An ethnographic approach. *Ethnography, 8*(1), 61–97.

Vigouroux, C. B. (2009). The making of a scription: A case study on authority and authorship. *Text and Talk, 29*(5), 615–637.

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, 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 (http://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, http://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, http://www.fon.hum.uva.nl/praat/) and ELAN (Sloetjes & Wittenburg, 2008, http://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 the Bird and Simons (2003) paper 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, http://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, http://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 datasets 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 February 2013, over 2,900 interview recordings are stored in and accessible through SLAAP, amounting to over 2,400 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 Libraries 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 worked 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 centralized 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 (http://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 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.

Vignette 12a
Making Sociolinguistic Data Accessible

William A. Kretzschmar, Jr.

The general practice in sociolinguistics, and it is not atypical of work in other social sciences, has been to collect data for our own studies and then keep the data private, or in other social sciences even destroy the data, in order to preserve the confidentiality of research subjects. Indeed, ethics boards – such as, in the United States, Institutional Review Boards (IRBs) – have set policies for the protection of human subjects that encourage the destruction of social data like ours, since audio or video recordings of interviews cannot be guaranteed to protect the privacy of the subject. Somebody might recognize a voice, at the least. One natural consequence of this practice has been to assume that, since nobody else will ever see them, we need to transcribe or otherwise manage our own data just for our own purposes, say by selective marking of features of interest. Our colleagues in other scientific disciplines, however, routinely have to share their data and their methods to permit replication of experiments by other researchers as a method of validation of research findings. We therefore find ourselves not in step with one of the basic tenets of modern scientific research, and we pay the price of too often being called a "fuzzy" or "soft" science.

Today, however, funding agencies often expect as a condition of a grant that we will make arrangements to keep our data over the long term so that our results can be consulted and validated by other researchers, and IRBs now recognize the need for us to do so by setting standards for "data repositories." As long as we are keeping our interviews, and keeping them safely to protect our research subjects, we should go one step further: we should give them all away. There are a number of issues we have to consider if we want to make our data available to a wider public (developed in detail in Bounds, Palosaari, & Kretzschmar, 2010; Kretzschmar et al., 2006; Kretzschmar & Potter, 2010), but these can all be covered under just two headings, audience and archiving. How we make the data available has a single answer: online via the internet.

First of all, we need to consider who will see our data if we do give it all away. We will all have two audiences, a general audience and a specialist audience. General users are interested in our interviews, at the most basic level, for the same reason that reality television is popular: we all like to see how other people live and think and, especially for our material, to hear how other people talk. To confirm this, we need only consider the proliferation of websites offering English dialect or accent samples (IDEA, AUE, Speech Accent Archive, and so on), most often not collected by sociolinguists or dialectologists. This natural curiosity may be enhanced for many users by a special interest in oral history or community culture, or by a particular interest in a special topic. This means that we should cater to the general audience by providing basic information about our research subjects so users can characterize what they hear, by providing our data in small pieces so that users do not have to wade through an entire interview to get to the parts they want to hear, and by providing finding aids so that users can locate the pieces they most want to hear. Specialists have these same requirements but in addition, in order to analyze the data to validate their own results or to make a new study, they need to be able to search the data for specific features, and they need to know how the data were collected and processed. We have designed our new Linguistic Atlas website (Kretzschmar, n.d.) using the LICHEN program, which we developed in collaboration with colleagues in Finland: the issue of making public corpora is international. LICHEN does not just have fixed means of access for users but is a "toolbox" that allows us to offer different tools for access and processing information that cater to different audiences.

Neither general users nor specialists are entitled to know personally identifying details about our research subjects or to hear personal or sensitive information that our subjects may have been incautious enough to talk about on a recording – no matter how interested in these things the users may be. Our research subjects themselves should understand, according to the principle of informed consent, that their interviews will be accessed by a wider public, and they should acknowledge such uses in advance. The following statement is the kind of agreement that subjects might sign to indicate their understanding:

> The interview obtained from you in this research may be used to conduct the research identified above and may be used as you have initialed above. Your interview may be made part of a larger repository to be accessed by researchers, possibly for a fee. There are no plans to provide financial compensation to you should this occur. By consenting to participate, you authorize the use of your interview, photographs, and likeness, and any of your copyright in those items, for the research described above.

According to funding agencies such as, in the United States, the National Endowment for the Humanities, distribution of grant products should occur for the "cost of reproduction," so nobody will make money from any small fees involved in reproducing WAV files for technical processing (see below; of course access to files online should be free), and our research subjects should understand that. The first audience for an interview must always be the person who grants it in the first place (or who thinks better of it and retracts the interview), and it is good practice to give our research subjects copies of what we produce from their interviews. In another kind of benefit, the families of some older subjects for our Roswell Voices projects have especially appreciated having recordings of their parents and grandparents, when our subjects subsequently passed away. Older interviews from before it was common to follow IRB human subjects guidelines are "exempt" from the informed consent principle, but there is no reason not to apply the same ethical guidelines about identifying details and sensitive information when we make older interviews accessible; doing the right thing is still right, even when we are not required to do it by an IRB.

The way we archive our materials should address these things directly. We need to keep original, unedited, and unfiltered copies of sound and video interviews for posterity, so that they can later be processed according to the unforeseen standards of some future time; these copies can be maintained in a "dark" archive, one with no public access. University librarians can help us out with this, following the new "institutional repository" movement in which libraries take an interest in archiving research products from their institutions. The University of Georgia library has "adopted" the Linguistic Atlas Project, but we have also been able to pass along grant funding to help the library acquire equipment for the purpose – as the song says, "God bless the child who's got his own." Material released to the public and to specialists needs to have all identifying details and sensitive information removed, "beeped out" of sound files (i.e., replaced by a tone of the same length as the material removed, using sound editing software) and not transcribed in written versions. Besides protecting their privacy, giving this kind of "public" version to our research subjects can increase their confidence in our good faith. Sound files in WAV format are too large to be distributed well on the web. We should all keep files in "lossless" uncompressed WAV format, suitable for acoustic phonetic processing, but smaller audio files, in segments lasting a few minutes in "lossy" compressed format (the new open-source standard is OGG, not MP3), can be made available safely and conveniently on the web – just like song downloads. Our new Atlas website offers over 100,000 of these approximately four-minute files, extracted from our long interviews. Besides maintenance of original multimedia files, full-text orthographic transcriptions should be made as soon as possible, with minimal editing (e.g., respelling to indicate pronunciation) or coding (e.g., special marking of features we are interested in), in order to serve as an index to the sound files. Our wider audiences need a clear, intelligible, consistent written record of an interview for their own purposes and should not have their access to the interview hampered by interpretations we may have placed on it for our own (Barry, 2008). Along with transcripts, we need to provide two sorts of metadata: information about our research subjects and information about how we collected and processed the data. Without such information, our data may eventually become unusable because future researchers will not be able to categorize subjects appropriately for their own purposes – which is already the case with interview data like those from the US non-profit organization StoryCorps or other oral history projects not managed by sociolinguists – or, as has already happened with some historical Linguistic Atlas material, future researchers may not be able to decode how the data were recorded and processed, so we at length decided not even to keep it in our archives. Guy Lowman had created an analysis comparing his southern England interviews to his American Atlas interviews, which would have provided an excellent comparison of British and American English, but even after years of trying, we could not decipher the way he had annotated and tabulated the material. Lowman, of course, was not the only one to develop his own idiosyncratic way of working with his materials without leaving a key, and none of us can afford that any longer if we keep future users in mind.

The web is really the only choice for wide distribution of our data. It reaches both the general and the specialist audiences. The web allows specialists to discover material relevant to their work, whether through an institution like the Open Language Archives Community (OLAC) or just by searching online; they can then make separate arrangements to get WAV files if they need them, to supplement the versions of files made available online. Storage of data for distribution online also helps with maintenance of archives, because it is much easier to keep copies of digital files on disk than to preserve analog media like tape or physical digital media such as tape or CDs. The act of giving all of our data away, ironically enough, makes the preservation of our data more secure.

References

Barry, B. L. (2008). Transcription as speech-to-text data transformation. (Unpublished doctoral dissertation.) University of Georgia, Athens, GA.

Bounds, P., Palosaari, N., & Kretzschmar, W. A., Jr. (2010). Issues in using legacy data. In M. Di Paolo & M. Yaeger Dror (Eds.), *Sociophonetics* (pp. 46–57). London: Routledge.

Kretzschmar, W. A., Jr. (Ed.). (n.d.). Linguistic atlas projects. Retrieved from http://www.lap.uga.edu

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*, 172–205.

Kretzschmar, W. A., Jr., & Potter, W. G. (2010). Library collaboration with large digital humanities projects. *Literary and Linguistic Computing, 25*, 1–7.

Vignette 12b
Establishing Corpora from Existing Data Sources

Mark Davies

Corpora are searchable collections of spoken and written language (nearly always in electronic format) which can be used for linguistic analysis. Ideally, the texts come from sources where, at the moment of speech or writing, there was no understanding that the materials would later be used in a corpus for linguistic analysis, since this helps to preserve the "naturalness" of the language.

Until recently, the largest publicly available spoken corpus was the 10 million words of spoken English in the 100-million-word British National Corpus (BNC). Other important corpora of spoken English are the Cambridge and Nottingham Corpus of Discourse in English, the Cambridge North American Spoken Corpus, the Santa Barbara Corpus of Spoken American English, the Switchboard corpus, and the CallHome corpus. Unfortunately, with the exception of the BNC, most of these corpora either are not publicly available (they are just used for in-house materials development) or are prohibitively expensive for most researchers (costing $1,000 or more).

Because of the issues with pricing and (lack of) availability, some researchers might consider creating their own corpora. Unfortunately, it is almost prohibitively difficult for individual researchers to create large spoken corpora "from the ground up." It takes a very long time and a great deal of money to design a corpus and find speakers, record the speech, and (especially) to carefully transcribe the speech and then revise and correct the texts. The only reason it could be done in the cases of the spoken corpora listed above is that there was typically a large research team and robust funding for creation of each corpus.

As a result, the most realistic alternative for most researchers is to create corpora from existing resources. This was, for example, the process that was followed in the creation of the Corpus of Contemporary American English [COCA] (Davies, 2009; 2011). Although this is the largest and most up-to-date publicly available corpus of English, it was created by just one person in less than a year. The corpus contains 425 million words of text (including 85 million words of spoken language – eight to nine times the size of the spoken portion of the BNC). And unlike any other corpus of English, COCA continues to be expanded: 20 million words of text (including four million words of spoken English) continue to be added each year. The remainder of this vignette will focus on some of the issues raised in the creation of COCA from existing resources, for the benefit of others who might want to follow a similar path.

Perhaps the most obvious question is what resources to use to find spoken texts. One approach might be to consider texts that are not actual "spoken texts" per se but attempt to model natural spoken language, including scripts for television series, radio, movies, and plays. COCA has more than 15 million words of text from these types of sources, and there are probably hundreds of millions of words of text from such resources freely available online. (For example, in just a day or two we created another 70-million-word corpus of scripts from US soap operas.) The question with these "pseudo-spoken" texts, however, is how closely they in fact represent actual spoken language. Many different phenomena in COCA – lexical, phraseological, and syntactic – show that while these scripts are probably the most "spoken-like" of all of the non-spoken genres (fiction, magazines, newspapers, and academic journals), there is still a noticeable difference between these texts and those from actual spoken English. (For this reason, these texts are categorized as "Fiction" in COCA.)

Another possibility might be to find interviews online, as we did while we were compiling the spoken component of the 100-million-word Corpus del Español and the 45-million-word Corpus do Português. There are at least three issues involved in using these resources, however: First, some of the texts that are the easiest to find come from speech types that are probably overly formal, such as political press conferences, and that may only slightly resemble natural, conversational speech. Second, there is a question of how much post-interview "editing" and cleanup has already been done to the texts to eliminate things like hesitation, false starts, and backchanneling. Third, creating such a corpus may involve a great deal of manual editing to extract the interviews from thousands of web pages on hundreds of websites, each with its own formatting for headers, footers, ads, and comments.

Recognizing these limitations, perhaps the best source for spoken language are the transcripts of unscripted speech on television and radio programs such as *Oprah*, *Jerry Springer*, *Geraldo*, *Good Morning America* (ABC), *60 Minutes* (CBS), *Larry King Live* (CNN), or *All Things Considered* (NPR). As I have already noted, more than 85 million words of speech from such resources were used in the creation of COCA, and these 85 million words of spoken data are just a small fraction of what is available online. For example, CNN alone has freely downloadable transcripts of all of its programs from the past 10 years or so, representing more than 250 million words of text.

These transcripts typically do not have the shortcomings of some of the formal interviews discussed above. First, the transcripts cover a wide range of speech types and topics, such as interviews with politicians, actors, or sports figures, or discussions about parenting, hairstyling, new electronic devices, or any number of other topics. This means that the vocabulary is quite diverse, and the style is more informal and natural than that used in press conferences and similar speech types. Second, the transcripts used in COCA have minimal editing to remove features such as hesitation and backchanneling. Third, the page format for all of the tens of thousands of transcripts is typically the same or quite similar, which reduces the problem in processing the texts.

There are two limitations with such transcripts, however. The first concerns the naturalness of the language. The speakers knew that they were on national

TV or radio and were therefore probably on guard to avoid non-standard features like double negation (*She doesn't have no reason*), double modals (*They might could do it*), lexical items and constructions like *ain't* or *had went*, and profanity (which would, in any case, be censored by the television or radio program). Nevertheless, as is discussed in Davies (2009), for most linguistic phenomena these transcripts still model normal everyday conversation quite well. For example, colloquial features like quotative *like* (*He was like, I'm not going with her*), *so not* ADJ (*She's so not interested in him*), or even the common *you know* (*He's, you know, kinda worried about her*) are much more common in the spoken data in COCA than in the data from other genres (fiction, popular magazines, newspapers, and academic journals).

The second concern about using transcripts is the difficulty in coding them for demographic information, e.g., age, gender, ethnicity, or socioeconomic status. There are more than 40,000 spoken transcripts in COCA, with at least two and perhaps as many as 10 or 20 speakers in each transcript, and someone would need to find demographic information for each of the hundreds of thousands of speakers. For lesser-known participants on these programs, this would likely not be possible, and even for those where it is possible, it would be extremely time-consuming (perhaps 25,000-plus hours) and very expensive (hundreds of thousands of dollars). For a smaller corpus (e.g., 100,000–1,000,000 words), it might be possible to code for speaker variables, but then the corpus might only be large enough for it to be possible to look at very frequent linguistic phenomena, such as discourse markers or very high-frequency grammatical constructions.

One way around this problem of sparse demographic coding would be to focus on comparing the different television and radio programs, rather than all of the speakers on these programs. Obviously, this would not give the level of demographic encoding that most sociolinguists are accustomed to, but it is likely the only possibility for large corpora that are created from existing resources. For example, one could easily and quickly create a 5-million-word *Oprah* or *Jerry Springer* corpus (which is presumably fairly informal) and compare it to a 5-million-word corpus containing more formal conversation on a program like *Face the Nation* or the *Newshour* on PBS, with perhaps an intermediate corpus from programs like *All Things Considered* or *Good Morning America* in the mix as well.

In summary, there are a wide range of sources that are publicly available, which allow researchers with even very limited funds and personnel to create very useful corpora of contemporary (spoken) language.

References

Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–190.

Davies, M. (2011). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*, 447–465.

Vignette 12c
Working with "Unconventional" Existing Data Sources

Joan C. Beal and Karen P. Corrigan

In this vignette, we share our experience of working with data collected at various times and according to varying methodologies to create the Newcastle Electronic Corpus of Tyneside English (NECTE) (for a fuller account, see Allen et al., 2007). In creating this corpus, we faced a number of challenges, some of which required us to devise new policies and protocols, albeit with advice from colleagues. Given the endeavors of sociolinguists working in the pre-digital age, there must be many important and useful collections of data languishing in cupboards, on shelves, or even under beds. We hope that this account of our experiences will inspire readers to rescue these data from "shedding the hard-won sounds of 20th-century speech in the constantly dispersing particles of ferric oxide of an obsolescent recording system" (Widdowson, 2003, p. 84).

The primary data behind NECTE were collected by two teams of sociolinguists, one working in the late 1960s and early 1970s on the Tyneside Linguistic Survey (TLS) (see Pellowe, Strang, Nixon, & McNeany, 1972, for the TLS methodology) and the other in the 1990s for the Phonological Variation and Change (PVC) project. The latter dataset posed fewer problems for the NECTE team, since it had been collected using what are still considered state-of-the-art methods and recorded in digital format (see Milroy, Milroy, & Docherty, 1997). We therefore concentrate on the challenges involved in processing the TLS data.

The first challenge was to find as many of the data and accompanying metadata as possible. The majority of the data had been left in the department of Newcastle University where the TLS team had worked. Unfortunately, the materials were not stored in controlled archival conditions but rather in unlabeled boxes in store-cupboards, in serious danger of deterioration. More data came to light only after our project began, when a former member of the TLS team brought back some recordings and index cards he had taken with him upon relocating. Although the original TLS data collection was carried out in accordance with the principle of random sampling, the NECTE team did not inherit the original random sample in this technical sense and instead inherited ad hoc remnants of it. Nevertheless, a majority of the interviews were, in fact, preserved. Moreover, the richness of the social data collected by the TLS team has ensured that NECTE users can make up their own balanced sample from the available material, as has already been done in publications such as Beal and Corrigan (2005a; 2005b). More recently, Barnfield and Buchstaller (2010) have sampled

NECTE alongside data from new interviews in the region collected since 2007 during the creation of the more recent instantiation of the corpus, the Diachronic Electronic Corpus of Tyneside English (http://research.ncl.ac.uk/decte/). In dealing with legacy materials, we really are making the best use not of "bad" data (Labov, 1994, p. 11) but of imperfect data.

The next challenge was compliance with 21st-century standards of ethics and data protection. The TLS researchers in 1969 had no internal ethics review board to satisfy, but their funding body (the Social Science Research Council) did have an ethics policy in place, so there was evidence that the subjects had indeed given informed consent to being recorded and to the recordings being made available to future researchers. These subjects (or indeed the researchers themselves) could, however, have had no idea that there would one day be such a thing as the internet and that the recordings might be available to anybody in the world at the click of a mouse, so great care had to be taken to preserve their anonymity by removing all names from recordings and transcripts and creating a table of names and ID codes accessible only by the project team and securely stored. There was still the question of whether a voice is ever truly anonymous, so restrictions had to be placed on access to NECTE to prevent the casual web surfer from happening upon the data. We are aware that these safeguards militate against open access, so the new additions to the corpus from 2007 onward use data for which full consent has been obtained.

Transcribing and tagging the data provided further challenges, as explained more fully in Beal, Corrigan, Smith, and Rayson (2007). Acting on advice from leading sociophoneticians, we decided against providing IPA transcriptions, but even orthographic transcription was not unproblematic. Given the objections to semi-phonetic spelling raised by, for example, Preston (1985; 2000), we chose to use standard British English spelling throughout, except where an item had no lexical or morphological equivalent in Standard English. Thus, the word *know* is transcribed <know> whether it is pronounced /na:/ as in traditional Tyneside Engish or /nou/ as in RP, rather than being written <knaa> as in dialect literature. However, the negation of *do* as /divnt/ is transcribed as <divvent> because it is morphologically distinct from *don't*. There are also many lexical items in the NECTE corpus that do not exist in Standard English, such as *gadgie* 'old man', *bairn* 'child', and *varnigh* 'nearly'. For these, we created a list of agreed spellings, based on entries in dialect dictionaries wherever possible (see www.ncl.ac.uk/necte/appendix2.htm).

With regard to tagging, we faced the challenge of adapting tagging software such as CLAWS, originally designed for use with corpora of Standard English such as the British National Corpus, to encode a corpus of Tyneside English, which has a number of distinctive dialectal morphological and syntactic features (see Beal, 1993, for an account). Fortunately, Paul Rayson and Nick Smith were able to adapt CLAWS8 by adding tags such as the following:

- pronouns, e.g., *wor* 'our' (= possessive form of personal pronoun); tagged APPGE;
- *mesel*, *hisself*, *theirself*, *theirselves*, etc. (= reflexive personal pronoun); tagged PPX1 or PPX2;

- auxiliaries: *div* = a regional variant of the auxiliary do, non-third singular present tense; tagged VAD0.

The NECTE data also include a wide range of discourse markers such as *wey*, *man*, *like*, *aye*, *well*, *uhhuh*, *huh*, *ah, you know*, and *I mean*. CLAWS8 had no existing tag for such features, so instead we used the tag for interjections, UH. This pragmatic decision allows a researcher interested in such discourse markers to identify them easily. The only features of Tyneside English morphology that could not be identified with tags were those with a surface form identical to a different morphological item in Standard English, such as *went* as past participle (*if I'd went*), *give* and *come*, *seen* and *done* as preterits, and *we* as first person plural object pronoun (*She sent we*). However, these could be identified in context, and any researcher interested in Tyneside English would be able to locate them. The main lesson we learned from the tagging exercise was that by seeking expert help, seemingly difficult problems can be overcome.

The final challenge for the NECTE team was to "future-proof" the corpus. Of course, we cannot predict the "shelf life" of digital data, but encoding NECTE in XML at the very least ensured that it would work on all platforms and with all software applications for the foreseeable future. At the time, this was a pioneering move, and we were concerned that some users would find XML too "unfriendly" a format. It does not, however, appear to be that problematic: the NECTE website has since been accessed successfully by a wide range of users, as the snapshot in Table 12c.1 demonstrates, and download or DVD requests remain frequent despite the fact that the materials were originally released back in 2005. Moreover, the fact that DVD copies of NECTE have been distributed to a large number of users, and that we have records of these, means that, in the event of some catastrophic loss of data, we could call upon these colleagues to retrieve them.

The NECTE corpus project was challenging, but it taught us that unconventional and imperfect data can nevertheless be transformed into a usable corpus, enabling researchers to mine these data in ways that could not have been envisaged

Table 12c.1 Snapshot of Website Activity, March 31, 2009–October 7, 2009

Type	*Number*
Total hits	**28,323**
Visitor hits	12,909
Spider hits	15,414
Average per day	148
Average per visitor	4.85
Total page views	**4,381**
Average page views per day	22
Average page views per visitor	1.64
Total visitors	**2,664**
Average visitors per day	13

by the scholars who originally collected them. We hope that this vignette will inspire readers to create similar resources, no matter how unconventional the data involved.

References

Allen, W. H. A., Beal, J. C., Corrigan, K. P., Maguire, W., & Moisl, H. L. (2007). A linguistic time capsule: The Newcastle Electronic Corpus of Tyneside English. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora*, Vol. 2: *Diachronic databases* (pp. 16–48). Basingstoke, UK: Palgrave Macmillan.

Barnfield, K., & Buchstaller, I. (2010). Intensifiers on Tyneside: Longitudinal developments and new trends. *English World-Wide, 31*, 252–287.

Beal, J. C. (1993). The grammar of Tyneside and Northumbrian English. In J. Milroy and L. Milroy (Eds.), *Real English: The grammar of English dialects in the British Isles* (pp. 187–242). London: Longman.

Beal, J. C., & Corrigan, K. P. (2005a). A tale of two dialects: Relativization in Newcastle and Sheffield. In M. Filppula, M. Palander, J. Klemola, & E. Penttilä (Eds.), *Dialects across borders* (pp. 211–229). Amsterdam: John Benjamins.

Beal, J. C., & Corrigan, K. P. (2005b). "No, nay, never": Negation in Tyneside English. In Y. Iyeiri (Ed.), *Aspects of negation in English* (pp. 139–156). Amsterdam: John Benjamins.

Beal, J. C., Corrigan, K. P., Smith, N., & Rayson, P. (2007). Writing the vernacular: Transcribing and tagging the Newcastle Electronic Corpus of Tyneside English (NECTE). In A. Meurman-Solin & A. Nurmi (Eds.), *VARIENG: Studies in variation, contacts and change in English I: Annotating variation and change*. Proceedings of the ICAME 27 Annotation Workshop. Retrieved from http://www.helsinki.fi/varieng/journal/volumes/01/beal_et_al/

Labov, W. (1994). *Principles of linguistic change: Internal factors*. Malden, MA: Blackwell.

Milroy, J., Milroy, L., & Docherty, G. J. (1997). Phonological variation and change in contemporary spoken British English. ESRC, unpublished final report, Department of Speech, University of Newcastle upon Tyne.

Pellowe, J., Strang, B. H. M., Nixon G., & McNeany, V. (1972). A dynamic modelling of linguistic variation: The urban (Tyneside) linguistic survey. *Lingua, 30*, 1–30.

Preston, D. R. (1985). The Li'l Abner syndrome: Written representations of speech. *American Speech, 60*(4), 328–336.

Preston, D. R. (2000). Mowr and mowr bayud spellin': Confessions of a sociolinguist. *Journal of Sociolinguistics, 4*, 614–621.

Widdowson, J. D. A. (2003). Hidden depths: Exploiting archival resources of spoken English. *Lore and Language, 17*(1&2), 81–92.

13 Working with Performed Language

Movies, Television, and Music

Robin Queen

As I write this chapter, a language story is circulating concerning "vocal fry." Tied to a paper to be published in the *Journal of Voice*, one of the central claims of the paper is that the use of vocal fry (or creak) has become more prevalent among younger (college-aged) women in the United States. In the discussion section of the paper, the authors surmise that young women are using vocal fry to mimic popular figures. This story was picked up by many different news and entertainment media organizations, and the tangential link to popular figures was elevated to a central point, as evidenced by titles such as "More college women speak in creaks, thanks to pop stars" (Dahl, 2011, p. 12). While there are many interesting components to this story and the media's reaction to it, in this chapter I draw attention to the idea that it is pop stars who are the drivers of this supposed innovation. In particular, this story encapsulates a fundamental question for working with performed language concerning the relationships between "real-world" linguistic variation and similar variation as it occurs in various media channels.

In the remainder of the chapter, I discuss what performed media are, the ways in which performed media can be a good source of data about language, and some of the places where performed media represent special challenges. In considering these questions, two main areas that require consideration remain in central focus. Working with performed media differs from working with other kinds of sociolinguistic data, first in terms of how the researcher theorizes the data, and second, how the researcher selects and organizes the data, including managing matters linked to copyright and "fair use."

Linguistics and Performed Language

As illustrated by the discussions of vocal fry, the idea that the media have a profound and fundamentally detrimental effect on language exemplifies one of the common assumptions made by non-linguists about language in the mass media. Liberman (2011) notes that concerns about how the media (and in particular new media technologies that facilitate mass communication) might affect language go back at least as far as the mid-19th century and no doubt further than that (see Milroy & Milroy, 1999). Linguists, on the other hand, have traditionally been fairly skeptical concerning the performed media as a source of interesting information about language variation. Chambers (1998), for instance, argues that there is virtually no effect of the media on language or language use (see also Labov, 2000).

The majority of studies that have used media data for thinking about linguistic variation have focused on print and broadcast news media or other forms of largely unscripted media such as talk shows, sportscasts, and broadcasts of live events. The assumption behind these studies, stated or not, seems to be that these types of media products are in some sense a more accurate reflection of "real" language than is the case for scripted, or performed, media forms (a topic dealt with in this chapter as well as by Weldon, Vignette 13a, and Adams, Vignette 13b). The fact that roughly 80% of the scholarship on language variation in the mass media is drawn from unscripted media sources can largely be explained by the assumption that performed language is less authentic than either unscripted language in the media or real-life communication (see Coupland, 2007, for a discussion of authenticity in the media). Thus, one of the primary considerations for working with performed media language concerns thinking carefully about how to frame the questions that can be answered with such data, given the various constraints that performed media present. A second consideration concerns delineating the data for analysis in a way that makes the analysis both motivated and manageable. Once those considerations are accounted for, the basic methods of analysis do not differ significantly from other forms of sociolinguistic study.

Performed Media

Before we move to some of the details of managing these two major considerations for working with performed language, it is worth considering what constitutes performed media language. At some level, of course, language itself is fundamentally a medium, an intermediary between my thoughts and yours. In the more conventional sense, and in the sense being used here, media are an intermediary, or a channel, between particular content and its audience. For instance, broadcast media are channels through which content is transmitted electronically to an audience that is not in the same place as the content. Print media, on the other hand, are channels in which content is delivered via some kind of material object, such as a book. Finally, news media are channels through which specific content – news – is delivered. These can include both print and broadcast channels.

All language in the media is primarily performed, or representational, in that it does not present "real-life," face-to-face conversation, the data most sociolinguists rely on for their research. That fact notwithstanding, a sizable body of research draws on data from the media. These studies, however, are largely taken from media that represent actual events (sometimes even as they unfold, as in live sportscasts; see Reaser, 2003) or that represent actual conversations (as found in talk shows or political broadcasts; see Carbaugh, 1990; Hay, Jannedy, & Mendoza-Denton, 1999; Mendoza-Denton & Jannedy, 2011). It is therefore important to draw a distinction between media linked to information and media linked to imagination. A documentary frames a narrative in particular ways and for particular purposes, but the people and events being represented are fundamentally representations of themselves (however perspectival). As discussed in Reaser and Adger (2007), for example, the documentary *Do You Speak American?* seeks to represent the actual linguistic variation that exists in the United States.

Imagined, or performed, media in the sense being used for this chapter involve people who are representing someone other than themselves, usually someone fictional or historical. I assume performed media to be tied largely to fictionalized content, though there is no clear-cut line for unambiguously distinguishing between fiction and non-fiction. For instance, virtual online communities, such as Second Life, or multiplayer video games, such as World of Warcraft, have non-fictional as well as fictional components (see Sadler, Vignette 3d). Reality television presents another genre of popular representation that is difficult to clearly delineate as fictional or non-fictional. For the most part, however, the term "performed language" refers to some kind of fictional representation.

A second distinction that is worth mention within the context of performed media relates to the mass nature of much media. Canonically, most people consider mass media when they think of "the media." Examples of mass media include broadcast, print, and electronic forms of content distribution that are available simultaneously to a broad range of recipients, regardless of those recipients' geographic or temporal location. These media are also typically either commercial in nature or supported through public funding such as taxation, something important to bear in mind. Non-mass forms of media are not designed or intended for a broad and largely unknown audience. They include personal letters, telephone calls, and texting. The distinction between mass and non-mass is largely idealized and gradient rather than specific and categorical, a characteristic made especially obvious in language that is somehow mediated through the internet.

The final distinction worth considering is between edited and unedited (or scripted and unscripted) mass media. As with the distinctions between information/imagination and mass/personal, the distinction between scripted and unscripted serves to define two edges of a continuum. The distinction matters because of the complex relationships between the writers, actors, and characters involved in producing scripted, performed language. At some level, characters are not themselves animators of language, since they are always the product of an author's, an actor's, a director's, and an audience's imagination. As Richardson (2010) writes,

> [television] repeatedly displays people talking, showing audiences how characters behave in the varying circumstances of their narratives. These stories, and the talk they give rise to, mediate between the familiar and the extraordinary, and engage the imaginative powers of their receivers as well as their creators.
>
> (p. 3)

Given the process by which it emerges, linguistic interaction is quite a bit different in the scripted media than in non-scripted media or in non-mediated interaction, and many of the features of spoken interaction, such as disfluencies and certain styles of interruption, are uncommon. In my own work (Queen, 2012), I studied a corpus of data drawn from daytime dramas in which the primary mode of activity is face-to-face communication among the characters and found very little conversational overlap or interruption.

The language in question for the rest of this chapter can thus be captured as language that occurs in mass media productions that are primarily scripted and emerge, at least initially, from a writer's imagination. The three major types of media in which this kind of performed language occurs are in music, television, and film. Music of course differs from television and film in that it is generally not narrative in the same ways that television and film are. At the same time, music is very much the product of a songwriter's imagination and is of course highly scripted (see Coupland, 2011). While both television and film represent media channels that frequently involve scripted, fictional representations, they also have specific differences that bear on the kinds of questions that can be asked of data drawn from them. Most critically, films generally represent no more than three hours of connected narrative. The narrative on television shows, on the other hand, can develop over the course of several months or years. These differences are nicely illustrated in Vignettes 13a and 13b, by Weldon and Adams respectively, which follow this chapter.

Framing Performed Media

Having discussed what performed media language is, I turn to some of the real benefits of using performed media materials. Richardson (2010) states:

> In mainstream television and film, there is a preference for realistic rather than stylized or poetic modes of talk in many genres. This approximate, and conventionalized, verisimilitude encourages audiences to take the easy road and hear *drama talk* as they hear everyday talk.
>
> (p. 5)

In other words, the mass media provide a vehicle for diffusing linguistic forms, particularly those that carry social prestige of some sort or another, far from their original source (Spitulnik, 1996). For the student of linguistic variation, the mass media thus offer a ready-made sandbox for exploring theoretical topics that span the range of sociolinguistic theory, but particularly those theoretical concerns tied to linguistic ideology (Lippi-Green, 2012), linguistic styles and stylization (Coupland, 2007; Queen, 2004; Richardson, 2010), and linguistic indexicalities (Beal, 2009; Bucholtz & Lopez, 2011; King & Comeau, 2011; Lippi-Green, 2012). The question of enregisterment is also frequently linked to data that circulate through the mass media (Agha, 2007; Bell, 2011; Johnstone, 2011; Squires, 2011). Fictional media can directly address ideologies of language, mainly as they relate to the indexical associations broadly assumed to hold in a community between types of people and how they speak. Similarly, fictional television characters frequently stand in for attitudes and values and are contrasted with each other along lines of social value (Bednarek, 2011, p. 12).

Fictional and otherwise scripted media can further provide a source of data against which to compare "real" language. Simpson (2003) reports that sample sentences used to illustrate word meanings in the *Oxford English Dictionary* (OED) frequently draw on first mentions within the media. For instance, the meaning 'great' for the word *magic* was first recorded in the script for the film *The Long Arm* (p. 190). Similarly, the OED often cites music lyrics as the textual basis for particular definitions. Tagliamonte and Roberts (2005) examine the use of different intensifiers in eight seasons of the television series *Friends*. They illustrate both that the patterns of intensifier use on *Friends* mirror those found in corpora of spoken English and that some aspects of intensifier use reveal innovations along predictable lines. Quaglio (2009) shows that the core linguistic features characteristic of an involved register such as face-to-face conversation are also characteristic of interactions on *Friends*. At the same time, face-to-face interaction is considerably more varied and includes elements, such as expletives, that are not found in the *Friends* corpus.

This particular pattern captures one of the broader differences between real-life linguistic interaction and that found in the scripted media, namely the lack of the full range of culturally structured variation. For instance, regional forms that are not necessarily stigmatized but that do not have broad indexical associations, such as the *needs* + past participle construction (*The dog needs fed*), are typically absent from performed media. Thus, the mass media offer a useful checkpoint for determining what linguistic characteristics within a linguistic community are stereotypes, markers, and indicators (Labov, 1972), such that we can predict that stereotypes will be prevalent, markers relatively rare, and indicators close to nonexistent in the performed media.

The mass media increasingly serve as a critical vehicle for the representation (and frequently the creation) of constructed languages (Chozick, 2011; Okrent, 2009). For instance, Dothraki was constructed for the television program *Empire of Thrones*, and Na'vi was constructed for the Oscar-winning film *Avatar*. The mixed language found in the television show *Firefly* consists of elements from both Chinese and English (Mandala, 2008). More famous examples include Klingon, created for the *Star Trek* series, and the various forms of Elvish constructed by J. R. R. Tolkien and animated in the *Lord of the Rings* trilogy. Of some interest here is not only the structure of the constructed language itself but also the ways in which a constructed language is mobilized as a quick index to Otherness, especially as found in non-human cultures. Interestingly, the actual use of the constructed language is characteristically fairly minimal (Elvish, for instance, occurs for less than 20 minutes across the over eight hours of the *Lord of the Rings* trilogy). In the case of Klingon, an active community of language users, as reported by Okrent (2009), provides excellent potential for an ethnography of speaking linked to a language constructed entirely for a media venue. While such communities may not be commonplace, they nonetheless illustrate a mechanism by which performed language becomes intertwined in real-life communities.

Using data drawn from performed language, especially music, provides an opportunity to see how performers use language as a vehicle for social justice. Alim's (2006) work on Hip Hop Nation Language presents a particularly salient example of how a focus on performed lyrics can highlight the tools artists use to combat discrimination and other forms of oppression via resistance. Lippi-Green (2012) similarly uses an analysis of a corpus of animated Disney films to highlight socialization into the dominance of the standard. In her analysis, Lippi-Green illustrates how Mainstream American English is typically the variety of English spoken by animated heroes, while animated villains are much more likely to use non-mainstream or non-American varieties. Further, sidekicks and secondary characters are much more likely to use non-mainstream varieties than are main characters. Such patterning helps socialize children to understand the social hierarchies tied to linguistic varieties. A similar pattern is also found in animated films not directed at children, though in those cases that hierarchy is much more likely to present a critique of such relationships than what is found in media directed at children. For example, King and Comeau (2011) show how the character Acadieman in the animated series *Acadieman* uses a local variety of Acadian French in order to valorize the experiences of young minority language speakers in New Brunswick.

Lastly, language variation in the performed media offers an excellent opportunity for application in the classroom. As Squires and Queen (2011) note,

> [b]ecause language is central to the way that characters are represented, actors are trained, and media are perceived, we believe that integrating mass media content with linguistics material – especially in sociolinguistics, but also across the discipline – is an appropriate way to engage students in a lively class while broadening their applications of course concepts.
>
> (p. 232)

For example, students are able to explore stereotypes, such as that gay men have both higher pitch and a wider pitch range relative to straight men. Through an analysis of the intonation of the sitcom character Jack on the television series *Will and Grace*, students were able to see that Jack's pitch did not differ significantly from Will's or from that of the straight male characters on the show, though there were other ways in which Jack was distinguishable linguistically from other characters.

The animated children's films mentioned also offer the possibility for students to explore various kinds of indexical associations between language variation and social types and persona. For instance, by examining the film *Cars*, students can see how language varieties, such as African American English, Chicano English, and Southern American English, are used to create characters and situations that illustrate concepts linked to indexical ordering such as presupposition and entailment (Silverstein, 1996). In this way, the media provide a means to help students connect to highly abstract theoretical concepts and see their everyday applications. In addition to examining strategies such as these, Squires and Queen (2011) also discuss selecting and managing a large corpus of performed language materials.

Turning Performed Media into Data

Most commonly, data for sociolinguistic research are gathered in the service of answering a particular research question and thus are, in a sense, specifically created as part of the research process. Even data gathered in the context of ethnographic fieldwork, which are less mediated by the research process than are data drawn from interviews, experiments, or surveys, are preserved explicitly for the purpose of addressing a research question. The opposite is the case for data drawn from the performed media. Language variation in the mass media becomes data long after it has been preserved. Thus, selecting data for analysis from the performed media is in many ways more serendipitous and thus requires careful consideration in terms of motivating the relative representativeness and specific selection.

The bulk of the existing literature using data from the performed media draws on a final product, namely the commercially released film, television program, or album. A relatively open area of research concerns the overall process by which language variation becomes a part of that final product. One intriguing possibility for analysis would involve documenting the emergence of language variation from an initial script or lyric through the various production stages and then in the final product. Adams, in Vignette 13b, compares language variation found in scripts and performed work, such as on the television series *Buffy the Vampire Slayer*, and Alim (2004) shows a truncated example of this type of process in his study of various individuals involved in the Hip Hop Nation.

Before the late 1990s, it was much more difficult to preserve media data for the purposes of analysis. Since the late 1990s, however, a range of technologies and new consumer markets has made it much easier to obtain performed media programming. For instance, it is a relatively trivial matter to obtain the entire run of a given television series, and much programming can be found at any time or place via streaming technologies available through the internet. Similarly, many different types of information about specific performed media outlets can be found relatively simply and quickly online. There is therefore a greater expectation for scholars to motivate the selection of performed media data being used for analysis. Thus, a researcher might use all the recorded output produced by a particular band (Beal, 2009; Clarke & Hiscock, 2009), a full run of a television series (Adams, 2004; Tagliamonte & Roberts, 2005), the entire oeuvre of a single actor (Bell, 2011), or a single season of a television series (Mandala, 2008; Richardson, 2010). Alternatively, a corpus of materials could be collected and then a specific set of criteria developed for selecting data for analysis. In my own study (Queen, 2004), I used this type of method to select data involving the dubbing of African American English, and Bucholtz and Lopez (2011) use it for the analysis of linguistic minstrelsy in Hollywood films. Lastly, Ensslin (2011) uses a survey method to select a group of video games for analysis based on a predefined typology of game genres.

Turning performed language into data typically involves some kind of transcription, and scholars must decide how much detail to transcribe. Most linguistic work based on data from the performed media has primarily relied on transcriptions of the language output alone. Increasingly, however, multimodal analyses call this practice into question. Generally speaking, most researchers engaged in qualitative research do not transcribe the entire set of performed language data being analyzed, while those working to quantify and statistically model their data do. Thus, an important early consideration for working with performed media involves the decision of how much of the data require transcription, and this decision will largely be tied to the research method and question being addressed. Fans have already transcribed many performed media products, which can be a useful starting point for turning the final media product into an analyzable set of data. For instance, Bedarnek (2011) discusses a searchable, fan-generated database of dialogue in 86 episodes of the series *Lost*. Similarly, the transcriptions used in Tagliamonte and Roberts (2005) as well as Quaglio (2009) were based on transcriptions of *Friends* available online. Once the data have been rendered into an analyzable format such as a transcription, they are no longer particularly different from any other kind of sociolinguistic data and can be analyzed along the same dimensions that real-life language can be analyzed.

The final area requiring special consideration for working with performed media data involves matters of copyright. In the United States, for example, researchers using performed media materials must be mindful of keeping the use of primary analytic material within the realm of Educational Fair Use copyright restrictions. All media materials I have used for analysis have been restricted to video and audio clips in which the source material (DVDs and CDs) is owned by either my university or me. Personally, I do not use material whose copyright status or publisher is unclear. A researcher may legally capture material; however, circumventing copy protection on a media product violates the (US) Digital Millennium Rights Act. Some universities and scholars have received exemptions to use clips for educational purposes, but as of the writing of this chapter, such exemptions are not available for the purposes of publication. Thus, care should be given to matters of copyright when considering capturing performed media data for the purpose of analysis.

The Society for Cinema and Media Studies (2012) has published guidelines that relate to fair use practices when using media materials for scholarly and educational purposes. According to the Digital Millennium Copyright Act, fair use allows the use of copyrighted material without asking permission from the copyright holder when the material is being used for criticism, scholarship, and education. In general, US case law concerning fair use has considered the question of whether the work was used for transformative purposes (for instance, transcribing parts of a film to illustrate beliefs about African American English as compared to providing entertainment for a commercial audience) and whether the user draws on only as much of the original work as necessary to fulfill the transformative goal. In terms of publishing such materials, there is a great deal of variation among academic presses: some assume that scholarly use of copyright materials is acceptable without securing permissions, while others require authors to secure permissions. Mainstream media companies have often been prohibitive about the use of any portion of a copyrighted work, and this prohibition has become stronger as digital reproduction, both legal and illegal, has become more widespread. The publication of short transcriptions has generally not been subject to legal scrutiny, except for cases involving transcriptions purchased from companies providing contract services to media producers. However, any scholar considering the use of these materials, particularly still images, should ensure that the publisher will permit it.

Conclusion

In this chapter, I have presented some of the challenges and many of the benefits of working with data from the performed media, particularly music, television, and video. Generally speaking, the primary challenges of working with such data concern managing the selection, capture, and presentation aspects of the media. A further challenge concerns the theoretical framing of the data and the highly circumscribed nature of the research questions that can be asked, given the nature of performed media. A full consideration of the various agents involved has, to date, not been undertaken but is clearly a fundamental component of understanding more about the relationship of performed media language to the language of everyday "real life." This consideration is of special importance given the fact that those who create language variation differ from those who animate it. On the other hand, a wealth of exciting and intriguing questions can be addressed using data drawn from the performed mass media.

References

Adams, M. (2004). *Slayer slang: A* Buffy the Vampire Slayer *lexicon*. New York: Oxford University Press.

Agha, A. (2007). *Language and social relations*. Cambridge: Cambridge University Press.

Alim, H. S. (2004). *You know my steez: An ethnographic and sociolinguistic study of styleshifting in a Black American speech community*. Durham, NC: Duke University Press.

Alim, H. S. (2006). *Roc the mic right: The language of hip hop culture*. New York: Taylor & Francis.

Beal, J. C. (2009). You're not from New York City, you're from Rotherham. *Journal of English Linguistics, 37*(3), 223–240.

Bednarek, M. (2011). Expressivity and televisual characterization. *Language and Literature, 20*(1), 3–21.

Bell, A. (2011). Falling in love again and again: Marlene Dietrich and the iconization of non-native English. *Journal of Sociolinguistics, 15*(5), 627–656.

Bucholtz, M., & Lopez, Q. (2011). Performing blackness, forming whiteness: Linguistic minstrelsy in Hollywood film. *Journal of Sociolinguistics, 15*(5), 680–706.

Carbaugh, D. (1990). *Cultural communication and intercultural contact*. Hillsdale, NJ: Lawrence Erlbaum.

Chambers, J. K. (1998). TV makes people sound the same. In L. Bauer & P. Trudgill (Eds.), *Language myths* (pp. 123–131). New York: Penguin.

Chozick, A. (2011, December 11). Athhilezar? Watch your fantasy world language. *New York Times*. Retrieved from http://nytimes.com/2011/12/12/arts/television/in-game-of-thrones-a-language-to-make-the-world-feel-real.html

Clarke, S., & Hiscock, P. (2009). Hip-hop in a post-insular community: Hybridity, local language, and authenticity in an online Newfoundland rap group. *Journal of English Linguistics, 37*(3), 241–261.

Coupland, N. (2007). *Style: Language variation and identity*. Cambridge: Cambridge University Press.

Coupland, N. (2011). Voice, place and genre in popular song performance. *Journal of Sociolinguistics, 15*(5), 573–602.

Dahl, M. (2011, December 12). More college women speak in creaks, thanks to pop stars. *The Body Odd*. Retrieved from http://bodyodd.msnbc.msn.com/_news/2011/12/12/9393348-more-college-women-speak-in-creaks-thanks-to-pop-stars

Ensslin, A. (2011). *The language of gaming*. New York: Palgrave Macmillan.

Hay, J., Jannedy, S., & Mendoza-Denton, N. (1999). Oprah and /ay/: Lexical frequency, referee design and style. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 1389–1392). Berkeley: Department of Linguistics, University of California, Berkeley.

Johnstone, B. (2011). Dialect enregisterment in performance. *Journal of Sociolinguistics, 15*(5), 657–679.

King, R., & Comeau, P. (2011). Media representations of minority French: Valorization, identity, and the Acadieman phenomenon. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique, 56*(2), 179–202.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (2000). *Principles of linguistic change: Social factors*. Malden, MA: Blackwell.

Liberman, M. (2011, December 31). Telegraphic language. Retrieved from http://languagelog.ldc.upenn.edu/nll/?p=3674

Lippi-Green, R. (2012). *English with an accent* (2nd ed.). New York: Routledge.

Mandala, S. (2008). Representing the future: Chinese and codeswitching in *Firefly*. In R. Wilcox & T. Cochran (Eds.), *Investigating* Firefly *and* Serenity*: Science fiction on the frontier* (pp. 31–40). New York: I. B. Tauris.

Mendoza-Denton, N., & Jannedy, S. (2011). Semiotic layering through gesture and intonation. *Journal of English Linguistics, 39*(3), 265–299.

Milroy, J., & Milroy, L. (1999). *Authority in language: Investigating Standard English* (3rd ed.). London: Routledge.

Okrent, A. (2009). *In the land of invented languages: Esperanto rock stars, Klingon poets, Loglan lovers, and the mad dreamers who tried to build a perfect language*. New York: Spiegel & Grau.

Quaglio, P. (2009). *Television dialogue: The sitcom* Friends *vs. natural conversation*. Amsterdam: John Benjamins.

Queen, R. (2004). "Du hast jar keene Ahnung": African American English dubbed into German. *Journal of Sociolinguistics, 8*(4), 515–537.

Queen, R. (2012). "The Days of Our Lives": Language, gender and affluence on a daytime television drama. *Gender and Language, 6*(1), 151–178.

Reaser, J. (2003). A quantitative approach to (sub)registers: The case of "sports announcer talk." *Discourse Studies, 5*(3), 303–321.

Reaser, J., & Adger, C. T. (2007). Developing language awareness materials for nonlinguists: Lessons learned from the *Do You Speak American?* curriculum development project. *Language and Linguistics Compass, 1*(3), 155–167.

Richardson, K. (2010). *Television dramatic dialogue: A sociolinguistic study*. New York: Oxford University Press.

Silverstein, M. (1996). Indexical order and the dialectics of sociolinguistic life. Paper presented at the Salsa III conference. Austin, TX.

Simpson, J. (2003). Reliable authority: Tabloids, film, email and speech as sources for dictionaries. In J. Aitchison & D. M. Lewis (Eds.), *New media language* (pp. 187–193). London: Routledge.

Society for Cinema and Media Studies (2012). *Society for Cinema and Media Studies' Statement of Fair Use Best Practices for Media Studies Publishing*. Retrieved from http://www.centerforsocialmedia.org/fair-use/related-materials/codes/society-cinema-and-media-studies-statement-fair-use-best-practices-

Spitulnik, D. (1996). The social circulation of media discourse and the mediation of communities. *Journal of Linguistic Anthropology, 6*(2), 161–187.

Squires, L. (2011). Enregistering internet language. *Language in Society, 39*(4), 457–492.

Squires, L., & Queen, R. (2011). Media clips collection: Creation and application for the linguistics classroom. *American Speech, 86*(2), 220–234.

Tagliamonte, S., & Roberts, C. (2005). So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech, 80*(3), 280–300.

Vignette 13a
Working with Scripted Data

A Focus on African American English

Tracey L. Weldon

Because so much sociolinguistic research (particularly in the variationist tradition) has been directed at tapping into the elusive “vernacular” (i.e., that most relaxed style, in which speakers are presumed to pay the least amount of attention to their own speech), few researchers have made use of scripted data as a sociolinguistic resource. However, language drawn from television commercials, sitcoms, films/movies, plays/performances, newscasts, or prepared speeches can be used for a variety of sociolinguistic endeavors, including the study of language ideology, linguistic stereotyping, accommodation, and audience design. My own work with scripted data has focused on the use of African American English (AAE) in the media and, in particular, on filmic representations of the variety (Weldon, 2010). In this vignette, I draw from this experience and the work of others in this small but growing area of research to reflect on methodological approaches to working with scripted data.

Because scripted data are not naturally occurring speech data, there is a limit to how much researchers can or should draw from them about real-world phenomena (see, for example, Queen, 2004, as well as Queen, Chapter 13, and Adams, Vignette 13b). In most instances, scripted speech is planned and rehearsed. It is often read and recorded. Thus, it lacks the spontaneity of naturally occurring speech. For purposes of character development, actors may use languages, dialects, or accents other than their own in performing scripted speech, often with the assistance of dialect coaches. And public speakers often bring voice to words prepared by speechwriters. As a result, scripted data are often not reflective of the (socio)linguistic backgrounds of the speakers themselves. Harper (2006) draws the following distinction between scripted and unscripted media: “Presumably, in unscripted media, the speaker has primary control over his/her language. Only in scripted media can one posit some degree of linguistic control or decision-making by someone other than the speaker (i.e., the studio executives, director, writer, etc.)” (p. 18). One of the main challenges in working with scripted data is thus determining how best to analyze speech that ultimately reflects the influence of multiple linguistic sources.

In my own work on representations of AAE in film, one strategy was to obtain original screenplays for the films under consideration. Several screenplays were available through university libraries. Others I purchased through services such as Script Fly (www.scriptfly.com). Having the original screenplays in hand not only facilitated the transcription process itself but also allowed me to see how the language of a given film differed from that which was originally scripted, thus providing some insight into the linguistic choices made by the actors and/or directors of the films, as opposed to the screenwriters themselves.

As a second strategy, I collected background information on various people associated with the films in question. Such information was particularly useful in examining questions of authenticity in filmic representations of AAE. A number of studies of AAE in the media have observed either a "whitewashing" effect by which general non-standard features are preferred over more ethnically marked features (see, for example, Fine & Anderson, 1980) or a tendency toward more minstrel-like characterizations of the variety (see, for example, Harper, 2006). As a working hypothesis, therefore, I considered whether more "authentic" representations of AAE might be found in films involving primarily African American actors, directors, producers, etc., to the extent that they were active members of the African American speech community and thus more familiar (consciously or not) with the rules and nuances governing the use of AAE (see also Harper, 2006; Wilkerson, 2000). Demographic and other background information of this nature were often readily available through websites such as Internet Movie Database (www.imdb.com) and Wikipedia.

Another strategy that some researchers have used to get a sense of the full linguistic range of actors involved in scripted productions, and thus their potential contribution(s) to a given script, is to consider their speech across a variety of films, filmic genres, or media genres, as well as their unscripted speech (e.g., in television or radio interviews). Harper (2006), for example, compared the variable use of specific features of AAE by four actors across five roles each in order to investigate actor variability and indexicality. Other researchers have gathered critical "behind-the-scenes" information from people directly involved in the scripted production. For example, in an analysis of *be* variability in film, Wilkerson (2000) interviewed writer, actor, director, and producer Laurence Fishburne, who provided key insights into decisions made about character development and language usage for two films that he was involved in. While such interviews are likely rare in the case of television shows, commercials, or movies, they are probably more feasibly obtained in studies of local newscasts, plays/performances, or public speeches (see, for example, Johnstone, 2009; Kendall & Wolfram, 2009).

In spite of the above-mentioned challenges and limitations, there are some inherent advantages to working with scripted data. One of the main advantages is their availability. Unlike naturally occurring speech, which usually must be recorded with the permission of the speaker(s), as well the approval of an ethics board (such as, in the United States, Institutional Review Boards), scripted data are often widely available and/or easily accessed, with little to none of the red tape typically associated with human subjects research. Such accessibility is particularly useful in the case of quantitative studies, where, for example, large samples of data may be needed.

As observed by Queen (2004), scripted speech also has the advantage of being highly stylized, which makes it ideal for the observation of language ideology and linguistic stereotyping: "Scripted productions may be more conducive than unscripted ones to the study of sociolinguistic indexicality because the stylized choices found in scripted productions are generally highly focused and easily manipulated indexes that can be (and are) taught to actors" (p. 517). Thus, while there is a limit to what scripted data can reveal about real-world speech, such data can be quite useful in examining real-world ideologies. In fact, much of the existing work on media representations of AAE, my own included, has been directed at better understanding the role of the media in reflecting and perpetuating stereotypes about the variety and its speakers (see, for example, Fine & Anderson, 1980; Green, 2002; Harper, 2006; Lopez, 2008; Rickford & Rickford, 2000; Ronkin & Karn, 1999). Particularly interesting with regard to language ideologies and linguistic stereotyping are studies of animated characters, where scripted language can be used to assign human characteristics (e.g., friendliness, anger, evil, intelligence, humor, compassion) to non-human characters (see, for example, Lippi-Green, 1997).

Because scripted speech typically has a predetermined target audience, it can also be ideal for examining linguistic style via accommodation and audience design. For example, in my research I found that smaller-grossing films aimed at primarily African American audiences presented more nuanced (and perhaps more authentic) portrayals of AAE and its speakers than larger-grossing films aimed at more mainstream audiences. In addition to audience, other style-influencing factors such as topic, setting, and referee design can be examined for their potential influence on linguistic choices made about phonological, lexical, rhetorical, and even discourse or conversational features in scripted data, bearing in mind, of course, the caveats mentioned earlier regarding scripted vs. unscripted speech.

If handled properly, scripted data are a potentially rich and largely untapped source of sociolinguistic investigation. With increasing access to various types of media and information resources associated with scripted data, sociolinguistic research in this area is likely to grow rapidly in the coming years and to bring new perspective to questions that have previously been considered only on the basis of analyses of naturally occurring speech.

References

Fine, M. G., & Anderson, C. (1980). Dialectal features of Black characters in situation comedies on television. *Phylon, 41*(4), 396–409.

Green, L. (2002). *African American English: A linguistic introduction*. Cambridge: Cambridge University Press.

Harper, L. (2006). The representation of African American Vernacular English in the media. (Unpublished master's thesis). University of California, Santa Barbara.

Johnstone, B. (2009). Stance, style, and the linguistic individual. In A. Jaffe (Ed.), *Stance: Sociolinguistic perspectives* (pp. 29–52). New York: Oxford University Press.

Kendall, T., & Wolfram, W. (2009). Local and external language standards in African American English. *Journal of English Linguistics, 37*, 5–30.

Lippi-Green, R. (1997). *English with an accent*. New York: Routledge.

Lopez, Q. (2008). Media's (mis)appropriation of Black speech. Paper presented at the New Ways of Analyzing Variation 37 conference. Houston, TX.

Queen, R. (2004). “Du hast jar keene Ahnung”: African American English dubbed into German. *Journal of Sociolinguistics, 8*(4), 515–537.

Rickford, J. R., & Rickford, R. J. (2000). *Spoken soul: The story of Black English*. New York: John Wiley.

Ronkin, M., & Karn, H. E. (1999). Mock Ebonics: Linguistic racism in parodies of Ebonics on the Internet. *Journal of Sociolinguistics, 3*(3), 360–380.

Weldon, T. (2010). Bougie banter: Representations of middle class AAE in film. Paper presented at the American Dialect Society conference. Baltimore, MD.

Wilkerson, R. (2000). African-American English in film: “Be” variability. In J. Auger & A. Word-Allbritton (Eds.), *IUWPL 2: The CVC of sociolinguistics: Contact, variation, and culture* (pp. 139–156). Bloomington, IN: IULC Publications.

Vignette 13b
Working with Scripted Data

Variations among Scripts, Texts, and Performances

Michael Adams

Scripted works in mass media (television, film, music) appear to be stable and reliable sources of linguistic evidence, but they aren't; at least, use of them requires some caution for at least some types of sociolinguistic inquiry.

Of course, if you have access to the script underlying the work in question, you know what's in the script. But what's in the script isn't always in the work, and vice versa. Also, audiences interpret the performed work differently than the script: people consult the script to affirm a prescriptive accuracy, but people watching a television show, for instance, may hear something "inaccurate." The parameters of "accurate" in linguistic experience are problematic: each hearing is authentic, even if it doesn't match what an author originally wrote into a script.

For instance, in "Welcome to the Hellmouth," the first episode of the television series *Buffy the Vampire Slayer* (Whedon and Smith, 1997), Buffy responds to Giles' cryptic "Something is coming. Something is going to happen here soon," with the typically sarcastic, "Gee, can you vague that up for me?" from which one can extract the unexpected phrasal verb *vague up*, interesting on its own as a contrived slang item, but also as one token among many in the phrasal verbing so characteristic of the show and so prevalent in youth speech at the time (*bail out*, *deal with*, *freak out*, *wig out*, etc.), as well as similarly popular clippings of those same phrasal verbs (*bail*, *deal*, *freak*, *wig*).

But did Buffy in fact say *vague up*? The episode first aired on March 10, 1997, on the WB Network: production values were shaky, videocassette was the mode of recording and replaying the episode, and perhaps Buffy said *vogue up*, though in context it makes no sense. Whoever wrote "How to Speak Buffy" in the official *Buffy the Vampire Slayer* magazine (Winter 1998) heard *vogue up* 'beautify', and in the late 1990s the "inaccurate" form appeared in *Buffy*-related chat rooms and posting boards. The error played a (very) small role in use and codification of slayer slang.

When collecting material for *Slayer Slang* (Adams, 2003), I initially recorded *vague up* – that's what I heard – but couldn't reconcile that with what others apparently heard. I listened to the videotape with its blurry audio over and over again. The shooting script for the episode confirmed *vague up* (Whedon et al., 2000, p. 33), as do enhanced versions of the episode (for instance, in DVD formats), but in the years between the episode and the script book, viewers couldn't know for sure. In other words, if the artifact under inspection for data is scripted but the script isn't available, some data may be problematic, because establishing the text will be problematic.

Eventually, access to the script establishes the scripted language, so a variationist study contrasting scripted speech and "real" speech will not suffer much from anomalies like **vogue up*. Tagliamonte and Roberts (2009, p. 298), evaluating use of intensifiers (*really*, *very*, and *so*) in the television series *Friends*, went into the relationship between script and performance quite carefully. The more quantitative the approach, the more likely it is that a researcher can exclude forms that do not authentically represent the variable in question. But even then one may encounter difficulties: does one have ***a** wiggins* 'unsettled feeling, anxiety, fear' or ***the** wiggins* 'the creeps'? The question is worth asking if one is interested in actuation (Adams, 2003, pp. 106–107) or in the perpetuation and development of the form in slang "after" *Buffy* (Peters, 2006). Early on in *Buffy*, the idiom is unsettled, and it may be difficult to know from listening which variant is performed in some instances; the article occurs in an unstressed position, so articulation of it may, especially in some recorded media, be obscured. The script will clarify which variant should have been used but not which was actually used, so it will account neither for production of the item in question nor for everyone's linguistic experience of the work. In some cases, a scripted work will actually have two texts, the script and the performance, and one has to be clear about which is the source of any data collected.

The textual problem is still more complicated. For instance, one might look for evidence of particular features, stylistic practices, unrecorded slang, humor, gendered speech, or language that constructs the relationship between Jennifer Check and Needy Lesnicky in the film *Jennifer's Body* (Dubiecki, Novick, & Reitman, 2009). But which *Jennifer's Body*? There is a theatrical cut (102 minutes) and a director's cut (107 minutes), and, according to the film's director, Karyn Kusama, the latter "is a more accurate portrayal of Diablo [Cody]'s script" (Wax, 2009), especially with regard to humorous dialogue. The director's cut offers five extra minutes of tokens or features. Can one account for language of the work on the basis of the theatrical release alone? Many films are released in both theatrical and director's cuts, and often the difference in time and tokens between them is greater than five minutes and what those minutes hold.

In fact, not only can there be more than one performance of a script, but there can be multiple scripts of a scripted work, or at least variation among productions of that work. In the film *Jawbreaker* (Silverman, 1998) – that is, in its theatrical cut – the villain Courtney Shane reassures her co-conspirators they will successfully cover up the accidental murder of a high school friend: "Remember, everything is peachy keen – peachy fucking keen," a useful datum to collect if studying infixing and interposing in American speech. The F-word can't be heard on non-cable television networks, however. When *Jawbreaker* aired on the USA Network on June 3, 2002, *peachy fucking keen* was overdubbed as *peachy fuzzy keen*, an item interesting for more than one reason (see Adams, 2002), but not least with regard to euphemism, language attitudes, and mass media. This sort of "re-scripting" happens all the time.

Peachy fuzzy keen is not an "incorrect" item. It's not even an unscripted item. It just doesn't appear in *Jawbreaker*'s shooting script. It belongs to a different textual version of the work in question. Those who heard it in all of its sociolinguistic significance did *not* hear what was in the original script: the television version constitutes their linguistic experience of the film, which then contributes to their overall linguistic experience. If one collects data from *within* a particular (and scrupulously identified) version of scripted work for the purposes of variationist analysis of a feature or variable, like Tagliamonte and Roberts (2005), one need not worry about variation among versions of the same work. But if one collects data to answer more general, qualitative questions focused on perception (rather than production), language attitudes, or the intersections of language and culture – that is, if one collects data *across* texts – a scripted work that appears in multiple texts involves both complications and opportunities.

Zimmer (2011) has demonstrated the difficulties of extracting lexical data from very early street-recorded hip hop cassettes. Some problems arise from the quality of reproduction, as in recordings of other scripted works, but others arise from the fact that different performances of the same song resulted in different texts. One's linguistic experience of the scripted work depends on which performance of which script one has heard, unless we rely on an entirely prescriptive notion of an original script's authority over a work's performance and reception. Careful investigations of the variety of linguistic experience enabled by scripted works, whether controlled and accountable (quantitative) or relatively open and flexible (qualitative), must consider variations among scripts, texts, and performances of a given work.

Textual problems like these are familiar to philologists (and textual critics – the people who edit Shakespeare's plays, for instance), who encounter textual variation all of the time and who write the history of one or another language on the basis of it. In collection of data from scripted works, sociolinguistics converges (if only tentatively) with philology, which is, in my view, a welcome development. Philology can inform sociolinguistic practice with regard to collecting evidence from scripted speech, whether a sociolinguist pursues a linguistic variable or engages in an encompassing sociosemiotic account of meaning and cultural value, such as "mediatization." As Agha (2011) explains, "Mediatized practices occur inside the media but also outside of them. And most mediatized objects are not associated with 'the media' at all" (p. 164). But some of them occur inside the media, and some of those depend on scripted works. When collected with eyes wide open to the textual complications, evidence from scripted works can prove useful in many modes of sociolinguistic inquiry.

References

Adams, M. (2002). Meaningful interposing: An accidental form. *American Speech, 77*, 441–442.

Adams, M. (2003). *Slayer slang: A* Buffy the Vampire Slayer *lexicon*. New York: Oxford University Press.

Agha, A. (2011). Meet mediatization. *Language and Communication, 31*, 163–170.

Dubiecki, D., Novick, M., & Reitman, J. (Producers), & Kusama, K. (Director). (2009). *Jennifer's body* [Motion picture]. USA: Twentieth-Century Fox Home Entertainment.

How to speak Buffy. (1998, Winter). *Buffy the Vampire Slayer, 2*(1), 50–51.

Peters, M. (2006, May). Getting a wiggins and being a bitca: How two items of Slayer slang survive on the *Television Without Pity* message boards. *Slayage, 20*. Retrieved from http://slayageonline.com/essays/slayage20/Peters.htm

Silverman, A. (Producer), & Stein, D. (Director). (1998). *Jawbreaker* [Motion picture]. USA: Columbia TriStar.

Tagliamonte, S., & Roberts, C. (2005). So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech, 80*, 280–300.

Wax, A. (2009, September 15). Exclusive: "Jennifer's Body" director's cut … and sequel? Fear Net. Retrieved from http://www.fearnet.com/news/news-article/exclusive-jennifers-body-directors-cut-and-sequel

Whedon, J. (Writer), & Smith, C. M. (Director). (1997). Welcome to the Hellmouth [Television series episode]. In G. Davies et al. (Producers), *Buffy the Vampire Slayer*. Los Angeles: WB Network.

Whedon, J., et al. (2000). *Buffy the Vampire Slayer: The script book: Season one, volume 1*. New York: Pocket Books.

Zimmer, B. (2011). The new rap language: The emergence of hip-hop lexis. Paper presented at the 18th Biennial Conference of the Dictionary Society of North America, Montreal.

14 Online Data Collection

Jannis Androutsopoulos

In the past 20 years or so, research on computer-mediated communication (CMC) in linguistics has examined language online from a variety of aspects. Specifically sociolinguistic issues include variation and style in digital written language, processes of innovation and change, language and social identities, multilingualism and code switching, and the relation of language, digital media, and globalization. This and other research on CMC evolves in constant interaction with the socio-technological evolution of the internet, which I divide into three broad stages: In the pre-web era – that is, until the early 1990s – CMC is largely restricted to interpersonal (dyadic or group level) exchanges carried out on applications (or modes) such as email, mailing lists, newsgroups, and Internet Relay Chat (IRC). In the early web era, from the mid-1990s to mid-2000s, the emergence of the World Wide Web introduces personal homepages, web discussion forums and corporate websites, followed by blogs. In the participatory web era, from the mid-2000s onward, people draw on the infrastructure provided by blogs, social networking sites, media-sharing sites, and wikis in order to both produce and consume web content. In the course of this development, digital media evolved from socially exclusive to almost ubiquitous in the Western world, and from a small set of options for interactive written communication to a rich repertoire of multimodal and multimedia choices. The various modes of digital communication introduced in these three "eras" accumulate in implicational ways, with each era adding on to the options offered by the previous one. These developments shape what is being viewed as typical "internet language," what is perceived as "research-worthy," and what counts as relevant online data.

Based on an inclusive view of sociolinguistics that encompasses variationist, interactional, and discourse-oriented approaches to language in society, this chapter summarizes a range of issues related to online data collection. While it is increasingly possible to draw on compiled and annotated CMC corpora (Beißwenger & Storrer, 2008), this chapter focuses on issues related to the individual collection of original data. As it is practically impossible to neatly separate data collection from broader issues of methodology, parts of the discussion address conceptual, methodological, and analytic conditions that may affect data collection.

The chapter first discusses how CMC challenges methodological assumptions in sociolinguistics and outlines data sampling criteria in the framework of

Computer-Mediated Discourse Analysis. The next two sections introduce two distinctions that impact how we approach language online: viewing CMC as "text" or "place" and collecting data "on screen" or through contact with users. Subsequent sections discuss issues related to the modes and environments being sampled, multimodality, social identities and participation roles, units and sequences of online data, and research ethics.

Traditions and Challenges in Online Data Collection

Language-focused CMC research faces the challenge of adapting traditions of scholarship to the technological, social, and pragmatic conditions of digital communication. Familiar methods cannot be just replicated in new contexts. This is fairly well understood with respect to specific frameworks. For example, the absence of directly accessible socio-demographic information on language users and the lack of spoken-language data impose limitations on variation analysis, which call for creative solutions (Androutsopoulos, 2006; Herring, 2001; Paolillo, 2001; Squires, 2012). Likewise, the transfer of conversation-analytic categories to online data is limited, owing to the technological restrictions of synchronous CMC, which cancel out the familiar turn-taking system. Researchers may examine how users themselves respond to these restrictions (an empirical problem), but also have to adapt their own use of analytic categories (a methodological problem). Regardless of framework, general issues regarding the collection of online data for sociolinguistic purposes include the following:

1. The online data of interest to linguists is overwhelmingly written language data. CMC research is therefore confronted with the marginal status of written language in sociolinguistics and at the same time contributes to raising interest in written language data.
2. Written language online is closely related to various semiotic resources, including typography, still and moving images, and screen layout; the media-richness of contemporary digital environments increases the impact of multimodality on meaning-making.
3. Modes of digital communication introduce new base-level units in online discourse. Units such as "message" or "post" must be taken into account when collecting and analyzing online data, and their relation to familiar syntactic and discourse-level units (sentence, clause, utterance, turn, adjacency pair) must be analytically examined.
4. In CMC, social contexts can be invisible or only partially retrievable from digital exchanges. Information on participants and their social relationships is often limited for both analysts and participants. New conventions of anonymous public exchange emerge, and traditional operationalizations of socio-demographic constructs may be of little use.
5. Digital language data can be strikingly heterogeneous, especially if researchers sample across the range of digital modes, each with their respective semiotic resources, that people use in their online practices.

6. Digital data are available in overwhelming amounts, making it difficult to select and focus on one specific sample or site of discourse.

These are empirical conditions that CMC researchers across disciplines have to live with and adapt to in terms of their methodologies. The following sections identify some "best practice" solutions or alternatives that respond to these issues.

Data Sampling in the Computer-Mediated Discourse Analysis Framework

Computer-Mediated Discourse Analysis was the first coherent framework for CMC research in linguistics (Herring, 2004; 2007). Herring's work includes a typology of media and social/situational factors for the classification of CMC data and an outline of six criteria for data sampling, which are reviewed here in detail, elaborating on Herring's own pointers (2004, pp. 351–354):

1. *Random sampling* means that each unit of communication from an available set of data has an equal chance of being selected. A "randomizer" tool can be used to select items from a numbered list of posts or messages, or items in specified intervals can be selected (e.g., every 10th message from a newsgroup). Random sampling enables representativeness and generalizability but may result in a loss of context and coherence (e.g., by truncating conversations).
2. *Sampling by theme* can be used to collect data from discussion forums or other thematically organized streams of online discourse (e.g., hash-tagged tweets). The sample can consist of all messages in a particular forum thread or category, which are then compared to an equal sample from another thread in terms of, say, language style or language choice. This method is useful within a framework that includes theme or topic as a relevant factor conditioning language variation or language choice (Androutsopoulos, 2007a). However, sampling by theme has the disadvantage of excluding other co-occurring discourse activities (e.g., other topics discussed by the same users) and is therefore less useful if we are interested in language style across CMC modes and genres.
3. *Sampling by time* is required for any kind of longitudinal analysis. Researchers interested in language change online can draw samples at regular intervals across the available archives of a given newsgroup or forum. Sampling by time offers data that are rich in context, but it may result in very large samples and/or truncate interactions.
4. *Sampling by phenomenon* focuses on particular linguistic features or patterns of language use. For instance, we could select only posts that contain emoticons or certain patterns of non-standard spelling. Such feature-based selection can be (at least partially) automated by means of a concordance or customized script (Siebenhaar, 2006). Herring's examples are discourse-level phenomena such as joking or conflict negotiation, which must be selected

manually. Sampling by phenomenon is the method of choice for features that do not occur frequently and could therefore be absent from samples compiled based on other criteria. It enables "in-depth analysis of the phenomenon" in question (Herring, 2004, p. 351) but may result in loss of context and rule out a distributional analysis.

5. *Sampling by individual or group* can be based on socio-demographic criteria, if available, or some kind of member ranking in the relevant online environment. It can enable analysis of selected users and user comparisons along familiar sociolinguistic lines. However, it excludes by definition exchanges with other users or groups.
6. *Sampling by convenience* – that is, selecting "whatever data are available" (Herring, 2004, p. 351) – was popular in some early CMC research. Beyond its obvious advantage, it lacks a principle of systematic selection and may yield unsuited samples.

All alternatives have advantages and disadvantages, and the eventual choice depends on the research question and methodological practicalities. These criteria do not preempt the type of analysis that will eventually be carried out. Some options (notably 2, 3, and 5) roughly correspond to familiar "external" or independent variables and result in datasets that will be later scanned for linguistic features of interest. Option 4 targets particular features straight away, possibly ruling out a systematic control of independent variables if it is not deployed in combination with other selection criteria. In practice, however, combinations of two or more criteria are common.

CMC as "Text" or "Place"

In a paper on qualitative online research, Milner argues that "the study of cultures online demands we decide whether we frame online interaction as 'place' or as 'text'" (2011, p. 14). Although Milner's research is in communication studies, his dyad of "place" and "text" can be productively adapted to sociolinguistic concerns. I suggest that from the perspective of language studies, "CMC as text" focuses on the vast archive of written language provided by the internet. It implies a tendency toward screen-based data, a view of digital modes as "containers" of written language, and a preference for *etic* (researcher-oriented) rather than *emic* (participant-oriented) classifications and categories. By contrast, a "CMC as place" perspective might approach digital communication as a social process and CMC environments as discursively created spaces of human interaction, which are dynamically related to offline activities. Here, online data from various modes and environments might be collected, taking into account their cross-connections in people's digital literacy practices. This approach is likely to prefer ethnographic observation and blended data.

The example of Twitter can be used to illustrate this distinction. Approaching Twitter as "text" may mean collecting a large set of data, possibly via data-mining techniques, and analyzing those data in terms of specific linguistic variables or categories, thereby distinguishing between, say, "private users" and "organizations" in terms of social variables. A “Twitter as place” view could examine how particular social actors use Twitter alongside other digital modes to report on or coordinate social action related to a particular event (say, a political rally or a natural catastrophe), thereby shaping the course and meaning of that event.

One reason the text/place dyad seems useful in a sociolinguistic context is, in my view, that it echoes the familiar tension between “system-oriented” and “speaker-oriented” approaches – in other words, the differential focus of sociolinguistic research on linguistic variation itself as opposed to speakers’ language practices. The text/place dyad does not determine the type of analysis to be carried out; rather, it defines an epistemological perspective, which likely entails a preference for particular research questions and techniques of data collection.

Screen- and User-Based Data Collection

In the second distinction, “screen-based” and “user-based” refer to the two main, and in my view complementary, sites of data collection in new media sociolinguistics. “Screen-based” data are produced by participants and collected online by the researcher, while “user-based” data are prompted by the researcher’s activities and produced through their contact with CMC users. A limitation to screen-based data may seem the norm in language-focused CMC studies, but this is neither self-evident nor uncontested within the discipline, let alone from an interdisciplinary perspective. Jones (2004) argues that the notion of context in CMC should not be reduced to what is happening on screen, but requires a shift of attention to the offline social activities in which CMC is embedded. From this viewpoint, CMC is shaped by a duality of situational context with simultaneous online and offline aspects. While a limitation to digital textual data may be motivated by research questions that focus on linguistic variation rather than language practices, CMC researchers commonly find that the interpretation of linguistic findings can benefit from some awareness of the social and situational contexts of the data. In my own research (Androutsopoulos, 2008), I have been interested in CMC users’ awareness of particular linguistic variants and choices as a complement to screen-data analysis.

Figure 14.1 represents the relation of screen- and user-based data on a continuum with intermediate positions, which correspond to various degrees of ethnographic engagement on the part of the researcher. They will be briefly discussed, moving from “left” to “right” on the figure. (The extreme-right position is not discussed, as I assume that research on CMC sociolinguistics will always encompass screen-based data.)

Collecting screen data depends on both the options provided by various modes and environments and the technological sophistication brought along by researchers (for an introduction to corpus linguistic approaches to the web, see Hundt et al., 2007; Sharoff, 2006). Synchronous applications such as IRC and instant messenger (IM) come with the convenience of logfiles. Forum pages can be manually downloaded and then have to be cleaned up from html code in preparation for concordance or other software treatment. Content from social networking sites can be saved in PDF files, or relevant portions can simply be

	Screen-based			User-based
Relation of researcher to source of data	No online observation	Systematic online observation	Online observation and contact to users	Contact to users without online observation
Resulting type of data	Online data	Online data	Blended data	Offline data

Figure 14.1 Screen-Based and User-Based Data in CMC Research.

copy-pasted. Besides these more or less simple techniques, large portions of screen data can be mined using web crawlers, application program interfaces (APIs), customized scripts or other resources. Digital data can also be delivered to researchers by users themselves, e.g., students, members of the general public who donate data, or acquaintances of one member of a research team (Dürscheid & Stark, 2011; Schmidt & Androutsopoulos, 2004; Tagliamonte & Denis, 2008; Tsiplakou, 2009). This option specifically concerns private digital data exchanged on one-to-one applications and comes with the bonus of available socio-demographic information. Depending on the research question, the selection of screen data may proceed following any of the six sampling criteria (or combinations thereof) reviewed above.

Strategies of online data collection differ not just in terms of technology but also with regard to the degree of researcher engagement with the relevant site(s) of online communication. The researcher's position on a cline between no or minimal observation to fully fledged familiarity with the online research site is in principle independent of the technique of screen data collection. Data mining of course rules out a simultaneous online observation; what is relevant, however, is whether any prior engagement with the original sites of this data has taken place, by which a selection of data to be mined has been determined. In the extreme opposite case (which I am tempted to label the "take the data and run" approach), a researcher may harvest large amounts of digital data without ever visiting the sites where they originate. However, a complete lack of familiarity with the original site of the data may limit the available contextual information, resulting in a preference for standardized (*etic*) user categorizations and perhaps a replacement of socio-demographic categories by modes.

Online observation refers to the process of "virtually being there," with or without active participation, and watching the digital communication you will eventually analyze as it unfolds in a website or a network of connections across sites. Online observation is implicitly part of much linguistic CMC research, but it is often not explicitly acknowledged. I distinguish three aspects of online observation: "revisit," "roam around," and "explore resources for participation." "Revisit" stands for paying regular, iterative visits to the selected site, tracing both routine activities and changes. "Roam around" suggests exploring the virtual ground, browsing around sites, sections, threads, or profiles. Whether to lurk or actively participate is open to debate in the literature (Garcia, Standlee, Bechkoff, & Cui, 2009; Milner, 2011). What is important, in my view, is that researchers do not end up analyzing their own data or data that occurred as a direct outcome of their own contributions. "Explore resources of participation" stands for trying out all resources afforded by an online environment of choice, such as search facilities, user lists, statistics, tags, and tag-related hit lists. Across these activities, online observation involves a systematization of vernacular digital literacy practice, and the collection of screen data is complemented by the digital equivalent of ethnographic fieldnotes (which may involve tools such as Zotero or Evernote).

Observation, the bottom line of any "virtual fieldwork," comes in degrees. I suggest that even limited online observation offers a (limited) degree of ethnographic grounding, which can be further expanded and refined, and whose benefit can only be assessed within a particular project. In the absence of direct contact with users, the ethnographic information gained will be limited to what can be elicited in, or inferred from, the online environment. But especially when it comes to public (and semi-public) web spaces where participants' mutual background knowledge is incomplete and fragmented anyway, systematic observation can offer considerable insights that can subsequently be used to interpret surface data, to identify new objects of analysis, or to articulate new research questions (Androutsopoulos, 2008). Such insights may concern intertextual references or running gags, common and rare discussion topics, the usual pace or rhythm of discursive activities, categories of participation (e.g., core and peripheral members), the distribution of particular features across members, the trajectory or career of particular threads, and so on. With a bit of luck, researchers may even witness trends in a community's online talk as they emerge (see Kytölä & Androutsopoulos, 2012). As in any ethnographic endeavor, systematic observation allows researchers to acquire some of the "tacit knowledge" underlying the semiotic practices of regular members.

"Blended data" refers to any combination of screen data and data collection through direct contact to selected users. I focus here on cyclical procedures of blended data collection, assuming that user-based data will come to complement and interpretively frame the analysis of screen-based data. User-based data are of course not "online data" in the narrow sense of the term. Depending on question and contact, their collection may even take the researcher far off the computer to the offline environments where the social activities that participants "entextualize" – that is, document and turn into digital text – can be observed (Jones, 2009).

Some user contacts offer access to data in the first place, while others are initiated and established after an initial period of online observation and screen-data collection. In the first case, contacts may be decided in advance, as part of the overall research design; in the latter case, their selection will depend on previous observation and selection, for example by focusing on core members or users who "stand out" in some way. Depending on the research question and the researchers' familiarity with social-scientific methods, user-based data can be elicited in direct (face-to-face) or mediated contact by means of various instruments, including interviews, group discussions, or questionnaires, or by observing people's literacy practices in front of their computers. Interviews (narrative or semi-structured) can be also carried out via Skype, phone, chat, or email. Each choice has implications in terms of further methods of data handling, including recording and transcription.

Cyclical procedures of blended data collection can begin with observation, followed by screen data collection and preliminary analysis, then establishing contact with selected participants. In the contact situation, samples of online content can serve as a prompt in order to elicit participants' awareness of and attitudes to language use online. The cycle can be extended, or repeated, by additional data collection, perhaps following new hints to language features or patterns identified in the interview. User contacts can thus be the last or an intermediate step between two layers of screen data analysis. My own experience includes various patterns of sequencing screen and user-based data. One pattern is to observe private homepages or discussion forums, then contact and interview their producers or webmasters, then return to and refine screen data analysis. Research on social networking sites may involve an initial contact (off- or online) with likely participants, gaining permission to access their profiles, then observing profile activities and collecting samples, carrying out preliminary analyses, and conducting individual or group interviews. In my own research on multi-party IRC (Androutsopoulos & Hinnenkamp, 2001; Androutsopoulos & Ziegler, 2004), a period of familiarization involving observation of and some active participation in the channel of choice was followed by contact to selected individuals by means of the one-to-one ("whisper") mode afforded by chat software; disclosing my researcher identity, I could then discuss language issues with these individual chatters or ask them to fill out a short questionnaire. In this case, screen and user-based portions of CMC data were collected in parallel and simultaneous, but separate, online activities.

Modes and Environments

Broadly defined as applications that offer a standardized user interface and a set of options for digital interaction, modes are key components of CMC for users and researchers. Modes are traditionally classified on the parameters of synchronicity (synchronous/asynchronous) and publicness (one-to-one, one-to-many or many-to-many), thereby distinguishing IM (synchronous, 1 : 1) from IRC (synchronous, many : many) from email (asynchronous, 1 : 1 or 1 : many) and so on (Herring, 2001).

Modes usually serve as invariant parameters for digital data selection, and a lot of data reported in the literature are specified for or even restricted to particular modes, e.g., IRC, IM, or email. Analysis of mode-specific online data ties in with the practice of dividing "internet language" into mode-specific components, which are then discussed in separate textbook chapters, and so on. In sociolinguistic practice, modes have also played the role of external (independent) variables, based on the assumption of more or less stable relations between modes and patterns of online language use. In particular, the hypothesis that synchronous modes of CMC resemble spontaneous spoken language more than asynchronous ones has been tested for variation of standard/vernacular and spoken/written features as well as for the occurrence of conversational code-switching (Androutsopoulos, 2007b; Paolillo, 2011). Such inter-mode analysis compares data from two or more CMC modes (e.g., messaging vs. email or chatting vs. newsgroups) while controlling other social and situational factors. By contrast, an intra-mode design compares data from the same mode across varying social and/or situational conditions (e.g., informal online chat to moderated chat sessions with politicians). Provided the primacy of mode effects on language over social and/or situational factors is not being assumed by default, modes offer an invaluable handle for data collection and exploration.

The usefulness of modes as building blocks of online data collection is weakened by the growing importance of participatory web environments, where old modes are integrated, and new genres cannot be distinguished on synchronicity and publicness alone. Such environments include online portals that host edited content and user discussion forums; social network sites with user profiles, walls, and groups; and content-sharing platforms for photos and videos. Owing to their sheer size and diversity of contributors, genres, and interactive applications, web environments create new problems of comparability. To put it simply, comparing YouTube to Vimeo as such makes little sense from a linguistic perspective. Rather, developing a meaningful comparison relies on systematic online observation by which to identify relevant types of content, genres, or users prior to the actual data collection. Examples include a comparison of three asynchronous genres on hip hop portals for colloquial markers in spelling (Androutsopoulos, 2007b), the analysis of status updates as a prominent small genre on Facebook walls (Bolander & Locher, 2011; Lee, 2011), and the selection of YouTube videos and comments based on user tags (Pihlaja, 2011).

Multimodality

Multimodality can be understood in at least three different ways in the context of CMC. First, it can refer to user activities during the production of and interaction with online content. In research that includes photographs or video recordings of users in front of their screens (see papers in Androutsopoulos & Beißwenger, 2008), methods of multimodal analysis of embodied interaction can be used to examine the relation between users' face expression and posture and the online content they type in or read. In a second sense that relates to the concept of mode in the previous section, multimodality refers to the simultaneous use of more than one application in people's digital literacy practice. Screen recording software can be used to document how users multitask on various applications and what this means in terms of, for example, style shifting. This technique is not (yet) widespread in sociolinguistics but could offer an interesting addition to blended data. In a third sense, multimodality refers to the coexistence of resources from more than one semiotic mode in digital content itself. The evolution of CMC brought about increasingly complex forms of multimodal communication, and while language-heavy modes such as email predominate in early language-focused research, the contemporary integration of written language with other semiotic resources (spoken language, audio, static and moving image, video, color, pictograms, typography, etc.) presents a methodological challenge. Researchers interested in self-presentation online have long been alert to how users draw on all semiotic resources at their disposal to construct their identities on homepages and blogs. In contemporary web environments, an increasing amount of written or spoken language comes embedded in visual or audiovisual texts (think of lolcat images and YouTube videos), and written-language exchanges are often prompted by multimodal texts, as can be observed on Flickr or social networking sites (Lee & Barton, 2011). Even when the research question is concerned with the language part, taking into account multimodal prompts may help interpret patterns of variation or style choice. In the absence of widely accepted standards for multimodal online data collection, page-long screenshots and automated video/comment download are viable techniques, though ethical considerations may restrict the types of content that can be downloaded (see also Sadler, Vignette 3d).

Social Identities and Participation Frameworks

CMC complicates the process of social identity ascription for both researchers and participants. Digital communication, especially of the public type, is often carried out anonymously and among interlocutors who lack information for mutual social categorization. This is a serious problem for any sociolinguistic analysis that depends on clear-cut socio-demographic information (gender, social class, etc.), but it can be addressed or circumvented in a number of ways. First, researchers can contact relevant users and collect socio-demographic information post hoc, though doing so is not always practically feasible, especially in public CMC. Second, researchers can take data offered by users themselves as a basis for speaker categorization. Depending on mode and genre, these data may range from fairly straightforward information to a range of indexical cues in screen names and associated virtual identity signs such as avatars, member profiles, or signatures. One challenge concerns online versus offline identities and whether to conceive of users as "behaving like" or rather "performing" a particular social identity; however, this issue goes beyond data collection. Alternatively, researchers can abandon external socio-demographic factors and turn to environment-specific categories such as regulars/novices or admins/normal users, to which sociolinguistic variation is then correlated (Paolillo, 2001). A further alternative is to focus on the discourse processes by which participants ascribe and negotiate social identities to selves and others, thereby drawing on interpretive methods of data collection and analysis.

This discussion suggests that the more we depart from "offline" socio-demographic variables as a basis for the sociolinguistic analysis of CMC data, the more we need to reconstruct participation roles in various digital modes and environments, thereby going beyond a medium-specific replication of the simple "sender : receiver" (i.e., writer : reader) model. This has been an issue for analysis rather than data collection so far. In his study of participation roles in French newsgroups, Marcoccia (2004) distinguishes between "host" and "casual sender" on the basis of various diagnostic criteria. Hosts send and reply to more messages than other senders, are often on friendly terms with each other, manage (e.g., initiate, regulate) online interactions, and often play the role of experts. On the reception side, Marcoccia distinguishes between the (explicitly) addressed recipient, the favored recipient (which he takes to coincide with the host), and the "eavesdropper" – that is, the non-addressed but ratified recipient commonly referred to as a "lurker." Data collection in public or semi-public online environments can anticipate these (or adequately modified) participation roles, especially in terms of their relation to institutional conditions of communication online and/or theoretical frameworks of choice.

Units, Sequences, Intervals

Sociolinguistic studies of CMC data usually focus on micro-linguistic and interactional units, and data collection is therefore geared toward collecting material that contains these units. However, familiar units of linguistic analysis are embedded and reframed in larger-scale units of digital mediation that are defined by CMC applications or environments. These include the units of messages (units in one-to-one exchanges) and post (units of contribution to public, multi-party exchanges), which are in turn embedded in larger, multi-authored structures such as threads or lists of comments. Messages and posts are indispensable units of data collection, but their relation to familiar linguistic or conversation-analytic categories such as sentence, utterance, or turn is neither trivial nor straightforward. For example, a conversational turn can be divided into several online posts, and one post can accommodate more than one turn depending on its composition.

Acknowledging messages or posts as an additional level in the organization of online data is, in turn, indispensable for working with sequences – that is, temporally arranged chains of posts or messages that are exchanged in a particular interactional configuration. A sequence is either collected as such in a public CMC environment (e.g., a Facebook wall conversation, or forum threads) or reconstructed ("zipped together") from data exchanged between separate digital interlocutors. Any research question that takes its cues from pragmatics and interactional sociolinguistics is more or less dependent on collecting sequences rather than isolated messages or posts. As a consequence, the interactional processes usually examined in sequential analysis (e.g., adjacency pairs) are reframed within a sequence of posts or messages. When researching code-switching online, for example, post-internal and post-external code switching (i.e., within or across posts) form an additional level of analysis that does not coincide with either turns or sentences (Androutsopoulos, 2007a). This reframing has also an impact on intervals – that is, the time distance between individual contributions in the flow of a dyadic or multi-party exchange. Much has been written on intervals from the viewpoint of constraints determined by technology, resulting in transmission gaps or leading to an order of posts that disrupts expectations of sequential coherence. But relatively little is known about the active management and interpretation of intervals by participants themselves (Jones, 2005; Schmidt & Androutsopoulos, 2004). In practice, the time-stamps contained in the online data or noted by researchers or participants are a useful resource for reconstructing intervals, which can be analyzed as indexes to participants' footings in text-based interaction.

A Note on Research Ethics

Respecting and protecting the privacy of informants is a basic legal and ethical requirement in social-scientific fieldwork. There is no general consensus on how to maintain privacy in CMC research, and ethics guidelines for researchers and students vary by country and institution. It is common sense among CMC researchers that we need to protect the anonymity of our informants by not directly disclosing their offline identities and avoiding any cues that may lead to their identification. Various modes, environments, and user groups pose different conditions for achieving this aim. Maintaining anonymity for private online data is easier than for public and semi-public data. Asking participants for permission to use and publish is the rule regarding private data, but it is not always feasible for data collected from or available on public sites of CMC. Moreover, the researcher's (technical) definition of what constitutes publicness may not be shared by participants, resulting in diverging interpretations of what data can be treated as in public domain. Some scholars treat publicly posted screen names (e.g., on YouTube) as publishable. However, these can be easily traced back to other publicly available utterances posted under the same screen name. Even when screen names are anonymized, verbatim quotations from publicly accessible material may also lead back to original posts via web search. A complete anonymization of public CMC data may be technically impossible. On the other hand, not all online communicators may wish to stay anonymous in academic publications; famous bloggers could be a case in point. This possibility should be understood not as an excuse not to anonymize but as a reminder that participant and researcher views do not forcibly coincide. Our ethical decisions must ultimately observe legal requirements of "privacy," but our considerations should not neglect informants' views on the shifting boundaries of privacy and publicness. Readers are also referred to the ethics guidelines of the Association of Internet Researchers (http://aoirethics.ijire.net) and Vignette 3d by Sadler on ethics in online data collection.

References

Androutsopoulos, J. (Ed.). (2006). Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics, 10*(4).

Androutsopoulos, J. (2007a). Language choice and code switching in German-based diasporic web forums. In B. Danet & S. C. Herring (Eds.), *The multilingual internet* (pp. 340–361). New York: Oxford University Press.

Androutsopoulos, J. (2007b). Style online: Doing hip-hop on the German-speaking Web. In P. Auer (Ed.), *Style and social identities: Alternative approaches to linguistic heterogeneity* (pp. 279–317). Berlin: Mouton de Gruyter.

Androutsopoulos, J. (2008). Potentials and limitations of discourse-centred online ethnography. *Language@Internet, 5*, article 8. Retrieved from http://www.languageatinternet.org/articles/2008/1610

Androutsopoulos, J., & Beißwenger, M. (Eds.) (2008). *Data and methods in computer-mediated discourse analysis*. *Language@Internet*, *5*, introduction. Retrieved from http://www.languageatinternet.org/articles/2008/1609

Androutsopoulos, J., & Hinnenkamp, V. (2001). Code-Switching in der bilingualen Chat-Kommunikation: Ein explorativer Blick auf #hellas und #turks. In M. Beißwenger (Ed.), *Chat-Kommunikation* (pp. 367–401). Stuttgart: Ibidem.

Androutsopoulos, J., & Ziegler, E. (2004). Exploring language variation on the Internet: Regional speech in a chat community. In B.-L. Gunnarsson, L. Bergström, G. Eklund et al. (Eds.), *Language Variation in Europe* (Papers from ICLaVE 2, pp. 99–111). Uppsala: Uppsala University Press.

Beißwenger, M., & Storrer, A. (2008). Corpora of computer-mediated communication. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics* (Vol. 1, pp. 292–309). New York: de Gruyter.

Bolander, B., & Locher, M. A. (2010). Constructing identity on Facebook: Report on a pilot study. *Swiss Papers in English Language and Literature, 24*, 165–185.

Dürscheid, C., & Stark, E. (2011). *sms4science*: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In C. Thurlow & K. Mroczek (Eds.), *Digital discourse: Language in the new media* (pp. 299–320). New York: Oxford University Press.

Garcia, A. C., Standlee, A. I., Bechkoff, J., & Cui, Y. (2009). Ethnographic approaches to the internet and computer-mediated communication. *Journal of Contemporary Ethnography, 38*(1), 52–84.

Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 612–634). Malden, MA: Blackwell.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online communities. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for virtual communities in the service of learning* (pp. 338–376). Cambridge: Cambridge University Press.

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet, 4*, article 1. Retrieved from http://www.languageatinternet.org/articles/2007/

Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.) (2007). *Corpus linguistics and the web*. Amsterdam: Rodopi.

Jones, R. H. (2004). The problem of context in computer-mediated communication. In P. LeVine & R. Scollon (Eds.), *Discourse and technology: Multimodal discourse analysis* (pp. 20–33). Washington, DC: Georgetown University Press.

Jones, R. H. (2005). "You show me yours, I'll show you mine": The negotiation of shifts from textual to visual modes in computer-mediated interaction among gay men. *Visual Communication, 4*(1), 69–92.

Jones, R. H. (2009). Dancing, skating and sex: Action and text in the digital age. *Journal of Applied Linguistics, 6*(3), 283–302.

Kytölä, S., & Androutsopoulos, J. (2012). Ethnographic perspectives on multilingual computer-mediated discourse. In M. Martin-Jones & S. Gardner (Eds.), *Multilingualism, discourse, and ethnography* (pp. 179–196). New York: Routledge.

Lee, C. K. M. (2011). Texts and practices of micro-blogging: Status updates on Facebook. In C. Thurlow & K. Mroczek (Eds.), *Digital discourse: Language in the new media* (pp. 110–128). New York: Oxford University Press.

Lee, C. K. M., & Barton, D. (2011). Constructing glocal identities through multilingual writing practices on flickr.com. *International Multilingualism Research Journal, 5*(1), 39–59.

Marcoccia, M. (2004). On-line polylogues: Conversation structure and participation framework in internet newsgroups. *Journal of Pragmatics, 36*(1), 115–145.

Milner, R. M. (2011). The study of cultures online: Some methodological and ethical tensions. *Graduate Journal of Social Science, 8*(3), 14–35.

Paolillo, J. C. (2001). Language variation on Internet Relay Chat: A social network approach. *Journal of Sociolinguistics, 5*(2), 180–213.

Paolillo, J. C. (2011). "Conversational" codeswitching on Usenet and Internet Relay Chat. *Language@Internet, 8*, article 3. Retrieved from http://www.languageatinternet.org/articles/2011/

Pihlaja, S. (2011). Cops, popes, and garbage collectors: Metaphor and antagonism in an atheist/Christian YouTube video thread. *Language@Internet*, *8*, article 1. Retrieved from http://www.languageatinternet.org/articles/2011/

Schmidt, G., & Androutsopoulos, J. (2004). *löbbe döch*. Beziehungskommunikation mit SMS. *Gesprächsforschung, 5*. Retrieved from http://www.gespraechsforschung-ozs.de/

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics, 11*(4), 435–462.

Siebenhaar, B. (2006). Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics, 10*(4), 481–509.

Squires, L. (2012). Whos punctuating what? Sociolinguistic variation in instant messaging. In A. Jaffe, J. Androutsopoulos, M. Sebba, & S. Johnson (Eds.), *Orthography as social action: Scripts, spelling, identity and power* (pp. 289–323). Boston: Walter de Gruyter.

Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech, 83*(1), 3–34.

Tsiplakou, S. (2009). Doing bilingualism: Language alternation as performative construction of online identities. *Pragmatics, 19*(3), 361–391.

This page intentionally left blank

Part IV

Sharing Data and Findings

This page intentionally left blank

15 Sharing Data and Findings

Christine Mallinson

The chapters and vignettes in Part IV, "Sharing Data and Findings," address concepts, decision points, and techniques related to collecting sociolinguistic data from various populations as well as to sharing sociolinguistic data and/or findings with the public. Language is often central to the rights and privileges that are afforded by social institutions, including education, the legal system, and the media. As a result, many sociolinguists have addressed issues of language-related concern and promoted social justice by applying findings from sociolinguistic research to public issues and by engaging with the public in specific outreach endeavors.

Many linguists have argued that seeking ways to apply our knowledge should be a central concern of the broader scholarly enterprise. In "The Socially Minded Linguist," Bolinger (1979, p. 404) enjoins linguists not to "stay aloof" from concentrations of power and inequality that are often also "questions of language." Similarly, Wolfram (2012, p. 111) notes that while linguistics has a reputation as being an esoteric and abstract field, "linguistic research ranging from neurolinguistic imaging studies to studies of language variation in the community should be of interest into the public"; he goes on to say that linguistic research would indeed be of more interest "if the public knew how connected linguistic research was to their everyday life." But this burden of communicating about linguistic research does not rest on the public; it rests on linguists, who must make engagement with the public a priority:

> If linguists firmly believe that understanding the nature of language is central to understanding human cognition and behavior, then we owe it to the profession to have more of a presence in public life.... Public education about language is not just a luxury for full professors... Personally, I think that it is a responsibility that we all must share if we desire to sustain and expand our discipline.... Besides that, sharing some of our reverence and respect for language is one of the most rewarding career experiences I could imagine.
>
> (pp. 111, 116)

Sociolinguists who recognize a scholarly responsibility and even an obligation to engage with the public point out that public engagement should be guided by theoretical and methodological principles that form a comprehensive, ethically grounded approach to the methods of engagement. A series of principles well known to most sociolinguists are Labov's (1982) principle of error correction and principle of debt incurred, and Wolfram's (1993) principle of linguistic gratuity; Cameron, Frazer, Harvey, Rampton, and Richardson's (1992) three models of research – ethical, advocacy, and empowering – have also been influential in sociolinguistics.

More recently, Wolfram (2012) makes several methodological and practical recommendations for sociolinguists who are seeking to connect with the public. He asserts that, not at the end of a project but rather "from the outset," sociolinguists should "consider how linguistic research might have a strategic public outreach dimension" (p. 112). In other words, principles of engagement should not be adopted ad hoc or post hoc but rather outlined in advance and revisited throughout a research project. Wolfram also recommends that sociolinguists "be visionary and entrepreneurial" in how we consider the public dimension of our work (p. 114). To achieve maximal impact, we should also seek to foster long-term and sustainable public engagement endeavors, such as producing media and curricular materials (p. 115); to do so, we must not work in a solitary capacity but rather collaboratively, with non-linguists, including journalists, artists, educators, students, and community partners (p. 116).

Sociolinguistic data collection is therefore not merely a discrete phase in the research process, disconnected from questions of engagement and application. Rather, as the chapters and vignettes that appear in Part IV make clear, all manner of data-related considerations are relevant and interrelated concerns, including the types of data that are collected, the manner in which data are collected, and whether and how data are disseminated and to which academic and public groups. Just as it is crucial to plan in advance how to effectively and efficiently collect reliable and valid data, it is also crucial, as we prepare to collect our data, to consider whether and how our research goals and our methods of data collection can best align with our goals for engagement and application.

Questions of ethics are also central when working with public groups and when applying sociolinguistic knowledge to public- and/or community-specific concerns. How can we best initiate, foster, and sustain ethically sound collaborations with various communities? How can we maximize our efforts for the public good? Underlying these types of questions, according to Roberts (2012), a professor of Public Engagement, must be a rejection of a deficit model in which the public is seen as needing to be educated and, instead, the implementation of a model of public engagement that centers on and privileges dialogue. The best people to engage with the wider public about science are scientists, Roberts states, but we need to arm ourselves with the ability to communicate in order to be most effective.

The chapters and vignettes in Part IV explore how effective dialogue between sociolinguists and the public can take place. The chapters provide deep theoretical grounding in how to conceptualize, plan, and implement methods of public engagement, including data collection and dissemination of data and findings. The vignettes provide real-world scenarios of sociolinguistic engagement, exploring successes and challenges. Together, these chapters and vignettes provide guidance as well as food for thought when considering how to apply sociolinguistic knowledge and build connections with the public in various domains, including communities, schools, and the media.

Community and Educational Engagement

In Chapter 16, “Community Activism: Turning Things Around,” Arapera Ngaha describes connections between community activism and sociolinguistics, particularly with respect to linguistic rights and language planning. Describing the Māori people’s struggle to save *te reo Māori* (the Māori language), Ngaha explores ethics in language research in relation to community activism and provides recommendations that center on limiting the power differential between researcher and researched. With a focus on transparency and reciprocity, she calls upon scholars to collaborate with community members to determine the purpose and methods of linguistic work, allow community activists to make most decisions unless linguists are called upon, and actively search for ways to repay the debts of time and insights that community participants provide.

In Chapter 17, “Sociolinguistic Engagement in Schools: Collecting and Sharing Data,” Anne H. Charity Hudley surveys various approaches that sociolinguists have taken to integrate linguistic research with educational outreach and social activism. Turning toward application, she explores models of sociolinguistic engagement for those who seek to collect data from and share data with those in schools. In order to collect reliable, valid, and relevant data in and for schools, Charity Hudley compels linguists to read widely in the field of education and related disciplines and to work collaboratively with scholars in these fields to share insights and methods. As early scholars of language and education have pointed out, including Hymes (1980, p. 139), “part of what we need to know in order to change is not known to anyone; teachers are closer to part of it than most linguists.” Charity Hudley makes the case for linguists to collect data from schools and students in ways that address issues of mutual concern, to work with research participants not merely as subjects but as partners, and to paint a comprehensive sociolinguistic picture that places language use within broader social contexts. Case studies and practical strategies highlight how sociolinguists can design research to be maximally useful to scholars as well as schools and communities, so that people who contribute data for research purposes can directly benefit from having shared it.

Vignettes 17a and 17b, by Green and Serpell respectively, provide examples of some of the types of data that can be collected in schools, from students. Both authors call for collecting data from students in ways that accurately report on their patterns and norms of language use. Without data that take into account speakers’ range of variation, as Green notes, our measures can lead to false assumptions, both about them and about our research models. Instead, comprehensive and holistic assessments of students’ linguistic and communicative competence can ensure more accurate student assessment, which is particularly important for students from historically underserved groups. As illustrated in Serpell’s vignette, the accurate collection and interpretation of a range of sociolinguistic data are also crucial to the design of an effective literacy curriculum and therefore to establishing educational policy that supports and serves a multilingual society.

In Vignette 17c, Starks emphasizes some of the unexpected difficulties of collecting data in schools, recalling the Pasifika Languages of Manukau Project, in which she worked with a school as an entry point into a multilingual Pacific and Māori community in South Auckland, New Zealand. At the beginning of the project, the principal, who supported bilingualism, provided access for the research and promoted acceptance for it in the community. But the following year, when a new principal took over and the school lost most of its funding for bilingual programs, some of the materials that Starks and her colleagues produced were used in ways that supported a very different political agenda. Even if sociolinguists are unable to plan for the unexpected, Starks's vignette makes clear the importance of anticipating the fact that micro-level research endeavors can have serious macro-level implications.

Engagement with the Media

Following these discussions on community activism and educational engagement, Chapter 18, by Jennifer Sclafani, introduces the topic of sociolinguistics and the media. Sclafani points out the importance of viewing the media not simply as an object of study but as a conduit for communication between sociolinguistics and the public. Two case studies from the United States are explored in depth: the 1996 Ebonics controversy and the more contemporary situation of terminology applied to immigrants working without proper authorization. The media can also be used by linguists to disseminate linguistic knowledge, such as through films and documentaries, or in other ways that make contributions to schools and communities. Sclafani concludes that we must remain aware of how the media amplify and mute certain voices, create spaces in which dominant language ideologies may be perpetuated and also contested, and share linguistic data and findings with other researchers as well as the public.

Three vignettes also engage with questions of media engagement, including the collection of sociolinguistic data from media sources and the dissemination of sociolinguistic findings through media outlets. In Vignette 18a, Scott F. Kiesling discusses his experiences with media coverage of popular academic topics, focusing on his work on the word *dude* and the dialect of Pittsburgh, Pennsylvania. He notes the importance of embracing and even seeking out publicity for one's research: not only does media coverage heighten awareness of linguistics as a science and of language as an important topic of professional inquiry, but it also helps scholars connect with laypeople who may share valuable information about language and language use. Media coverage can also help dispel language ideologies; therefore, Kiesling says, it is worth trying to shape public perceptions about language, even if given only a soundbite during which to do so.

In Vignette 18b, Clive Upton reports on the BBC Voices Project (2004–2007), in which 60 journalists interviewed groups of speakers across the United Kingdom and collected local and personal words for 38 different everyday concepts; a website also allowed the public to contribute data and discuss language-related matters. In addition to being used for academic research, the large amount of data that were collected yielded sociolinguistic findings that were discussed on BBC radio and television. In Vignette 18c, Andrew D. Wong explores the semantic change of the Chinese term *tongzhi* ('comrade'), which in the late 1980s came to refer to sexual minorities. Wong built a corpus of articles from the *Oriental Daily News* between 1998 and 2000 to examine how *tongzhi* is used, while grappling with the selectivity with which print media present and represent certain voices and texts. Sharing his findings with *tongzhi* activists, Wong realized that his work speaks to issues surrounding how power is exercised and contested through language. Thus, sociolinguists can analyze language that is used in the media in ways that also contribute to public understanding of its political and societal implications.

Conclusions

Within Part IV, "Sharing Data and Findings," the chapters and vignettes make clear that sociolinguists can do both research and outreach, but that we should plan carefully and strategically in advance. If we conduct research that has an engagement component, we must set clearly outlined goals in which we consider questions of ethics, ownership, relevancy, responsibility, and resources. As we study research methods, we must also study methods of engagement, addressing such issues as how to communicate with various publics, how to collect data from and share findings with them, how to establish and manage a public persona in addition to one's scholarly persona, and how to navigate outreach-related challenges that may arise. As the authors in Part IV suggest, rather than being wary of engagement or viewing it as an add-on that comes only after the scholarly research has been completed, sociolinguists have much to gain and much to contribute when we weave public engagement into our research. Being part of the public conversation depends on making use of the skills of discourse and communication – skills that are manifestly at our disposal as sociolinguists to activate and employ.

References

Bolinger, D. (1979). The socially-minded linguist. *Modern Language Journal, 63*(8), 404–407.

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., & Richardson, K. (1992). *Researching language: Issues of power and method*. London: Routledge.

Hymes, D. (1980). *Language in education: Ethnolinguistic essays*. Washington, DC: Center for Applied Linguistics.

Labov, W. (1982). Objectivity and commitment in linguistic science. *Language in Society, 11*(2), 165–201.

Roberts, A. (Interviewee). (2012). Five minutes with Alice Roberts: "During my academic career I've encountered considerable opposition to engagement with the public" [Interview transcript]. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2012/03/02/5-minutes-with-alice-roberts/

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

Wolfram, W. (2012). In the profession: Connecting with the public. *Journal of English Linguistics, 40*, 111–117.

16 Community Activism

Turning Things Around

Arapera Ngaha

"Community activism" is a term derived from political activism and was used most effectively through the rise of community engagements in the fight for Black liberation by African Americans through the 1960s. Mobilization of the Black community was community activism in action. In the field of sociolinguistics, that same mobilization of language communities to seek justice pertaining to linguistic rights and linguistic freedom is indicative of community activism. Linguistic or language rights and freedom as addressed in this work relate to "small" and "minority" language communities that seek to retain and maintain their own language, and cultural practices that help define who they are as a distinct ethnic identity. When the language ceases to be used as the lingua franca of the community and inter-generational language transfer is no longer the norm, the language is seen to be in danger. For many of the world's small and minority language communities, it is not until that point is reached, or is seen to be imminent, that the community members themselves recognize the severity of language loss and are motivated to find ways to support language maintenance and revitalization. But when access to resources and assistance to address their linguistic concerns is blocked or is difficult to obtain, community activism has sometimes been invoked in order to find a way forward. Sociolinguistics has two distinct links with community activism in pursuit of linguistic supports for an endangered language: concerns around language rights and research ethics.

This chapter discusses these matters as they pertain to sociolinguistics and community activism. The context for the Māori people and *te reo Māori* (the Māori language) is briefly outlined, followed by illustrations of Māori community activism, where such action has resulted in gains for the ongoing survival and maintenance of *te reo*. Applications of sociolinguistic research methodologies that have progressed language revitalization efforts for *te reo Māori* may also provide encouragement and support for other minority language communities as they search for ways to help their language survive within a majority language society (see Nicholls, 2001, and Taylor, 2001, with Aboriginal languages in Australia; Nyika, 2008, in Zimbabwe; and Kuter, 1989, with Breton). The examples from the Māori experience may provide insights for language planning and language maintenance concerns as well as for researchers hoping to work in these communities.

Minority Language Rights

May (2003) provides a comprehensive overview of the issues surrounding minority language rights (MLR). He supports the arguments for MLR but notes that the arguments offered have usually been situated within the language and theorizing of biological and ecological contexts – the language ecology movement. He argues that rather than view language rights through this context, which belabors the negativity of language loss, endangerment, and the inevitability of language death if languages are not supported, a shift to viewing these matters through sociohistorical and sociopolitical lenses provides a more stable argument for language rights (p. 96). Advocates for linguistic human rights argue that these are also an extension of basic human rights (Wee, 2010, p. 49).

The essentialist model of language rights is made through the link between language and identity: without the language, elements of the culture are lost, or misunderstood. In this model, language is still considered the primary means of transmitting the values and beliefs of a culture from one generation to the next (Fishman, 2001; Marsden, 2003; Ngaha, 2011). The danger, as articulated by May (2003), is that such a model can present a romanticized view of the language that suggests an idealistic view of reversing language shift and may promote unrealistic language maintenance goals and objectives. May is concerned that in the push for MLR, proponents inadvertently provide opportunities for their critics to promote instead a mobility argument. In that argument, minority languages, because they are unsupported in the majority language society, cannot provide access to opportunities that are available to speakers of the more powerful majority language. These inequities highlight the power relations that are evident in society and are often catalysts for initiating community activism.

Detractors of the essentialist view argue that language is but one element of identity, that markers of identity can change over time, and that language therefore cannot be considered a core value of identity (Edwards, 1985; Kuter, 1989; Nash, 1987). Song (2003) suggests that a more "pervasive marker of identity is that sense of belonging which includes having a common history" (p. 14). This lends weight to the argument for situating MLR discourse within a sociohistorical and sociopolitical framework, but it also assists the push for nationhood and establishing a national or universal language within any given community detracting from equity arguments for minority languages. If we consider that language performance provides indicators of identity that are always situated within particular contexts (Mullany, 2006, pp. 157–159; Omoniyi, 2006, p. 13; Tabouret-Keller, 1997, p. 315) then placing MLR arguments in the sociohistorical and sociopolitical arena may well serve to take the emotive and romanticized elements out of the argument for MLR.

McIntosh (2005) and Omoniyi and White (2006) maintain that identity formation provides a frame of reference, "a point of departure from which we can assemble and negotiate who we are" (Ngaha, 2011, p. 15). That frame of reference allows for some retention of values and beliefs, markers of identity such as language, or discarding of such markers as appropriate for the circumstance of the individual. Ngaha (2011) promotes a flexible and fluid model of identity where language is one of a number of markers of identity that become more or less pronounced according to the circumstance of the individual at any one moment in time. Borell (2005) notes how participants in her study of Māori youth in South Auckland chose locality as a more appropriate defining feature of their identity than *te reo*, their traditional language. And, when choices are made about which language is used in specific contexts, the individual chooses the language best suited to her or his particular context and circumstance – what Edwards (1985) describes as "economic rationality."

This discussion highlights the position that minority languages find themselves in, where language shift is more pronounced when the minority language has severely limited support in society, and language survival is determined through the unequal distribution of power relations within that society. Minority languages are constantly battling to avoid being subsumed and acculturated into the majority language society, where policies and practices invariably privilege the majority languages. May (2003) urges that the battle for MLR continue, for it is imperative that minority language proponents take an active role in helping make change by challenging the social and political arena that seeks to deny these rights. It is only by being a part of the process that the balance of power can be altered.

Ethics in Language Research

Cameron, Frazer, Harvey, Rampton, and Richardson (1992) are often cited for their work on models of research used in sociolinguistics within communities. Their proposed models move from a position of "research on" the community – where the researcher plays the role of the objective observer, the "outsider," in the process – to a "research on and for" advocacy model and then to a "research on, for, and with" model, which is ostensibly an empowerment model. Although this model suggests power sharing, the aims, goals, and outcomes of the research are still primarily determined by the researcher(s), albeit filtered through community engagement in the process. The community has a limited measure of power, but the cultural capital invested by the community into the research tends to far outweigh any return benefit.

Recognition of the community must be transparent and the benefit gained from the research has to be real in its members' eyes: it has to be of value to them, an advantage to them, and it must be purposeful. It is important that the researcher(s) understand the value of this transparency and that the researcher(s) understand the community. Such insights cannot be achieved by the researcher(s) alone, as "insider" assistance is imperative for gaining commitment to the research process and outcomes. Action research or participatory research with communities has been explored in a range of disciplines, with varying degrees of success (e.g., Borell, 2005; Cashman, 2006; Hudson, 2005; Levy & Kingi, 2003). One model of community-based, participatory research advocated by Maiter, Simich, Jacobsen, and Wise (2008) is based on a more reciprocal approach that advances equity and that, I believe, comes much closer to addressing the concerns of minority and indigenous communities. Maiter et al. (2008)

note: “Reciprocity is not only necessary to accomplish research in an ethical manner, but it is also illuminating, since the process of negotiating priorities and learning what study participants expect to obtain from co-operating with researchers reveals valuable cultural knowledge” (p. 308). For Māori, reciprocity (*utu*) is a value well understood in our cultural context, a process that advocates equity and respect for all parties.

Members of the community must be engaged in the research at all stages, and the ethical responsibility of the researcher must be to the target minority group themselves, “from identification of community needs to interpretation of results and implementation of any proposed actions” (Skuttnab-Kangas, 1990, p. 98). Crystal (2000) underlines the need to place the community at the center of research in language revitalization, and Nyika (2008) on language revival of Zimbabwean notes that “grassroots mobilisation is based on the argument that the affected language community should be at the centre of language revitalisation efforts” (p. 4). These sentiments are echoed in work with Aboriginal languages in Australia (Nicholls, 2001; Taylor, 2001) and in New Zealand with *te reo* and the Māori community (Hudson, 2005; Mead, 2004; Ngaha, 2011).

Building strong and respectful relationships between the researcher(s) and the target community is the first step in conducting sound community research. To do so requires community engagement at the outset. It requires the researcher(s) to make contact with senior members of the target community and discuss with them their ideas regarding a possible research project. At this very first stage, it is important to listen, to hear what the community’s views on such ideas and proposals might be. The researcher(s) must be open-minded, prepared to work collegially, willing to work in ways that may be outside of their own comfort zone and research experience thus far, and flexible in their ways of thinking about the context and environment within which the research will be situated. The researcher(s) must also be prepared to have their ideas and proposals reshaped, revised, or even dismissed altogether.

The potential for clashes is high when outsider researchers attempt to carry out research without gaining the confidence and respect of the target community. The two parties come from different value bases and will have differing perspectives on the roles of both the researcher(s) and the community members, and different expectations for the research outcomes. If the challenge of differing ideologies coming together in this way is not addressed at the outset, it is likely that dissatisfaction and disruption will pervade all aspects of the work (Cashman, 2006; Ngaha, 2011). In-depth discussions at the earliest opportunity with community elders and interested parties, enlisted by the community, will go a long way to resolving other questions about the research, including the methodology, data-gathering processes, data analysis, and dissemination of the research findings, as well as benefits accruing from the research.

In order to acknowledge the cultural capital invested by the community through its members’ input, it is important to do more than simply include community members as researcher(s) or research assistants and train them in a predetermined research plan. Cashman (2006) suggests using a stronger principle than Labov’s (1982) principle of debt incurred, which he considers too passive.

Instead, he urges the use of Wolfram's (1993) principle of linguistic gratuity, which puts more onus on the researcher to actively seek out opportunities to repay the debt incurred; or, in Wolfram's words, "investigators who have obtained linguistic data from members of a speech community should actively pursue positive ways in which they can return linguistic favors to the community" (p. 227). Cashman urges researchers to use their professional skills to advocate for the community in the courts, in education, and especially in domains where land or resource ownership is a concern. Mutu (2011) talks about how researchers have been of great value in reporting indigenous people's stories as documentary evidence in the Waitangi Tribunal Claims process, ensuring that the views of Māori have been advanced.

The Māori People, *Te Reo Māori*, and Community Activism

Māori are the indigenous people of Aotearoa (New Zealand) and have been in these islands for more than 1,000 years. They settled in their tribal groupings primarily in the coastal regions and adapted their lifestyle from that of the tropics to a harsher subtropical climate. The advent of the first European settlers in the late 1700s and early 1800s forced further change; in 1840, the British sought to gain Aotearoa as a southern outpost for the British Empire and entered into a treaty with Māori for peaceful settlement of this land. The Treaty of Waitangi was signed in good faith by the indigenous Māori, as a treaty of peace and friendship with the British colonizers (Mutu, 2010). The British made guarantees in the treaty to protect Māori assets, including their chiefly autonomy, their land, and other assets Māori considered of value, such as their language.

Over time, encroachments upon Māori assets, primarily land, resulted in Māori being severely marginalized (Waitangi Tribunal, 1986; Walker, 2004). Land loss through legislative means, some dubious land sales, and wars, both tribal and global, took their toll on the Māori population and their asset base. Rural landholdings were much reduced, and subsistence farming resulted in many *whānau Māori* (families) being forced to move to the cities to seek employment. As a direct result of relocation, urbanization, and – ultimately – dislocation from their communal language stronghold, *te reo Māori* became marginalized, and major language loss resulted (Benton, 1997; Waitangi Tribunal, 1986; Walker, 2004).

It was not until the 1970s, when the first National Language Survey (NZCER Survey) was conducted to assess the state of the language, that Māori recognized how severe the decline in *te reo* use had become. Language shift was advancing rapidly, and since *whānau Māori* were no longer speaking Māori in the home, intergenerational language transmission was declining. The fear was that if nothing was done to help promote and support the learning of *te reo Māori*, the language might die out (Benton, 1997). Māori addressed that issue by setting up local pre-school centers that functioned under *tikanga Māori* (Māori values and customary practices) and were conducted in *te reo*. The government was challenged to support this initiative, and in 1981 the first government-funded Māori-medium pre-school (Kohanga Reo – "language nest") was established. Many more throughout the country quickly followed. A few years later, Kura Kaupapa

Māori, Wharekura, and later still Whare Wānanga (Māori-medium primary schools, high schools, and universities) were established. However, despite advances made in establishing Māori-medium education, support for *te reo* in mainstream education remains severely lacking. There are very few compulsory Māori language courses sustained in New Zealand public schools, but one notable gain has been that in 2010, King's College in Auckland, one of New Zealand's most prestigious private schools, implemented compulsory *te reo Māori* classes at years 9 and 10 (students aged 13–15) for all students, regardless of ethnicity. Such inroads in the domain of education were achieved as a result of Māori community activism.

The Waitangi Tribunal (hereafter referred to as the Tribunal), a commission of inquiry set up to hear Māori grievances in relation to Treaty of Waitangi matters, received a claim in 1986 that held the government to account for its failure to protect *te reo Māori*, a guarantee made through the signing of the Treaty of Waitangi. Much of the evidence provided to support the claim came from the NZCER Survey Report, from researchers and professionals in the field of education, and from within Māori communities, and especially native speakers of the language, who feared losing the language altogether. The Tribunal found in favor of the claimants and made a number of recommendations to the government, several of which were implemented by the Crown. The Māori Language Act 1987 resulted from the Tribunal's recommendations, giving recognition to the Māori language as an official language of New Zealand and requiring the government to preserve and protect *te reo Māori* (Benton, 1997; Waitangi Tribunal, 1986).

To meet the government's responsibilities in this regard, several government agencies were charged with fostering the use of *te reo Māori* in public forums such as in public service, the courts, the media, and education. Government departments that had strong involvement with Māori clients were encouraged to implement bicultural awareness staff training programs, including basic *te reo Māori* classes. In the 1980s, bicultural training was a compulsory component of staff development for all staff in the former Department of Social Welfare, but by the mid-1990s the department had restructured, and bicultural awareness was no longer deemed an imperative. Bicultural practices are maintained in some areas but are not official public service policy. In the courts, intent to speak Māori must be notified to the court two weeks prior to a court appearance to ensure translation facilities can be provided.

The media have also played a major role in supporting *te reo* revitalization, through the print media, Māori Iwi radio stations, and Māori television (MTV). Māori newspapers have been in circulation since the 1860s, tribal (Iwi) radio stations have been operating around the country for more than 20 years in some regions, and in March 2004, after a lengthy battle for resources and market position for television airtime, MTV was launched. MTV screens programs in *te reo*, some with English subtitles, bilingual programs, and programs about other indigenous peoples delivered in that indigenous language with English subtitles. The journey toward the establishment of MTV was fraught with problems and disjuncture (Hollings, 2005), but unrelenting community activism ensured that it became a reality. In 2010, MTV secured the rights to broadcast the All Blacks' games live on free-to-air television in the 2011 Rugby World Cup, a coup unprecedented for a small indigenous television station.

Research in Māori Contexts

Researching within Māori communities has its own particular challenges, and researchers intending to work within Māori communities must learn about the community from their viewpoint. Smith (1999) suggests that Māori are among the most-researched indigenous communities worldwide, and they have become wary and discerning about what research is undertaken in their community, who is carrying out that work, and for what purpose. Māori tell stories of researchers who have taken advantage of the people's generosity and departed with material that has provided the researcher with academic kudos but has given little or nothing back to the community. All too often, acknowledgment of the community by outsider researchers has been overlooked. It is not surprising that Māori communities have become suspicious.

Ethics is about values, and values guide the way we see and engage with the world. For Māori, ethics are based on *tikanga*, a process by which the right thing is done, and done in the correct manner. As a guide to informed and ethical research within Māori communities, Hudson, Milne, Reynolds, Russell, and Smith (2010) have produced a set of guidelines or a framework for researchers and ethics committee members who might be considering engagement in Māori community research. They cite a base or rationale for a different way of working:

> In a research context, to ignore the reality of inter-cultural difference is to live with outdated notions of scientific investigation. It is also likely to hamper the conduct of research and limit the capacity of research to improve human development.
>
> (p. 2, quoting National Health and Medical Research Council, 2003)

Māori ethical frameworks derive from Māori creation stories, lessons learned through the exposure to these experiences, and the ensuing *tikanga* or traditional customary practices that evolved. *Tikanga* provides the blueprint for the manner in which Māori conduct themselves, their ways of living, their ways of being and knowing, and their ways of understanding their world. Marsden (2003) considers that these stories were "deliberate constructs employed by the ancient seers and sages to encapsulate and condense into easily assimilable forms of their view of the World, of ultimate reality and the relationship between the Creator, the universe and man" (p. 56). Māori values must be incorporated into the ways in which research is derived, developed, and managed, and only by engagement of the target Māori community in all facets of the research can this process be achieved appropriately. When Māori have the power to control the use of their information, control their input, use their own processes that empower the community, and assert their right to retain *tino rangatiratanga* (self-determination), the end result benefits all parties (Kennedy & Cram, 2010; Levy & Kingi, 2003; Ngaha, 2011).

When a language-related research need is identified, it is the role of the sociolinguist to assist members of that community to clarify what the community hopes to achieve through the research. The sociolinguist, often someone not of that community – an outsider – is the conduit through which tools, resources, knowledge of processes, and analysis can be provided to the community. It is important to emphasize that the outsider should assist only when invited. Anything less serves to disempower and disenfranchise the community. Community members often have the skills and expertise to carry out this work and will derive the research plan using their own *tikanga* processes, which include healthy and robust ethical oversight. What the community may need assistance with is to understand and possibly be trained in the academic and Western processes of linguistic fieldwork and analysis. In this way, the community members are then better equipped to discern the best use of the tools and resources at their disposal, the researcher(s) gain deeper knowledge of the community, and their context and respectful relationships are developed.

For example, some years ago at a church gathering I posed a research idea I had been considering to a small group of *kaumātua* (elders) to get some feedback on how useful it might be to explore. All these *kaumātua* were native speakers and had been actively engaged in the implementation and support of Māori-medium education initiatives for many years. I had been thinking about ways to increase the number of speakers of *te reo* and wondered what Māori might think about encouraging and supporting non-Māori (people who had no *whakapapa Māori* [Māori genealogy]) to learn *te reo*. That proposition generated much discussion, often very heated debate, but after several discussions over many days, I was told that the general feeling was that this was a worthwhile topic to explore. The *kaumātua* then proceeded to outline a research plan and process that I might follow.

These *kaumātua* became mentors for me and the research team in this project. They were a core advisory group who monitored the research from the beginning "germ of an idea" that they helped shape and mold into the research plan, through fieldwork and analysis, and up to the reporting of the research. At each *hui* (gathering) or community focus group in this study, we were accompanied by at least one of the *kaumātua*, who often facilitated parts of the hui. Their contributions in these *hui* were invaluable, as native speakers who could engage with other speakers of *te reo* on a level that was outside of the expertise of some of the team, who were fluent but not as accomplished speakers of *te reo* as the *kaumātua*. Their work in the analysis of the data was also extremely valuable, as they ensured that the subtle nuances of language and language behaviors were addressed appropriately. Marsden (2003) notes that it is only from within the community itself, from within the Māori worldview, that these aspects can truly be understood, and Hudson (2005), Mead (2004), Smith (1999), and Watson (2006) suggest that Māori narratives will always privilege the truth from the cultural viewpoint of Māori.

The reports back to the community and to the academic community were also different. In all our community engagements, *tikanga Māori* was observed. These traditional values and customary practices guided the process when we presented ourselves and our reports to the community. They were delivered both orally and in printed form and discussed bilingually with questions and comments directed to the researchers *kanohi ki te kanohi* (in face-to-face interactions). The academic reports, however, were achieved through my doctoral thesis, conference presentations, and published papers.

Conclusion

Community activism has developed because of the perceived need to make change to secure the rights of communities and has been instrumental to the journey that Māori have taken to revitalize their language. That journey began with addressing the state of *te reo* after the findings from the NZCER survey of the language undertaken during the 1970s revealed that a major language shift had taken place across the whole Māori population, but most markedly in urban regions. The implementation of Kohanga Reo was a "flax-roots" response – a response grounded in this land that came from the indigenous context – to Māori elders' concerns that their *mokopuna* (grandchildren) returning from the cities to traditional homelands for family celebrations and holidays were not speaking *te reo*. The establishment of the Māori Language Act 1987 has furthered Māori revitalization efforts and aided the growth and development of Māori language medium schools. These models of Māori language medium schooling have been the catalyst for a number of indigenous peoples to build their own language learning programs. Again, these advances were only achieved by the efforts of Māori community and supporters within the wider New Zealand society who actively sought and pushed for justice for Māori. Academics and researchers in a range of disciplines, particularly in the fields of law, education, politics, sociology, linguistics, and sociolinguistics, have aided the work by producing the data, writing the reports, and sometimes offering a voice to support Māori endeavors.

The most successful changes have come through projects that have retained Māori engagement at the center and addressed the research work from within a Māori paradigm, within a Māori worldview: Māori community engagement has been at the core driving the research and driving the overall plan, while the support of others has fed into the core but remained at the periphery. These changes have brought about clear gains for the Māori community and *te reo* revitalization. Telling these stories provides clear illustrations about ethics, values, and goals, and, for the "outsider" researcher who chooses to engage in research within minority or indigenous communities, guidance for just how to do so respectfully.

References

Benton, R. A. (1997). The Māori language: Dying or reviving? A working paper prepared for the East-West Center Alumni-in-Residence Working Paper series. Wellington: New Zealand Council for Education and Research.

Borell, B. (2005). Livin in the city ain't so bad: Cultural diversity for young Māori in South Auckland. In J. H. Liu., T. McCreanor, T. McIntosh, & T. Teaiwa (Eds.), *New Zealand* *identities: Departures and destinations* (pp. 191–206). Wellington: Victoria University Press.

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., & Richardson, K. (1992). *Researching language: Issues of power and method*. London: Routledge.

Cashman, H. R. (2006). Who wins in research on bilingualism in an anti-bilingual state? *Journal of Multilingual and Multicultural Development, 27*(1), 42–60.

Crystal, D. (2000). *Language death*. Cambridge: Cambridge University Press.

Edwards, J. R. (1985). *Language, society and identity*. Malden, MA: Blackwell.

Fishman, J. A. (2001). *Can threatened languages be saved? Reversing language shift, revisited: A 21st century perspective*. Clevedon, UK: Multilingual Matters.

Hollings, M. (2005). Māori language broadcasting: Panacea or pipedream. In A. Bell, R. Harlow, & D. Starks (Eds.), *Languages of New Zealand* (pp. 111–130). Wellington: Victoria University Press.

Hudson, M. (2005). A Māori perspective on ethical review in (health) research. In Ngā Pae o te Māramatanga Centre of Research Excellence for Maori Conference Organising Committee (Ed.), *Tikanga rangahau mātauranga tuku iho: Traditional knowledge and research ethics conference 2004* (pp. 57–78). Auckland: Ngā Pae o te Māramatanga.

Hudson, M., Milne, M., Reynolds, P., Russell, K., & Smith, B. (2010). *Te Ara Tika – Guidelines for Māori research ethics: A framework for researchers and ethics committee members*. Auckland: Health Research Council of New Zealand.

Kennedy, V., & Cram, F. (2010). Ethics of researching with whānau collectives. *MAI Review, 3*, 1–8.

Kuter, L. (1989). Breton vs. French: Language and the opposition of political, economic, social, and cultural values. In N. C. Dorian (Ed.), *Investigating obsolescence: Studies in language contraction and death* (pp. 75–89). Cambridge: Cambridge University Press.

Labov, W. (1982). Objectivity and commitment in linguistic science. *Language in Society, 11*(2): 165–201.

Levy, M., & Kingi, T. K. (2003). *Facilitating effective Māori participation in research: Experiences of the National Mental Health Classification and Outcomes Study Monitoring and Review Group* (CAO-MMRG). Wellington: Ministry of Health Research and Development.

Maiter, S., Simich, L., Jacobson, N., & Wise, J. (2008). Reciprocity: An ethic for community-based participatory action research. *Action Research, 6*(3), 305–325.

Marsden, M. (2003). God, man and the universe, a Māori view. In T. A. C. Royal (Ed.), *The woven universe: Selected writings of Rev. Māori Marsden* (pp. 2–23). Masterton: The Estate of Rev. Māori Marsden.

May, S. (2003). Rearticulating the case for minority language rights. *Current Issues in Language Planning, 4*(2), 95–125.

McIntosh, T. (2005). Māori identities; Fixed, fluid, forced. In J. H. Liu, T. McCreanor, T. McIntosh, & T. Teaiwa (Eds.), *New Zealand identities: Departures and destinations* (pp. 38–51). Wellington: Victoria University Press.

Mead, A. (2004). Public policy: Ethics and mātauranga Māori. In Ngā Pae o te Māramatanga Centre of Research Excellence for Maori Conference Organising Committee (Ed.), *Tikanga rangahau mātauranga tuku iho: Traditional knowledge and research ethics conference 2004* (pp. 121–144). Auckland: Ngā Pae o te Māramatanga.

Mullany, L. (2006). Narrative constructions of gender and professional identities. In T. Omoniyi & G. White (Eds.), *The sociolinguistics of identity* (pp. 157–172). London: Continuum.

Mutu, M. (2010). Constitutional intentions: The Treaty of Waitangi texts. In M. Mulholland

& V. Tawhai (Eds.), *Weeping waters: The Treaty of Waitangi and constitutional change* (pp. 13–40). Wellington: Huia Publishers.

Nash, M. (1987). *The cauldron of ethnicity in the modern world*. Chicago: University of Chicago Press.

National Health and Medical Research Council. (2003). *Values and ethics: Guidelines for ethical conduct in Aboriginal and Torres Strait Islander health research*. Canberra: NHMRC.

Ngaha, A. (2011). Te Reo, a language for Māori alone? An investigation into the relationship between the Māori language and Māori identity. (Unpublished doctoral dissertation). University of Auckland.

Nicholls, C. (2001). Reconciled to what? Reconciliation and the Northern Territory's bilingual education program, 1973–1998. In J. Lo Bianco & R. Wickert (Eds.), *Australian policy activism in language and literacy* (pp. 325–341). Melbourne: Language Australia.

Nyika, N. (2008). Language activism in Zimbabwe: Grassroots mobilisation, collaborations and action. *Language Matters, 39*(1), 3–17.

Omoniyi, T., & White, G. (Eds.). (2006). *The sociolinguistics of identity*. London: Continuum.

Skutnabb-Kangas, T. (1990). Legitimating or delegitimating new forms of racism: The role of researchers. *Journal of Multilingual and Multicultural Development, 11*(1–2), 77–100.

Smith, L. T. (1999). *Decolonizing methodologies: Research and indigenous peoples*. London: Zed Books.

Song, M. (2003). *Choosing ethnic identity*. Cambridge: Polity Press.

Tabouret-Keller, A. (1996). Language and identity. In F. Coulmas (Ed.), *The Handbook of sociolinguistics* (pp. 315–326). Oxford: Blackwell.

Taylor, A. (2001). The cost of literacy for some. In J. Lo Bianco & R. Wickert (Eds.), *Australian policy activism in language and literacy* (pp. 149–162). Melbourne: Language Australia.

Waitangi Tribunal. (1986). *Report on Te Reo Maori Claim, WAI 11*. Wellington: Government Printer.

Walker, R. (2004). *Ka whawhai tonu matou: Struggle without end* (rev. ed.) Auckland: Penguin.

Watson, S. (2006). The stories people tell: Teaching narrative research methodology in New Zealand. In S. Trahar (Ed.), *Narrative research on learning: Comparative and international perspectives* (pp. 61–76). Bristol: Symposium Publishers.

Wee, L. (2010). *Language without rights*. New York: Oxford University Press.

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

17 Sociolinguistic Engagement in Schools

Collecting and Sharing Data

Anne H. Charity Hudley

This chapter surveys the methods needed to conduct effective research in schools. I focus in particular on methods that are transformative in that they have changed how sociolinguistic research is conducted and have directly affected the educational experiences of students and educators who participate in and benefit from such research.

Research for Educational Purposes

As noted by observers (Rickford, 1997) and researchers themselves, support for seminal sociolinguistic surveys (e.g., Labov, Cohen, Robins, & Lewis, 1968; Wolfram, 1969) was granted by the United States Department of Education with the goal of better understanding the language patterns of students in diverse communities. Labov's early research inspired other studies on the nature of language variation for African American students and implications for testing and reading (1972a). Similarly, Wolfram (1969) and other early work intersected with the research of speech language pathologists, teachers of English as a Second or Other Language, and the Center for Applied Linguistics. These studies primarily sought to access the speech that was most different from school language and standardized language; secondarily, they attempted to assess language variation within a speaker, which set the methodological model for sociolinguistic examinations since that time. These early studies also led to a battery of common linguistic assessments, including interview questions, word lists and reading passages, and linguistic insecurity tests.

The original interview questions used by Labov et al. (1968) in the Harlem Study were designed to compel the speaker to produce narratives. Questions centered on family, school, games, girls, fights, and counting out. The classic questions "Have you ever gotten blamed for something you didn't do?" and "Was there ever a time that you thought, this is it, I'm going to die?" emerged from this era, and various researchers have since expanded on or adapted these questions. The classic Labovian model also presents a strand of questions to use with younger children and older interviewees to help them remember when they were younger. For example, modules give a sketch of questions to ask children and adults about the games they play. Such interview questions have often been used in school settings, whereas most interviews are done outside the classroom, often in informal settings such as the cafeteria or on the playground or schoolyard.

Word lists and reading passages have also been used as a measure of stylistic variation in sociolinguistic studies in schools, but reading ability presents a distinct challenge with such materials. Baugh (2001) describes difficulty in assessing stylistic variation among adult speakers with limited reading ability, and beginning readers and pre-readers share this problem. A greater sensitivity to the intersection of linguistic variation and linguistic insecurity due to reading level should be monitored when using such tests. In most studies, the two issues are confounded. Linguistic insecurity tests (Labov, 1972b) are, in a sense, the most school assessment-like of the general sociolinguistic assessments, as they seek to discover elements of speech that people would most like to correct. These tests may not, however, fully expose linguistic insecurities in populations with more education, where language-related anxieties may be more masked.

Supra-linguistic Methods

Some of the most robust sociolinguistic methods used in schools are not particular to linguistic elicitation but are informed by the entire situation in which linguistic information is gathered. Labov (1972a) notes that how a researcher executes the methods of working with children in schools is just as important as the linguistic merit of the methods themselves. The height of the interviewer, the rapport between students and interviewer, and the topics being addressed all play a role. Labov describes children whose entire demeanor changed when the interviewer put himself on the same height level as the children, shared snacks with them, and talked about non-school and slightly taboo subjects that put them at ease. (Such observations relate to Bourdieu, 1990, on school as a socializing agent and a broker of cultural capital.)

Extensive school-oriented sociolinguistic interviews and assessments have also led to new methods of thinking about language variation and change. Eckert (1989) examined what happened inside as well as outside school, by studying students who adhered or did not adhere to school structure as a basic organizing principle of their social groups. For about 20 years after such work, sociolinguistic work on education had a strong theoretical focus that often eclipsed the original focus on education and language use within schools. While early sociolinguistic work on education was rich in its narrative content and context, the subsequent theoretical turn made it increasingly difficult for educators to relate the methods used in sociolinguistics, which were designed to assess students' richest vernaculars, with the modes of assessment that are used in schools and with which educators are most familiar.

Intervention Methods

Early sociolinguists also developed materials to be used as interventions for students. The most famous attempt is the Bridge Readers (Simpkins, Holt, & Simpkins, 1977), designed to scaffold children's home language onto the language of school. The books used features of African American English (AAE) and rich narratives but were met with controversy in schools because of orthographic representations of AAE and the vernacular nature of a number of the stories. Recently, sociolinguists have participated in producing multimedia materials, such as *American Tongues* (Alvarez & Kolker, 1988) and *Do You Speak American?* (Cran, Buchanan, & Anthony, 2005), and have created accompanying teaching guides (Reaser, Adger, & Hoyle, 2005). These materials follow a linguistic awareness model but garner wider interest. Many materials that focus on dialect awareness, however, have not directly addressed how educators can specifically work not to discriminate against students who speak varieties of English or other languages.

Other methods of intervention seek to align assessment and instruction in schools with insights from sociolinguistics. Sociolinguistic information is too rarely coupled with specific examples of what a given variety consists of and how to use that information to aid students in developing their reading, writing, and speaking skills. Even further removed from most sociolinguistic materials is discussion of how information about language variation can be used to help socialize students into academic or professional culture. As a result, most work of this type has been produced by non-sociolinguists.

Research methods that most effectively lead to direct student intervention are tied to local, national, governmental, political, and academic guidelines, but much more work remains to be done. Sociolinguists, along with psychometricians and speech-language pathologists, have pointed out cultural and linguistic biases in school assessments, particularly on standardized tests, but there is not much literature showing changes in tests as a result. Most current research consists of sociolinguistic analyses of the tests themselves, which is meant to bring greater awareness about the macro- and micro-level issues that students face when taking such tests. Educators in psychology and speech language pathology (e.g., Hoover, Politzer, & Taylor, 1995) first posed questions about how best to apply sociolinguistic insights in the classroom, and a subsequent generation of educators and sociolinguists are revisiting these questions in their methodology (see, for example, Charity Hudley, 2009; Charity Hudley & Mallinson, 2011; N. P. Terry, 2008). As Godley, Sweetland, Wheeler, Minnici, and Carpenter (2006) observe, for sociolinguistic research to be most applicable, each teacher must be presented with information that is easy to access and implement. Educators who are also experienced in sociolinguistics often provide the most integrative models (e.g., LeMoine, 2001; Sweetland, 2005).

Language and Educational Policy and Planning

One of the most famous applications of sociolinguistic insight to language and educational policy is documented by Labov (1982), who describes the methods that linguists who testified used to convince Michigan courts that the children of Ann Arbor, Michigan, deserved instruction that was sensitive to their own language varieties. Subsequently, in the Ebonics controversy, linguists worked with school districts, provided statements to the national press, and wrote resolutions to gain favor and recognition for AAE as a legitimate variety (Wolfram, 1998). These methods were mostly modes of recovery as no sociolinguists had been included in the original planning of the program (Delpit & Perry, 1998).

Despite more than 40 years of research, stronger methods are needed to help sociolinguists disseminate information about language and education into the legal and educational systems. Linguists have worked in California since the Ebonics controversy to create curricular materials that are sensitive to AAE and to require additional support for students who use AAE who may have difficulty with phonological awareness and with the structures of standard academic oral and written English (California Curriculum Commission, 2008). The recent introduction of the Common Core Standards, widely implemented across the United States, presents further opportunity for language policy-related materials. For example, Kenji Hakuta and his research group at Stanford (LEEP) are developing Common Core materials that are specifically designed for English language learners (Stanford University School of Education, 2011). Other recent endeavors include the Workshop on the Role of Language in School Learning (Welch-Ross, 2010), which explored language development and its effects on school achievement. Of particular interest was the degree to which group differences in school achievement might be attributed to language differences, and whether language-related instruction might help to close gaps in achievement.

Linguists can also employ specific methods to be more proactive in helping local schools with their language planning. Some of that work merits further documentation here, owing to the split in ideas about what is and is not publishable in linguistics and whether or not efforts in applied linguistics should be separated from other types of linguistic research, both in the publishing arena and in academic departments. In the United States, the National Council of Teachers of English (NCTE), including linguists Geneva Smitherman and Elaine Richardson, has been instrumental in integrating information about language variation into state policy. In another example, the Center for Applied Linguistics has produced materials such as the Sheltered Instruction Observation Protocol (SIOP) program for English language learners (Center for Applied Linguistics, 2011), which is widely used in classrooms throughout the country. Such methods are less known among sociolinguists than, for example, Walt Wolfram's methods for dialect awareness. What started as the eighth-grade Ocracoke curriculum has now been expanded to reflect language variation across North Carolina and is part of the social studies curriculum statewide (Wolfram & Reaser, 2007). Wolfram recognized that social studies curricula are open to materials that reflect local culture, and, using that knowledge, got the curriculum approved. As this example shows, it is important to know the local educational landscape in order to make such natural connections.

Along these policy-setting lines, educational and professional groups have also set language policy. In the United States, the American Speech and Hearing Association has set sociolinguistically informed guidelines for diverse speakers that are specific to those who seek to assess and diagnose diverse speakers in a standardized manner. Stockman (1996) notes that establishing language norms is crucial to helping African American children who may be language impaired as against children who are merely speaking their home variety (see also Green, Vignette 17a). In another example, the assessment methods created for the Diagnostic Evaluation of Language Variation™ (DELV) were designed to focus on the elements of language that are least likely to vary according to descriptions of AAE. Spaulding, Plante, and Farinella (2006) found that the DELV is only reliable 80% of the time, however, and it should be combined with other measures of language assessment. They urge speech-language pathologists to use a sociolinguistically informed battery of tests including informal assessment of more casual and fluid speech, as it is still difficult to provide accurate assessment and intervention.

Collecting Data: Current Methods in Schools

1. *Ethics.* When undertaking research in schools, it is very important to be familiar with the guidelines set forth by the Linguistic Society of America (2009) but also those used by researchers affiliated with organizations such as the American Education Research Association (2011), the American Psychological Association (2011), and the policies of local schools and school districts, which may vary in standard and practice. Depending on the information gathered, the 1996 Health Insurance Portability and Accountability Act (HIPA) (US Department of Health and Human Services, 2011) and the Family Educational Rights and Privacy Act (FERPA) (US Department of Education, 2011) also may protect student information. In addition, sociolinguists must often obtain approval from school principals and district administration. It is critical to implement comprehensive, ethically sound methods to protect educators who often do not have the sovereignty in their positions that sociolinguists do. The time that educators spend working with researchers should also be compensated at a rate commensurate to their salary or with a trade in equivalent resources.

2. *Qualitative methods.* Qualitative methods used by sociolinguists have often centered on the ideological role and social function of language in school and how educational, social, and emotional factors relate to language variation and language ideology. Ideological and communication-related issues may also arise as teachers of one language background teach material and encounter students from other, different language backgrounds (Cross, DeVaney, & Jones, 2001). Myth-busting approaches, in which sociolinguists dispel incorrect assumptions or beliefs held by educators about language, have been common among sociolinguists who wish to address the need for ideological change on the part of educators (Mallinson & Charity Hudley, 2011). A stronger methodology would include the presentation of more relevant facts and facts showing how language variation impacts different aspects of school. This critical process more comprehensively builds educators' language awareness, understanding, and application to their own local classroom and school settings.

Qualitative methods that involve educators and students have allowed sociolinguists to get a deeper sense of students, schools, and communities in ways that are arguably overlooked in more general assessments of speech communities. Labov (1995) notes that students' confidence in the alphabet can affect their educational success. Other research, including Paris (2011) and Ferguson (2001), uses qualitative methods to investigate teachers' and students' language and language use. The anti-bullying movement has also brought attention to the language of bullying (Teaching Tolerance, n.d.). Another area that warrants more research is the analysis of behavioral records and what are often nebulously reported as behavior infractions to measure the degree of language ideology and linguistic conflict that were involved, following Kochman (1983).

3. *Quantitative methods.* Quantitative, school-focused research has used surveys, including Likert-type scales, often with follow-up interviews, to measure educators' language attitudes (Blake & Cutler, 2003). Surveys have also been used to discern what relationship the language ideologies of educators might have to students' school experiences and to help educators grasp critical linguistic concepts about the role of language in the reading, writing, and school socialization process.

Quantitative methods have also been used to measure students' use of particular linguistic features and correlate such use with their achievement on standardized tests. Charity, Scarborough, and Griffin (2004) offered a glimpse of students' most formal language, as used in sentence imitations, and compared it to their reading scores on the Woodcock Johnson Reading Mastery Test. In another example, J. M. Terry, Hendrick, Evangelou, and Smith (2010) related the use of third person singular *-s* to increased challenges with mathematics word problems. Labov (1995) correlated spontaneous speech samples with reading errors using his DX reading program, and Labov and Baker (2010) questioned traditional methods of analyzing reading errors, linking many presumed errors to language variation. Dialect density measures have also been popular, especially among speech and hearing scientists, to provide rough estimates of the use of non-standardized features where comparable assessments across a large body of students are needed (Language Development and Disorders Laboratory, 2007).

Theory on the effect of the cognitive load of language difference and language variation is central to the use of sentence imitation as a sociolinguistic measure (Fraser, Bellugi, & Brown, 1963; Radloff, 1991). The theory states that the limitations of working memory intersect with the evaluation of language and other academic assessments. In their work with adolescent boys in Harlem, Labov et al. (1968) give examples from three types of stimuli (memory tests) to determine whether there was an interference of cognitive load. Most of their findings focused on syntactic and morphosyntactic mismatches in the structure of Standard American English (SAE) and African American Vernacular English (AAVE). In my own research (Charity, 2005), I extended the model to include a more elaborate examination of story retelling. At the end of the sentence repetition task, I had the teacher turn back to the start of the story and tell the child, "Now tell the story back to me. Tell me everything you remember about what happened. Do the best you can." The testing instructions directed the teachers not to ask specific questions about the story and to prompt the child only with the words "anything else?" and with positive reinforcement. The story retells are coded with the Urban Minorities Reading Project coding system (Labov, Charity, & Robin, 2002), and aligning the studies allowed for comparison that provided greater impact for policy making based on the findings.

More research is needed that combines the insights of sociolinguistic methodology with school-based assessments. Information from school records and materials can tell us a great deal about the use of language in instruction. Running records for reading errors and writing samples are very popular with educators and could shed light on the intersection of language and school performance. While such investigations might seem obvious, no such studies exist.

It is also important to pay attention to the differences in the acceptable statistical methods between sociolinguistics and education research for quantitative analyses and publications to be usable and relevant across disciplines and in policy making. There is a need for more random sampling if findings are to be most applicable to schools and larger entities. Within school settings, methods for determining demographic and other important student-related information, such as test scores, may be limited. When conducting broad quantitative studies, sociolinguists must work with schools to gather more specific data than what might be gathered just by asking students themselves and to use data reported by parents or guardians, teachers, and school administrators to ensure that the data are most valid, reliable, and comparable.

4. *Mixed methods.* Comprehensive, mixed methods studies are few, and more are needed. For example, mixed methods studies could examine how educators' own linguistic patterns correlate with those of their students and with their linguistic ideologies to lend insight into mismatches between production and perception. Such studies could also examine educators' ideas about the role of language in materials and information that are mandated and in materials that they choose to use to see where systematic change or pedagogical insight might be helpful. Mixed methods studies involving students could provide a snapshot of a school, class, or social group within a school or school system. Examinations of student performance in specific contexts (e.g., in science, technology, and math classrooms, in oral reports, in specific teaching materials) are greatly needed. Studies that aim to correlate the language of educators with that of students could shed light on the interaction of educators and students over time in relation to student achievement and school-based ideology. Such work would give sociolinguists better insight into the role of school in the language socialization process and in the development of even the most basic sociolinguistic principles.

In public school settings in particular, it is also important to note what researchers might learn about methodology from seemingly non-compliant and taciturn students and educators. How they would be counted in a linguistic analysis is quite a challenge since most of the time they would be excluded from the examination altogether. From an academic standpoint, non-compliance could be marked as 100% failure, since a standardized test administrator might not attempt to seek alternative means of measurement for such children. Taciturn children are often referred to special education, and if they do stay in the mainstream classrooms they may also be labeled as non-compliant or as "problem students." Sociolinguistically speaking, there is much to learn from students whose language use or silence has marked them as needing academic separation or remediation.

Methods for the Future

In the education literature, the framing of linguistic variation is marked by the term "culturally and linguistically diverse" (CLD) populations. It would benefit sociolinguists to integrate our work with research in education that centers on CLD populations. For example, Klingner et al. (2005) give special consideration to accurate assessment of student language in exceptional populations, a topic on which research is particularly scarce. Experimental methods to determine the relationship of language variation to academic success for CLD students would also allow for greater insight into the role of language in the educational process.

Recent general methodological trends in sociolinguistics are also applicable to school settings. In the traditional participant-observer anthropological approach, researchers visit a community for an extended period of time and then return to their home community to spread the knowledge to the rest of the world. In later models, anthropologists became more involved in the daily life and social mechanisms of the community over a longer period of time. In order to help solve social problems in a community on a larger scale, scholars often need to have insider knowledge and use social connections to help implement change in the community – starting with the community of the university. Rather than following a model of engagement that is built around one-time lectures in communities or schools, sociolinguists must work with scholars in education and related fields as well as with educators to create methods for disseminating sociolinguistic information that are sustainable over time.

Sharing data across linguistics and education programs, publications, and outreach is an underused methodological approach. Such an approach would streamline the integration of sociolinguistic insights into educational research and practice and would also allow for educational insights to be integrated in sociolinguistic approaches. Along these lines, Mallinson and Charity Hudley (2010) compel linguists and other academics to partner with educators in CLD schools, districts, and communities; disseminate accurate linguistic knowledge to educators of CLD students; explore best practices for communicating linguistic information to educators; assess the results of providing linguistic training to educators; and apply these findings to educational policy. Most critically, sociolinguists cannot just "drop in" and do education-related research and outreach in an effective and wide-scale manner. The approaches will seem disjointed, and the initiatives will often fail or be tainted with misunderstanding, as the Ebonics controversy and others suggest. Successful initiatives depend on building local alliances – for example, with just one colleague in education at a local college or university, one local organization, or one school. Another critical step is to find out who makes the decisions about educational changes in a given school, school district, city, state, or country and start with them to effect school and/or governmental policy change. Such partnerships not only give sociolinguists an immediate, broad social network but also allow us to understand local and political contexts (see also Serpell, Vignette 17b, and Starks, Vignette 17c).

It is also critical for sociolinguists to form partnerships with colleagues in schools of education at our colleges and universities. When doing so, we must discuss what is taught to linguistics students about education, culture, and diversity; what is taught to education students about language, culture, and diversity; and what is taught to everyone else, at undergraduate and graduate levels. Too often, linguistics departments have structurally disregarded education, while education departments have tended to view linguistics as too irrelevant or abstract to include in their general curricular offerings.

Collaboration with colleagues in education and related disciplines also helps connect sociolinguists with programs that are already established in schools and communities, especially those that provide professional development to educators and services to students. For sociolinguists to undertake such big projects alone would require massive amounts of funding and infrastructure. Instead, through partnerships, linguistic knowledge can be disseminated as part of existing development and outreach. Other faculty who work in schools of education and policy fields can also provide sociolinguists with critical information regarding the most effective practices for communicating to educators, such as helping figure out who is the best first contact in a school or community. It is also important to reach out to broader networks of educators, especially at community colleges, historically Black colleges and universities, and Latino-serving institutions. Most colleges and universities have informal and formal partnerships, so it is important to seek them out and find ways to participate. Research across campuses and with communities can also be more comprehensively connected. Service learning and community engagement initiatives can help with such connections (Charity Hudley, 2010), as can working on interdisciplinary applied research in school contexts with colleagues from related disciplines, especially psychology, anthropology, and sociology. Changes to the National Science Foundation's (2007) mandates for broader impact will hopefully spur more research of this nature. It is also important to produce publications and materials that have impact across fields.

Students and faculty alike should also look to research from related fields including special education, speech-language pathology, writing and rhetoric, and ethnic studies. Undergraduate students should be encouraged to seek double majors and to obtain teaching certification, even for research-focused students. On the graduate and postgraduate levels, students should take courses and obtain postdoctoral fellowships that help train junior scholars across disciplines; they should also be encouraged to obtain teaching certification and to do research and outreach that involves educators from the beginning, not just later in their careers. Sociolinguists at all stages of their careers should be encouraged to mix it up! Faculty should also be incentivized to step outside of their professional boxes. Co-teaching courses and lecturing across disciplines regularly can foster synergy across disciplines. Sociolinguists can also attend conferences designed for education researchers and classroom teachers; in the United States, the National Council of Teachers of English and the ASCD (formerly the Association for Supervision and Curriculum Development) hold important annual conferences.

Conclusion

Stronger networks across sociolinguists, other scholars, and educators are needed to design research that addresses theoretical questions and practical concerns so that findings are applicable across schools and communities. Such methods are the best way to build on previous insights and approaches, rather than inadvertently reinventing the research wheel. In a concrete sense, sociolinguists are still working to address those education-related goals of assessing linguistic variety and addressing educational needs that were the impetus for early sociolinguistic work. As this chapter and this volume suggest, the real challenges in the next generation of sociolinguistic research will be in the comprehensive “how to” in order to achieve maximal impact and relevance.

References

Alvarez, L., & Kolker, A. (Producers/Directors). (1988). *American tongues* [Motion picture]. PBS. United States: Center for New American Media.

American Education Research Association. (2011). AERA code of ethics. Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/CodeofEthics/tabid/10200/Default.aspx

American Psychological Association. (2011). Ethical principles of psychologists and code of conduct. Retrieved from http://www.apa.org/ethics/code/index.aspx

Baugh, J. (2001). Applying linguistic knowledge of African American English to help students learn and teachers teach. In S. L. Lanehart (Ed.), *Sociocultural and historical contexts of African American English* (pp. 319–330). Amsterdam: John Benjamins.

Blake, R., & Cutler, C. (2003). African American Vernacular English and variation in teachers’ attitudes: A question of school philosophy? *Linguistics and Education, 14*(2), 163–194.

Bourdieu, P. (1990). *Reproduction: In education, society and culture*. London: Sage.

California Curriculum Commission. (2008). K-8 Reading/Language Arts/English Language Arts Criteria. Retrieved from http://www.cde.ca.gov/ci/rl/im/documents/aavestatementlabov.doc

Center for Applied Linguistics. (2011). Sheltered instruction observation protocol. Retrieved from http://www.cal.org/siop

Charity, A. H. (2005). Dialect variation in school settings among African American children of low-socioeconomic status. (Unpublished doctoral dissertation). University of Pennsylvania, Philadelphia, PA.

Charity Hudley, A. H. (2009). Standardized assessment of African-American children: A sociolinguistic perspective. In M. Farr, L. Seloni, & J. Song (Eds.), *Ethnolinguistic diversity and education: Language, literacy and culture* (pp. 167–193). New York: Routledge.

Charity Hudley, A. H. (2010). Cultivating socially minded linguists: Service learning and engaged scholarship in linguistics and education. Paper presented at the American Dialect Society conference. Baltimore, MD.

Charity Hudley, A. H., & Mallinson, C. (2011). *Understanding English language variation in U.S. schools*. New York: Teachers College Press.

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with School English in African American children and its relation to early reading achievement. *Child Development, 75(5)*, 1340–1356.

Cran, W., Buchanan, C., & Anthony, S. (Producers), & Cran, W. (Director). (2005). *Do* *you speak American?* [Television series]. Princeton, NJ: Films for the Humanities and Science.

Cross, J. B., DeVaney, T., & Jones, G. (2001). Pre-service teacher attitudes toward differing dialects. *Linguistics and Education, 12*, 211–227.

Delpit, L., & Perry, T. (Eds.). (1998). *The real Ebonics debate: Power, language, and the education of African-American children*. Boston: Beacon Press.

Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. New York: Teachers College Press.

Ferguson, A. A. (2001). *Bad boys: Public schools in the making of black masculinity*. Ann Arbor: University of Michigan Press.

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of Verbal Learning and Verbal Behavior, 2*, 121–135.

Godley, A. J., Sweetland, J. Wheeler, R., Minnici, A., & Carpenter, B. D. (2006). Preparing teachers for dialectally diverse classrooms. *Educational Researcher, 35*, 30–37.

Hoover, M., Politzer, R., & Taylor, O. (1995). Bias in reading tests for black language speakers: A sociolinguistic perspective. In A. G. Hilliard (Ed.), *Testing African American students* (2nd ed., pp. 51–68). Chicago: Third World Press.

Klingner, J., Artiles, A., Kozleski, E. B., Utley, C., Zion, S., Tate, W., & Riley, D. (2005). Conceptual framework for addressing the disproportionate representation of culturally and linguistically diverse students in special education. *Educational Policy Analysis Archives, 13*, 1–38.

Kochman, T. (1983). *Black and white styles in conflict*. Chicago: University of Chicago Press.

Labov, W. (1972a). Academic ignorance and black intelligence. *Atlantic Monthly*. Retrieved from http://www.theatlantic.com/past/docs/issues/95sep/ets/labo.htm

Labov, W. (1972b). *Language in the inner city: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1982). Objectivity and commitment in linguistic science: The case of the Black English trial in Ann Arbor. *Language and Society, 11*, 165–202.

Labov, W. (1995). Can reading failure be reversed? A linguistic approach to the question. In V. L. Gadsden & D. A. Wagner (Eds.), *Literacy among African-American youth: Issues in learning, teaching, and schooling* (pp. 39–68). Creskill, NJ: Hampton Press.

Labov, W., & Baker, B. (2010). What is a reading error? *Applied Psycholinguistics, 31*, 735–757.

Labov, W., Charity, A. H., & Robin, E. (2002). *Urban minorities reading project scoring system*. Philadelphia: US Regional Survey.

Labov, W., Cohen, P., Robins, C., & Lewis, J. (1968). *A study of the nonstandard English of Negro and Puerto Rican speakers in New York City*. Cooperative Research Report 3288. Vols. I and II. Philadelphia: US Regional Survey (Linguistics Laboratory, University of Pennsylvania).

Language Development and Disorders Laboratory. (2007). Measures of dialect density. Retrieved from http://www.lsu.edu/cdlab/training/dialect_density.html

LeMoine, N. (2001). Language variation and literacy acquisition in African American students. In J. Harris, A. Kamhi, & K. Pollock (Eds.), *Literacy in African American communities* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum.

Linguistic Society of America. (2009). Linguistic Society of America Ethics statement. Retrieved from http://lsadc.org/info/pdf_files/Ethics_Statement.pdf

Mallinson, C., & Charity Hudley, A. H. (2010). Communicating about communication: Multidisciplinary approaches to educating educators about language variation. *Language and Linguistics Compass, 4*, 245–257.

Mallinson, C., & Charity Hudley, A. H. (2011). How K–12 educators apply sociolinguistic knowledge in the classroom. Paper presented at the New Ways of Analyzing Variation 40 conference. Washington, DC.

National Science Foundation. (2007). Merit review broader impacts criterion: Representative activities. Retrieved from http://www.nsf.gov/pubs/gpg/broaderimpacts.pdf

Paris, D. (2009). "They're in my culture, they speak the same way": African American language in multiethnic high schools. *Harvard Educational Review, 79*(3), 428–447.

Radloff, C. F. (1991). Sentence repetition testing for studies of community bilingualism: An introduction. *Notes on Linguistics, 56*, 19–25.

Reaser, J., Temple Adger, C., & Hoyle, S. (2005). Using *Do You Speak American?* for educator training and professional development: Guide and materials. Retrieved from http://www.pbs.org/speak/education/training/

Rickford, J. R. (1997). Unequal partnership: Sociolinguistics and the African American speech community. *Language in Society, 26*, 161–197.

Simpkins, G. A., Holt, G., & Simpkins, C. (1977). *Bridge: A cross-cultural reading program*. Boston: Houghton Mifflin.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech and Hearing Services in the Schools, 37*, 61–72.

Stanford University School of Education. (2011). A language project building on the common core state standards and the next generation science standards initiatives. Retrieved from http://www.stanford.edu/group/ell/cgi-bin/drupal/

Stockman, I. J. (1996). Phonological development and disorders in African American children. In A. G. Kambi, K. E. Pollock, & J. L. Harris (Eds.), *Communication development and disorders in African American children: Research, assessment, and intervention* (pp. 117–153). Baltimore: Paul Brookes.

Sweetland, J. (2005). Evaluation of contextualized contrastive analysis in language arts instruction. (Unpublished doctoral dissertation). Stanford University, Palo Alto, CA.

Teaching Tolerance. (n.d.) Bullying basics. Retrieved from http://www.tolerance.org/supplement/bullying-basics

Terry, J. M., Hendrick, R., Evangelou, E., & Smith, R. L. (2010). Variable dialect switching among African American children: Inferences about working memory. *Lingua, 120*(10), 2463–2475.

Terry, N. P. (2008). Addressing African American English in early literacy assessment and instruction. *Perspectives on Communications Disorders and Sciences in Culturally and Linguistically Diverse Populations, 15*(2), 54–61.

US Department of Education. (2011). Family Educational Rights and Privacy Act (FERPA). Retrieved from http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

US Department of Health and Human Services. (2011). Health information privacy. Retrieved from http://www.hhs.gov/ocr/privacy/

Welch-Ross, M. (2010). Language diversity, school learning, and closing achievement gaps: A workshop summary. Retrieved from http://www.nap.edu/catalog.php?record_id=12907

Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Washington, DC: Center for Applied Linguistics.

Wolfram, W. (1998). Language ideology and dialect: Understanding the Oakland Ebonics controversy. *Journal of English Linguistics, 26*(2), 108–121.

Wolfram, W., & Reaser, J. (2007). *Voices of North Carolina dialect awareness curriculum*. Raleigh: North Carolina Language and Life Project.

Vignette 17a
Beyond Lists of Differences to Accurate Descriptions

Lisa Green

One topic that is understudied but is beginning to receive more attention is the development of language use by children growing up in non-mainstream American English speech communities. Children developing such varieties of American English invariably produce structures that are identical to those in Standard American English, structures that are subtly different from those in the standard, and structures that are maximally different from them, and complete and realistic descriptions of the varieties include all of these properties.

More research on developmental patterns and norms of language use by children in these speech communities will provide insight into questions beyond those about the extent to which these patterns differ from the standard. Such research has important classroom implications and application. Given that non-mainstream dialects of American English have tended to be characterized strictly by the way they differ maximally from Standard American English, with little or no attention being paid to the similarities and subtle differences, actual non-mainstream American English speech is often described dichotomously, as either reflecting one kind of feature or another, each type being associated with the standard *or* non-standard. For instance, an African American English (AAE)-speaking child who uses a feature that is not associated with Standard American English is said to be speaking AAE at that point in time, but the instant the same child produces a feature that is also compatible with the standard, she or he is seen as having shifted into Standard American English.

Although it provides a concrete way to characterize a speaker's language use or a way to identify when a child displays more "dialect" use on the one hand and more "standard language" use on the other (where more is represented by a simple count of features, not by reference to the nature of linguistic structures), this dichotomous approach is misleading. From the standpoint of anyone having to make the determination that a student is a dialect speaker and having to quantify what makes the child a dialect speaker for instructional purposes, it makes sense to be able to identify, count, and to assign concrete features to one variety or another, the standard or non-standard. In this way, it is useful to refer to features of AAE and other dialects of American English that are maximally different from those in standard English as a means of underscoring the fact that the child speaker systematically uses a variety that can be identified – in part – by distinct features that are different from those in classroom American English.

However, descriptions based on lists of maximally different features are only useful if their limitations are clear. Such descriptions do not (1) represent a speaker's system of language use or the speaker's knowledge of his or her linguistic system, or (2) admit any overlap between the non-standard dialect and the standard variety, or (3) separate developmental dialectal patterns from adult patterns.

There is a problem, but the problem is not just with the lists themselves; the problem is related to the way the lists of maximally different features are understood and the conclusions that are drawn from them. When such descriptive lists are used for purposes other than as a reference to the stark differences between the standard variety and non-standard varieties, then there is a risk of substituting the lists for the linguistic systems, but they are not good substitutes for linguistic systems. They fall short in indicating what speakers actually know about their language varieties. One of the major shortcomings of the focus on descriptive lists is that it significantly detracts from opportunities to give a complete picture of what constitutes the non-standard variety. That is, the distinguishing features are underscored, but there is no hint about ways in which they are manifested systematically in speech in well-defined linguistic environments. Along these same lines, there is no mention of the overlap between features of a non-standard variety of American English and standard English, such that, in certain linguistic environments, the non-standard variety exhibits patterns that are identical to those in the standard. As it turns out, there are ways in which features from the two are identical. Consider the AAE feature zero copula exemplified in the following example:

> She Ø$_{copula}$ not tall. ('She's not tall')

The overt copula is very much a feature of AAE but never gets counted as one in the lists of maximally distinct AAE features. Speakers of AAE would most certainly use the overt form, as shown in the example below, to emphasize the affirmative of the negative statement (i.e., *She Ø$_{copula}$ not tall.*) above.

> She **is** tall.

Developing AAE-speaking children also almost always produce the overt copula when it is the final word in a sentence. Consider the following example from a five-year-old developing AAE speaker, reported in Green (2011, p. 41):

> I know what color this is.

Both the zero copula (*Ø$_{copula}$*) and overt copula (e.g., *is*) are part of the AAE grammar. Children use the overt form as part of their native AAE, not necessarily as an instance of shifting from AAE zero copula to the standard overt copula. As such, AAE and standard American English are similar in that they both use overt copula. A complete description of AAE must also include overt copula as a feature alongside zero copula.

Two additional cases in which the focus on "difference" makes it difficult to consider the full range of AAE patterns, including those that might overlap with standard English, are habitual *be* (as in *They be running.* 'They are generally running') and zero third person singular *-s* marking (as in *She leave early.* 'She leaves early'). Both of these constructions indicate habituality or that an event occurs with some regularity. That is, AAE speakers use habitual *be* as well as plain present tense verb forms such as 'leave' to indicate the regular occurrence of an event. Feature lists draw our attention to the fact that in AAE, verbs are not generally marked for third person singular agreement, but what is overlooked is that those verb forms can also indicate habitual meaning. Associating this feature with verbs is important for a number of reasons. It helps give a clearer picture of the way habituality is marked in AAE and highlight the pattern speakers might use when they do not use habitual *be* to mark events that occur regularly. It is reported in Green (2011) that four- and five-year-old developing AAE-speaking children do have some knowledge of habitual *be*, and some children use it in spontaneous speech, but they have not mastered the marker at that stage. In the example below, Rayna, a five-year-old developing AAE-speaking girl, expresses habituality with plain present tense verbs:

> And they pi—and guess what? They blow bubbles on-on-on Squidward when Squidward sleeping. And when Squidward blowing, they they pick the house up with a blowing with a blowing thing.
>
> (p. 41)

On the other hand, Akila, a five-year-old developing AAE-speaking girl, uses both a plain present tense verb and aspectual *be* to mark what appear to be habitual contexts:

> Cause when I when I watch Blues Clues, my eyes be like this.

In order to get this information, linguists have to move beyond lists of differences from the standard to methods of analyzing datasets for systematic language use.

More than the absence of *-s* is noteworthy in relation to plain present tense verbs. Sometimes these verbs are used with plural subjects, so third person singular *-s* marking is irrelevant. It is useful to have information about habitual marking in AAE because listeners do not always perceive it in AAE. One clear example is illustrated by an area high school teacher during a workshop. In explaining the meaning of habitual *be* in AAE, I noted that the sentence *I be tired* actually conveys the meaning that the speaker is tired from time to time. I noted that the sentence means something like 'I'm usually tired'; however, the teacher expressed firmly and confidently that I was wrong. At the start of the second half of the workshop, the teacher reported to the group that he had been able to talk with one of his students during the lunch period and confirm that *I be tired* does indeed refer to recurrence of being tired, not just being tired at the present moment. I imagine that the main reason the teacher found it so difficult to accept the story about the habitual meaning was that the sentence *I be tired* is very close in form to *I am tired*, a sentence that is more familiar to him, and it was hard to abandon the more familiar meaning.

Finally, when single descriptive lists of maximally distinct features are offered, it is never clear whether the features are representative of adult language or are a conglomerate of adult and child language. For example, negative concord occurs in many non-standard varieties of English, and it also occurs in the speech of children who are developing standard as well as non-standard American English (Bellugi, 1967). The following negative concord structure was produced by a four-and-a-half-year-old child growing up in an AAE speech community, as reported in Green (2011, p. 124):

> I don't have no training wheels.

When children from AAE-speaking communities who are in developmental stages produce negative concord, is it representative of an AAE feature or universal language development? As things stand, there is no guide to the way language development factors into identification of dialects, so it is not always clear which features mark dialects in the developmental phases or full-fledged adult language. The conclusion is that it is useful to underscore these apparent dialectal features, but researchers are also responsible for considering them in the context of realistic – not idealistic – and age-appropriate language use for developing AAE speakers.

Information about developmental patterns is crucial in establishing accurate descriptions of language use by children growing up in speech communities in which varieties of non-mainstream American English are the norm. Such descriptions, which go beyond lists of maximal differences, will contribute to research in linguistics and will be useful in areas of practical application in education. For instance, given claims about the effect of language use on academic achievement, it is important to understand that child speakers of dialects also move through developmental stages on their way to the mature linguistic system. They do not begin with idealistic adult AAE features, so to look at their language as one would look at adult AAE is problematic. While they may not have mastered Standard American English, these speakers are developing the linguistic systems in their speech communities, which should also be taken into consideration, especially because their own linguistic systems will likely be factors in use and mastery of other systems.

The following generalizations might be useful guidelines in thinking about developmental language patterns. (1) By age three, children are beginning to develop unique linguistic patterns in their speech communities. (2) In early stages of language development, some language patterns reflect general child language development and others reflect patterns that are consistent with adult speech in the children's communities. (3) Children show consistent patterns of development of the language in their communities, and some of these patterns differ from those associated with Standard American English (or the variety that will be used and accepted in the classroom); other patterns are identical to those in the standard variety.

More accurate descriptions of early dialectal patterns could prove useful for developing educational instruction in pre-school and elementary school programs, and they could significantly increase our understanding of speakers' progression of language use – especially given claims about their increase and decrease of dialect throughout school. One way to move toward more accurate descriptions is to expand lists of features to include information about different strategies children have for indicating a certain meaning, such as habituality, that might lead to a better description of a child's overall system of AAE use – not just a list of the way it differs maximally from the standard.

References

Bellugi, U. H. (1967). The acquisition of the system of negation in children's speech. (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.

Green, L. J. (2011). *Language and the African American child*. Cambridge: Cambridge University Press.

Vignette 17b
Linguistic Flexibility in Urban Zambian Schoolchildren

Robert Serpell

Educational policy is sometimes informed by the stereotype of a language as a discrete and largely autonomous system of rules. In a young nation-state that incorporates a variety of sociocultural groups, the medium of instruction in basic schools can become a focus of contention that is intensified by such simplistic assumptions. The study described in this vignette illustrates the complexity and fluidity of the linguistic repertoire needed by children to achieve full communicative competence in the multilingual African city of Lusaka. Sociolinguistic analysis of Zambian society in the second decade of its political independence showed that two of the indigenous Bantu languages had become established as lingua francas for the two major multiethnic urban areas: Bemba had become the lingua franca of the Copperbelt, and Nyanja was the lingua franca of the fast-growing capital city of Lusaka. Individual multilingualism was widespread in the adult population, with urban residents claiming fluency in an average of 2.8 different languages, one of which was usually English and another was one of the lingua francas (Mytton, 1974).

Informal observation of discourse patterns in Lusaka suggested to me that, as in many other postcolonial states, Zambia's speech community had a stratified repertoire, with the official language of the former colonial administration, English, assigned many of the sociolinguistic functions of the H(igh) code described by Ferguson (1959) in his classic account of diglossia, while the various indigenous Bantu languages served most of the functions of the L(ow) code. In the first decade following the declaration of political independence from Britain, the Zambian government introduced a national policy of immersion in English for all children entering first grade. By the end of the decade, a number of undesirable consequences of this English-medium policy were receiving attention, including a high prevalence of outright failure to achieve basic literacy by the fourth grade (Serpell, 1978). A "national debate" was proclaimed by the government in 1976–1977 to inform a comprehensive process of "Educational Reform," including as one of several key topics the issue of medium of instruction.

I conducted the study described in this vignette in the context of that national policy debate. My objectives were (1) to demonstrate the feasibility of using several languages concurrently in the first-grade classrooms of Lusaka's public schools catering to a multiethnic urban population, and (2) to explore the situational and cognitive influences on young urban children's communicative competence in three linguistic codes at the beginning of their school careers. To attain these objectives, I designed a field experiment to simulate systematically some naturally occurring speech varieties and discourse practices. Scribner (1976) has ably defended the use of such "situated experiments" as resources for cross-cultural research, when interwoven with ethnographic research. A compelling example is her classic study with Cole of the character of three different cultural practices of literacy in Liberia (Scribner & Cole, 1981).

The present experiment combined an *ex post facto* design, comparing children from different home language backgrounds, with a repeated measures, within-subject design. The independent variables of linguistic medium and conceptual domain test were systematically varied as orthogonal factors, and the dependent variable was the degree of communicative competence displayed by the child. Forty-two boys and girls about eight years old were purposively recruited from first-grade classrooms at three public primary schools in Lusaka based on self-report that their home language was either Nyanja or Bemba. Each child was given three short tests individually in a fixed sequence. The first test (Information) consisted of questions about the child and her or his home; the second test (Play) consisted of phrases, mainly simple commands, that might be used among children playing together (with a ball, and so on); and the third test (Picture) consisted of questions about the content of a picture such as might be asked by a teacher in the course of a lesson. The tests were administered by an experienced female schoolteacher whose home language was Nyanja and who was also fluent in Bemba and English. Each test was administered in a different language, and the six possible permutations of the three languages – Nyanja, Bemba, and English – were counterbalanced across participants. The switches from one language to the next were not announced or explained, but followed a brief natural pause as the tester turned from one page of the schedule to the next and changed the topic from the conceptual domain of Information to Play or from Play to Picture. The tester, Phides Nguluwe, brought to the task a highly developed set of communicative skills for interacting with young children that seemed to me, as an observer, to put the children at ease, lending validity to this method of systematically sampling various registers of contemporary speech in Lusaka.

Based on naturalistic observations, we proposed the following hypotheses. (1) Lusaka schoolchildren would accept without question the switching by a teacher from one code to another, and they would often reply to the teacher in a different linguistic code from that in which the question was phrased. (2) Different conceptual domains would be handled by the children more efficiently in one code than in another. More specifically, (2a) grade 1 children of Nyanja- and Bemba-speaking families in Lusaka would handle questions about home and playful interaction better in Nyanja and Bemba than in English, and (2b) these children's command of English would be better in dealing with pictures than in answering questions about home or in the realm of play. (3) Grade 1 children from Bemba-speaking homes in Lusaka would have a better command of Nyanja than of English. (4) Children from Nyanja-speaking homes in Lusaka, where the lingua franca is Nyanja, would not have as good a command of their second language, Bemba, as the command of Nyanja shown by the children of Bemba-speaking families.

Our results confirmed hypothesis (1). A total of 82 code switches were recorded from the 42 children. Although these constitute only a small proportion (about 9%) of the 924 occasions on which switches could have occurred, it is noteworthy that the switches were not confined to a few individuals: 33 of the 42 children switched at least once. Detailed analysis revealed two major contexts in which the children switched code: first, the use of English words to answer questions posed in an indigenous language about the picture, which they had met before in English-medium lessons; and second, semantically correct answers in the child's home language to a question posed in another language.

Statistical analysis of the test scores by children of each family-language sample yielded significant support for each of hypotheses (2), (3), and (4), except that relatively high scores in English were recorded on the Play test. Two complementary factors may help explain this finding. First, the atmosphere of the Play test, although it elicited a number of smiles from the children, was clearly not truly analogous to the play situations with which they were familiar. Second, a number of instructions in this test (e.g., "come here," "show me…") were almost certainly included in these children's English classroom exercises.

Our study clearly demonstrated, among first graders in Lusaka, that for a number of purposes, Nyanja and Bemba speakers (as defined by their parents' ethnicity) found each other's languages more effective media of communication than they did English. Arguably, the social framework in which we elicited the children's speech was too asymmetrical to be truly representative of general language usage in urban Zambia. An interview to which one party brings a predetermined set of questions clearly constitutes an atypical sequence of verbal utterances. Social encounters normally generate a context out of interaction and the framework of discourse is negotiated between the participants. But even in this constrained setting, we were able to document a high degree of flexibility in the strategies of communication adopted by the children when dealing with linguistic forms over which their control was incomplete. Some children spontaneously translated some of the questions put to them into their home language before replying. In the picture test, several children launched into reading aloud the words beside the picture as soon as the book was opened, while others responded to questions about the picture by reciting long stock phrases without regard to their relevance (e.g., Teacher: "Who is this?" Student: "They are sitting on the chair").

Thus, the data collected in this study provided food for thought to educational policymakers as they sought to design a curriculum that is flexible enough to capitalize on the various language skills that Zambian children bring to school, while fostering communicative competence adequate for Zambia's multilingual and rapidly modernizing society. Decisions about public educational policy are often constrained by political factors that extend well beyond the scope of evidence that a technocratic perspective might wish to prioritize. The Educational Reform process in Zambia took more than two years to reach its conclusions, articulated in a final report that departed radically from the interim draft report in several respects, including the medium of instruction for the lower primary grades. Despite a carefully phrased set of recommendations from the University of Zambia citing evidence that included the present study, the policy of immersion in English was retained for a further 20 years before a less public and more evidence-based process of reform was introduced (Tambulukani, Sampa, Musuku, & Linehan, 2001). Current policy mandates initial literacy instruction in the medium of one of the Zambian languages, selected in accordance with prevalent usage in different zones. Ten years on, the initial evidence of improved learning under these conditions has been hard to replicate, and concerns have been raised about the choice of linguistic varieties relative to current usage in various (notably urban) zones. The design of an adequate school curriculum for supporting and extending the communicative competence of children growing up in a multilingual society continues to pose a significant challenge for applied sociolinguistic research.

References

Ferguson, C. A. (1959). Diglossia. *Word, 15*, 325–340.

Mytton, G. (1974). *Listening, looking and learning: Report on a national mass media audience survey in Zambia (1970–73).* Lusaka: University of Zambia Institute for African Studies.

Scribner, S. (1976). Situating the experiment in cross-cultural research. In K. F. Riegel & J. A. Meacham (Eds.), *The developing individual in a changing world* (Vol. 1, pp. 310–321). Chicago: Aldine.

Scribner, S., & Cole, M. (1981). *The psychology of literacy*. Cambridge, MA: Harvard University Press.

Serpell, R. (1978). Some developments in Zambia since 1971. In S. Ohannessian & M. E. Kashoki (Eds.), *Language in Zambia* (pp. 424–447). London: International African Institute.

Tambulukani, G., Sampa, F. R., Musuku, R., & Linehan, S. (2001). Reading in Zambia: A quiet revolution through the Primary Reading Programme. In S. Manaka (Ed.), *Proceedings of the First Pan-African Reading for All conference, August 1999*. Newark, NJ: International Reading Association/UNESCO.

Vignette 17c
Engagement with Schools

Sharing Data and Findings

Donna Starks

Sociolinguists are often under the misconception that because they went to school and/or have accompanied their children to school, they have some understanding of schools as institutions. This is as true as statements such as "all multilinguals are linguists." Research in schools is exhilarating and rewarding and also frustrating, as decisions about whether, when, and how research is to be conducted are sometimes made with lightning speed and at other times left for months in a pile on the principal's desk. Anyone intending to conduct research in schools should acknowledge that schools do not work to university deadlines. At one school, a principal apologized for not yet getting around to distributing a set of questionnaires. The apology came six months after the project had been completed. The opposite occurred in another project that I report on here. It illustrates a common trend in school research: to expect the unexpected in terms of both opportunities and consequences.

Because schools are often a center point of communities, they are one of the best ways to gain access to the wider community. They provide contacts with like-minded souls who share an understanding of academia, with legitimate power holders who know the community and who have community members' respect, and with community language liaisons who are able to translate the language of the researcher into a variety that the community can both understand and relate to. In the Pasifika Languages of Manukau Project, we worked with a primary school as an entry point into a multilingual Pacific and Maori community in South Auckland, New Zealand. The principal at the school had a strong interest in promoting his bilingual programs and, importantly, saw connections between his language programs and our research. Because of this support, he opened up the school hall to run information sessions to introduce our research agenda to the community and later allowed us to use the venue to report back our findings, a key to establishing credibility in the community. He also helped us befriend the teachers in the bilingual classes, who provided links to their respective communities. To help establish and build connections, I regularly "hung out" in the school lunch room during morning tea breaks and brought cake from the local bakery in an attempt to create small talk and thereby build links with the teachers' worlds and their broader community. As my work with the Pasifika Languages of Manukau Project was community based rather than school based, the ensuing school-based research happened by chance.

Funding is a constant issue in schools. At one of the morning teas, I offhandedly mentioned that our Pasifika Languages of Manukau Project had wanted to look at English use in the community, but given funding cuts and strong community interest in a fourth Pasifika language, we could not conduct our research with the dominant European group. Within the span of the remainder of the tea break, the teachers had misinterpreted this as a need to collect English reading data and asked whether I could use the students in the classes for any of my research. In an attempt to clarify my stance but at the same time support their initiatives, I mentioned that I had another small project to set up an English linguistic diversity resource for my phonetics students, and perhaps I could transfer some of the remaining funding from that project for the reading tasks to create a collection of Pasifika student voices. The teachers quickly volunteered their classes the following week. University research usually doesn't happen this quickly, particularly in cases such as this, where we need to make modifications to existing projects in order to include younger participants. But the teachers were not deterred and suggested I take two weeks to organize things along these lines. They also suggested that my research funding be translated into $20 tokens that students could use toward an existing school fund-raising venture to purchase musical instruments for the school orchestra.

Within weeks, I had the requisite university approvals and was somewhat ready, but the teachers were even more so. When I arrived with a 15-year-old Niuean girl who would engage with the students and record the reading data, the classes had already been briefed about the project, and the students were primed both to be involved in the English diversity project and to participate in the fund-raising activity. A classroom had been set aside, and I was presented with a list of all students present on the day, their gender and ethnic backgrounds, and the order in which they were to participate in the recording. The teachers also informed me that they had decided to exclude students younger than 10 years old, so that students who participated in the project could enjoy the reading rather than struggle with it. These decisions as to when and how to conduct the project were controlled by the teachers' schedules and how they felt the research aligned with their class and their classroom activities on that particular day. Although I felt like a bystander in my own project, I went with it. Within 36 hours, I had recordings of 40 students, and the school had funds for a new saxophone. This example illustrates how schools are fast-paced microcosms where decisions that fit within the school curriculum are embraced, once trust has been established. Delighted, I presented the teachers and principals with a copy of the CD containing the student readings to celebrate the different Pasifika voices in their classes. The CD became part of their shared teaching resources.

The vignette does not have a happy ending, however. As in many places, school funding for the bilingual programs is constantly under threat, and the school lost the majority of its bilingual programs the following year when a new principal took over the school administration. She argued that poor English literacy scores at the school needed to be addressed. The school resources, of which the CD I produced was one, were used together with a long list of additional evidence to advocate for greater Standard English literacy and the abolition of most of the bilingual education programs at the school. This experience shows that any research in schools can have macro-level consequences far outside of the researcher's intentions. It also highlights the extra care we need to take when promoting linguistic diversity in classroom-based research.

Both the sociolinguistic ethos and New Zealand university research guidelines for working with indigenous communities encourage us to give language resources back to the participants. In the short term, the CD was a resource, through which student voices were enjoyed and celebrated. Yet in the longer term, the vernacular voices on the CD were used as one piece of evidence to justify ending a bilingual education program for its speakers. This situation raises an interesting dilemma: how can we as sociolinguists give resources to schools to help celebrate student diversity yet avoid presenting schools with resources that can be used for school agendas that focus on the promotion of a single hegemonic Standard English curriculum and consider bilingual programs and linguistic diversity to be a threat? Although it is impossible to think through all ethical dilemmas in advance, we need to be aware of the complexities and political agendas that exist in educational arenas. It is not enough to be attuned to the current views of teachers and principals; we need to be prepared for when political agendas change. In thinking back about what happened, I often wonder whether the outcomes would have been different if I had not provided the CD or if perhaps I had included a short, carefully written description of the importance of vernacular voices on its cover. What might you have done differently?

18 Sociolinguistics in and for the Media

Jennifer Sclafani

Language, Media, and Social Awareness

My first introduction to sociolinguistics came not in a college classroom but in my home as a child, where I was constantly reprimanded by my mother for not pronouncing my "r"s. I remember being confused by her scolding; why did everyone else in my suburban Boston neighborhood – including my father, an English teacher! – seem to drop this sound left and right? "People will think you're stupid if you talk like that," my mother would explain to me.

One day, I asked my father why he dropped his "r"s. "I'm a product of my environment, not my education!" was his retort. He would later invoke this response in reaction to so many of my smart-aleck comments about his vernacular grammar over the years. ("You *says* to her, Dad? Do you teach your students that?" or " 'I should *have gone*,' not 'I should *of went*.' I'm glad you're not the one taking the SATs next week!")

Years after these formative conversations in my family, I read William Labov's (1972a) seminal studies on the sociolinguistics of Martha's Vineyard and New York City department stores. Aside from being astounded by the ingenuity of these studies that empirically presented what I had intuited over the years, these foundational works also made me realize that although my parents' views on language had always seemed to be diametrically opposed to each other, they were actually in a sense both "right." One might say that, in a dialogical manner, my parents taught me what I now believe to be the two most important principles of sociolinguistics: (1) We are judged by others on the basis of the way we talk, and (2) We are linguistic products of our environments. When I left my hometown as a young adult and began to experience new linguistic norms and the social consequences of speaking differently from those in my new geographic and social environments, a third and equally important principle presented itself to me: (3) Through the language we use, we have the power to enact social change.

My research has been driven by these three principles, which have directed me toward the study of language use in a variety of media contexts, from newspaper discourse and televised talk shows to news-sharing websites and online discussion forums. Because of the immediacy, ubiquity, and ever-increasing number of channels through which we interact with the media on a daily basis, these outlets are a powerful source not only for directly informing us about what is happening in the world (or at least for providing a particular interpretation of it), but also for indirectly informing us about the appropriate discourses through which we articulate "facts" and come to understand the "natural order" of things. The media also play a significant role in influencing the ways in which we conceptualize our relationships with other individuals and social groups. In sum, the ideological potential of the media to reinforce or reconstrue the sociolinguistic lessons we learn at home, along with their capacity for enacting social change, has led me to focus my interest in language variation on this particular context of use.

Studying Language and the Media

The relationship between sociolinguistics and the media can be seen as a mutually reflexive one. On the one hand, a discipline that defines its field of study as language in its social context must include a consideration of language use in public and semi-public mediated contexts; on the other hand, the media can be viewed as a megaphone that projects certain types of language – either certain regional, social, and ethnic varieties or broader genres and big "D" Discourses (Gee, 1990) – while muting others. For this reason, it is useful to consider the media not only as an object of study but as a medium through which we can connect with others and communicate critical sociolinguistic concepts and knowledge about the nature of language to a wide range of audiences.

This chapter reviews past work in sociolinguistics that both investigates language use in the media and uses the media to disseminate knowledge about language. I begin by focusing on two specific cases: (1) the media representation of "Ebonics" following the 1996 Oakland School Board controversy, along with linguists' participation in these discourses, and (2) the current debate in the media over terminology to refer to the status of immigrants residing and/or working in the United States without proper authorization (i.e., "illegal immigrants," "illegal aliens," or "undocumented workers"). Both of these cases have drawn the attention of linguists, who have used the media as a source of data through which to study circulating discourses on these topics. In addition, sociolinguists have engaged with the media in an effort to share experts' research-based points of view on these issues with broader audiences. Following these two case studies, I discuss some of the ways in which sociolinguists have proactively made use of the media to disseminate knowledge about language variation, for educating traditional students and the wider public. I conclude the chapter by suggesting areas for further development along these lines of research.

Sociolinguistic Studies of the Media

The analysis of language use in various media contexts has been a growing field in sociolinguistics over the past few decades. Having expanded the scope of inquiry beyond language use in the print and broadcast news (e.g., Bell & Garrett, 1998; Fairclough, 1995; Heritage, 1985; Van Dijk, 1988), sociolinguistic research on the media now includes journals and volumes dedicated to the study of language in a host of new media outlets (e.g., *Journal of Computer-Mediated Communication*; *Language@internet*; Baron, 2008; Crystal, 2006; Thurlow & Mroczek, 2011), especially various internet-based genres of discourse, such as email, chat, blogging/vlogging, social networking, video sharing, and gaming (see also Sadler, Vignette 3d).

Case 1: Ebonics in the Media

Studies that focus on the construction, negotiation, and reinforcement of dominant language ideologies in media contexts, especially ideologies regarding marginalized social groups and their associated language varieties, have been of particular interest to linguists who also wish to use the media to disseminate scientific knowledge about language. African American English (AAE), the most studied and politicized social dialect of American English (Baugh, 2000; Wolfram & Schilling-Estes, 2006), has received a great deal of attention from both linguists and the popular media over the years. Despite the fact that the structural regularities and the distinct pragmatic features of AAE have been studied and documented by sociolinguists for a half-century (some early in-depth studies include Dillard, 1973; Labov, 1972b; Smitherman, 1977; Wolfram, 1969), the acknowledgment of the variety as something other than "slang" or "street speech" still rarely makes it beyond introductory linguistics courses.

Following the highly publicized Oakland Ebonics controversy of 1996, however, the sociolinguistics of AAE was suddenly thrust into the media limelight. Not surprisingly, the focus of research on AAE turned from a descriptivist approach to a critical reflection on the representation of AAE in various media contexts around this time (Baugh, 2000; Perry & Delpit, 1998). For example, Baugh (2000) examines the fallout of the controversy in popular magazines such as *The Economist*, *Newsweek*, and *Mad*, finding that linguistic satire and racist humor outweighed thoughtful reflection on the issue in the press. Ronkin and Karn (1999) and Pandey (2000) have both examined ideologies of Ebonics on the internet, though in contrasting communities, following the Oakland controversy. Ronkin and Karn examine non-linguists' parodies of Ebonics on the web, finding that they overuse a limited number of stereotypical features (which the authors dub "Mock Ebonics") and discursively link features of the variety with the socioeconomic status of many of its speakers. Pandey examines discourse among linguists on the topic of AAE, finding that even experts use a number of "othering" devices in an online discussion thread on the topic, which serves to separate specialist versus lay understandings of the variety. All three of these studies examining various aspects of public discursive representations of Ebonics are united in that they emphasize a disconnect between the empirical study of language variation and the powerful "commonsense" beliefs about the language variety that abound in the world around us.

Rickford (1999) brings together these two spheres of expert and lay discourses on language, taking the perspective of a media insider during this particular debate. He reflects on his own involvement with the media following the Oakland School Board decision and shares lessons he learned, making suggestions about how linguists can take an active role in what he calls "the Great

Language Debates of our Times" (p. 267). Rickford recalls the frustration of many linguists over the negative representation of AAE in the press despite concerted efforts to inform the public (e.g., through the Linguistic Society of America's Resolution and numerous op-ed pieces and letters to the editor written in national newspapers), and he reminds us:

> We seem to have forgotten what advertisers of Colgate toothpaste and other products never forget: that the message has to be repeated over and over, anew for each generation and each different audience type, and preferably in simple, direct and arresting language which the public can understand and appreciate.
>
> (p. 271)

In his dual role as a producer and consumer of media, Rickford sends a message that is an important one to keep in mind when considering how sociolinguists can effectively utilize their understanding of how media discourse works as they construct discourse for the media. In essence, if we do not want to remain on the sidelines of the Great Language Debates of Our Times, we must put into practice our understanding of "audience design" (Bell, 1984) and tailor our messages in ways that effectively communicate with the general public. This idea is emphasized by Kiesling in Vignette 18a, who advises us to keep in mind that we must tailor our points (that may have been made in the span of an article or a book), to be reiterated in a sentence of an article or a 20-second sound bite of a news program. Furthermore, Laforest (1999), whose work on the Québécois French has been featured in Canadian-media language debates, emphasizes that we must not only think about simplifying our points but also consider our affective stance and tone when interacting with the media:

> But above all, TV has to entertain. The winner of a TV debate is the one who knows how to be melodramatic, make people laugh or cry; it's a game that academics, with nothing but their theses, arguments and counterarguments to draw upon, are usually not very good at.
>
> (p. 278)

In essence, from the Oakland Ebonics controversy and other similar language debates, we have learned that if we want specialist discourses on language to be projected rather than muted by the media megaphone, academics must step out of their comfort zone of detached scholarly objectivity and invest as much emotional charge into these debates as non-specialists do.

Case 2: Immigration Terminology in the Media

The Ebonics controversy is of course not the only great language debate during which linguists' views on language have been mediated to lay audiences. Quite recently, the issue of US immigration policy has received a great deal of attention in the press, especially because it was a central issue distinguishing the candidates vying for the Republican bid for the 2012 US presidential election. Sociolinguists have examined the representation of immigration in the media extensively, especially in newspaper discourse, in a number of different national contexts. For example, Santa Ana's (2002) monograph, vividly titled *Brown Tide Rising*, details an extensive study on the metaphoric representations of Latinos in newspaper coverage of California's immigration legislation in the early 1990s. In this study, Santa Ana demonstrates the power of the press to perpetuate racist ideologies of Latinos in the United States by discursively erasing (cf. Irvine & Gal, 2000) important intra-group distinctions and dehumanizing an already racially and socioeconomically marginalized group of people. This issue is of course not particular to the United States: critical studies of language use depicting immigrants and asylum seekers have also covered other national and transnational contexts (e.g., Baker & McEnery, 2005; El Refaie, 2000).

The controversial topic of what terminology should be used by journalists in making reference to immigrant groups surfaced in the American press when the National Association of Hispanic Journalists (NAHJ) launched a campaign to change the recommended terminology to refer to immigrants residing in the United States without authorization in *The Associated Press Stylebook*, the journalism industry's standard language guide. The NAHJ expressed opposition to the terms "illegal immigrants" and "illegal aliens," which they deem to be politically charged and dehumanizing to immigrants. (See also the internet campaign "Drop the i-word": http://colorlines.com/droptheiword/.) The NAHJ sought to replace these terms with "undocumented workers" (Carmichael & Burks, 2010), which they view as both politically neutral and referentially accurate. This debate is ongoing; at the time of writing this chapter, *The Associated Press Stylebook* recommends avoiding "illegal" as a noun but recommends its use in adjectival form over "undocumented worker" (Associated Press, 2011, p. 137).

Linguists and communication scientists (myself included) have been called upon to weigh in on this issue and have provided cognitive linguistic and sociolinguistic perspectives on the importance of the choice of referring terms journalists use in the attempt to report objectively on such divisive issues. Newspaper articles and columns have even addressed the controversy over the *Associated Press Stylebook* guidelines, sharing linguists' perspectives on how terminology such as "illegal" (especially as a noun, but also in adjectival form) are deemed offensive by the individuals to which they refer and how they discursively frame controversial policies in a way that favors a particular (anti-immigrant, not just anti-*illegal* immigrant) point of view (Carmichael & Burks, 2010; McIntosh, 2011).

However, it should be noted that despite *The Associated Press Stylebook*'s current recommendation against the use of the word "illegal" as a noun, the trend has not necessarily caught on among wider audiences, especially outside of newspaper journalistic practice. For instance, in a 2011 nationally televised US Republican primary debate (CNN, 2011), reference to "illegals" (as a noun) surfaced a dozen times in the span of a few minutes during a heated discussion on the topic between candidates Rick Perry and Mitt Romney. Carmichael and Burks (2010) also point out that the use of the term "illegal alien" (which *The*

Associated Press Stylebook recommends against) is on the upswing, appearing nearly four times as frequently in US newspapers in 2010 as it did in 2000. Clearly, the negotiation of appropriate language use on the topic of immigration in the United States is still ongoing, is being fueled by a diverse range of political ideologies, and has not yet given way to a prevailing trend in the press. It is also evident, however, that official journalistic perspectives have not been picked up more widely, especially where it counts: by potential future national and international leaders.

I have outlined two cases in which language-related topics that have been of interest to sociolinguists for quite some time were catapulted into the press, and I have touched upon the engagement of scholars with journalists to share linguists' perspectives on linguistic diversity and the ideological implications of language with the broader public. The vignettes that follow this chapter provide three further examples of the nexus between sociolinguistics and the media, one in a US context (Kiesling, Vignette 18a), one in the United Kingdom (Upton, Vignette 18b), and one in China (Wong, Vignette 18c). The issues highlighted in this chapter, namely of how to successfully frame complex issues for non-specialist audiences, how to deal with frustration over the media "getting it wrong" or "missing the point," and what to do when wider audiences remain ignorant of linguists' perspectives, are also reiterated in Kiesling, Upton, and Wong's reflections on their own studies of and interactions with the media.

For this reason, it is of ultimate importance for linguists to continue to weigh in on current language debates in any way possible, be it through direct contact with the press, our own media channels (e.g., personal and professional websites, blogs, and social networking channels), and, perhaps most importantly, in our daily non-mediated interaction with students. Just as with our interactions with the media, it is important to frame sociolinguistic research appropriately for the classroom context. In fact, my work dealing with the representation of Ebonics in the press (Sclafani, 2008) was inspired by interactions with my college students, who would respond positively when I introduced the topic of "African American English" in a sociolinguistics class but would practically shudder when I used the word "Ebonics." In the next section, I introduce some ways in which linguists have used the media for educational efforts.

Media, the Community, and the Classroom

In addition to considering the media as an object of study, researchers have made strategic use of the media to share their work with students and the public, as well as to give back to the communities that have served them in their research endeavors. The most extensive and longest-running research project working toward disseminating academic research to the public and serving the local community is the North Carolina Language and Life Project (NCLLP, http://www.ncsu.edu/linguistics/ncllp/), spearheaded by Walt Wolfram at North Carolina State University. The NCLLP has extensively documented linguistic variation in North Carolina over the past two decades and has created a number of educational books and videos for the general public that celebrate the diverse cultural heritage of the state, including general-interest books and documentaries on the history and local brogue of Ocracoke Island (Wolfram & Hutcheson, 2008; Wolfram & Schilling-Estes, 1997), the African American communities of North Carolina (Wolfram, Rowe, & Grimes, 2006), the Lumbee Native American community (Wolfram, Dannenberg, Knick, & Oxendine, 2002), Appalachian cultural heritage (Wolfram & Hutcheson, 2006), and, most recently, the Spanish-speaking population of the state (Wolfram, Cullinan, & Hutcheson, 2011).

In addition to these materials, Wolfram and his colleagues have made extensive efforts over the years to disseminate information on linguistic variation both in person and through a variety of media outlets, for example at museums, state fairs, and in the state public educational system. Keeping in mind the importance of educating youth on issues of dialect diversity and language-based discrimination, they have developed the Dialect Awareness Curriculum (DAC) (Reaser & Wolfram, 2007), which teaches middle school educators and students about sociolinguistic issues of local interest and includes topics such as language attitudes; dialect patterns; regional, social, and ethnic variation; style shifting; and the history of language.

Recognizing the difficulty of bringing "extra" material into public school classrooms, which are tightly constrained by state regulations, Reaser and Wolfram designed the DAC to better meet the standard course of study for the state's eighth-grade social studies curriculum, and they attained the endorsement of the program by the North Carolina Department of Public Instruction. They have also made all DAC materials, including texts, videos, sound files, and interactive games, available free of charge (www.ncsu.edu/linguistics/dialectcurriculum.php).

Despite the plethora of books, DVDs, and sound recordings produced by the North Carolina Language and Life Project, the transferability of the materials to curricula outside of North Carolina is somewhat limited, owing to the strong local emphasis of the work. This is not to say the materials cannot be used in other geographical and social contexts; I have successfully used these materials with my students in Athens, Greece, as a case study of dialect diversity in the United States, for example. Other examples of sociolinguistic media endeavors that appeal to wider audiences include recent documentary films such as *Do You Speak American?* (*DYSA*) (Cran, Buchanan, & Anthony, 2005) and *The Linguists* (Kramer, Miller, & Newberger, 2009), which involve linguists describing and demonstrating the work they do to understand and document regional and social varieties of American English and endangered languages around the world, respectively. Both of these documentaries include additional resources for individual and classroom use (www.pbs.org/speak/; www.thelinguists.com). For example, *DYSA* has several interactive web-based games in which students can test their knowledge of American English dialects and find links to additional resources on topics such as language and ethnic identity, teaching Standard English to dialect minorities, literature and voice, and standard and official language movements.

Alongside these recent documentaries, two films of sociolinguistic interest that are now dated but still relevant and entertaining are *American Tongues* (Kolker & Alvarez, 1987) and the BBC-produced *Crosstalk* (Gumperz, Jupp, & Roberts, 1979). *American Tongues*, like *DYSA*, focuses on regional dialects of American English, and *Crosstalk* features Gumperz's interactional sociolinguistic analysis of the conversational basis of interethnic miscommunication and stereotyping of ethnic minorities in Great Britain. Unfortunately, these older documentaries are now difficult to find and do not capture the shifting demographics of the past couple of decades in the United States and United Kingdom. This is one area in which sociolinguists can contribute new work toward media endeavors.

It is also worth noting that *Crosstalk* is one of the few sociolinguistic documentaries that deal explicitly with discourse-level differences and resulting cross-cultural miscommunication. While regional phonological and lexical particularities, which are central to the other documentaries mentioned thus far, are often perceived as endearing to laypeople when performed for demonstrative purposes, features such as the distinct intonational contours of Indian and Pakistani speakers of English emphasized by Gumperz in *Crosstalk* are not often perceived in the same way as "quaint" local vocabulary and regional accents. Neither are gendered differences in conversational styles, which have been highlighted in documentaries featuring Deborah Tannen's work (Tannen, 1995; DiNozzi & Tannen, 2001). As both Gumperz and Tannen point out, it is at this level of language structure, where differences are less noticeable to an untrained ear and more likely to be perceived as related to intellect or character rather than to language, that differences can make or break relationships and have the potential to result in systemic discrimination against certain (oftentimes already marginalized) groups.

Looking Forward

How might we move forward in both the sociolinguistic study of the media and in sharing the findings of sociolinguistic research through a variety of media channels? Considering the fact that the media constitute a key site in the construction and reproduction of language ideologies, linguists should continue to critically investigate new forms and uses of language in these contexts, in both realistic (news) genres and fictional ones (television programming, film). As mediated channels of communication expand and evolve alongside technological innovation, linguists must continue to ask questions about how these channels act to project and mute particular voices, registers, and varieties of language. The variety of media channels categorized under the umbrellas of blogs and social networking services that have appeared in the past few years have undoubtedly caused some shift in the ideological potential of traditional news media outlets, or at least have created new discursive space in which dominant ideologies of language can be contested and stances can be posited in multimodal formats. Indeed, new research is currently investigating the language structures, social functions, and ideological components of these new discursive spaces (see, for example, Thurlow & Mroczek, 2011).

On a similar note, let this discussion also be a note of encouragement for linguists to continue to weigh in on the language debates of our times through a variety of media outlets. I have discussed only two cases here, but there are many other current debates that critically hinge on language, either directly, as in the two cases I have presented, or indirectly through the inherently political nature of the language of representation. Sharing the findings of sociolinguistic work with the public, especially work related to language variation at all levels, from phonological to discourse-level phenomena, should be a priority in scholarly work. In addition to communicating with the media and introducing media in the classroom, the creation of more documentary films is an excellent way to share findings and need not be expensive or time-consuming. Many universities have professional-quality equipment available for free rental when used for academic purposes, and in my own teaching experiences I have found that students delight in putting their extracurricular interests and skills in design, shooting, and editing to work for assignments in introductory linguistics courses. With free video-sharing websites and the strategic placement of key words, the fruits of sociolinguistic research have the potential to be reached by a global audience.

Over the past few decades, sociolinguists have gotten off the media sidelines and have made extensive efforts to share the findings of research with the public through a variety of channels and modes: spoken and written, and in live and mediated contexts. These successful efforts have resulted in the proliferation of resources that are easily accessible – and in some cases, highly visible – to non-academic audiences. Sociolinguists must continue to weigh in on language debates in the media, even if it means taking on traditionally non-scholarly (i.e., emotionally charged) stances. As a result, the findings of empirical research on language can be understood, appreciated, and put to use in a variety of institutional contexts, such as business, law, and education. It is only when understandings of language-based variation, evaluation, and discrimination are made relevant to the daily lives of everyday citizens that we can use language to achieve the goal of social change.

References

Associated Press. (2011). *The Associated Press stylebook and briefing on media law 2011.* New York: Basic Books.

Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics, 4*, 197–226.

Baron, N. S. (2008). *Always on: Language in an online and mobile world*. New York: Oxford University Press.

Baugh, J. (2000). *Beyond Ebonics: Linguistic pride and racial prejudice*. New York: Oxford University Press.

Bell, A. (1984). Language style as audience design. *Language in Society, 13*, 135–204.

Bell, A., & Garrett, P. (Eds.). (1998). *Approaches to media discourse*. Oxford: Blackwell.

Carmichael, K., & Burks, R. A. (December 13, 2010). Finding the right language: What should journalists call immigrants in the U.S. without papers? *Maryland Newsline*. Retrieved from http://www.newsline.umd.edu/politics/specialreports/immigration/illegal-immigration-terminology-120910.htm

CNN. (2011, October 18). Western Republican presidential debate [Television broadcast]. A. Cooper (Moderator). M. Bachmann, H. Cain, N. Gingrich, R. Paul, R. Perry, M. Romney, & R. Santorum (Participants). Las Vegas, NV.

Cran, W., Buchanan, C., & Anthony, S. (Producers), & Cran, W. (Director). (2005). *Do you speak American?* [Television series]. Princeton, NJ: Films for the Humanities & Science.

Crystal, D. (2006). *Language and the internet*. Cambridge: Cambridge University Press.

Dillard, J. L. (1973). *Black English*. New York: Vintage.

DiNozzi, R. (Producer/Director), & Tannen, D. (Writer) (2001). *He said, she said: Gender, language and communication* [Video recording]. Los Angeles: Into the Classroom Media.

El Refaie, E. (2000). Metaphors we discriminate by: Naturalized themes in Austrian newspaper articles about asylum seekers. *Journal of Sociolinguistics, 5*, 352–371.

Fairclough, N. (1995). *Media discourse*. London: Edward Arnold.

Gee, J. P. (1990). *Social linguistics and literacies: Ideologies in discourses.* London: Falmer Press.

Gumperz, J. J., Jupp, T. C., & Roberts, C. (1979). *Crosstalk: A study of cross-cultural communication* [Video recording]. London: National Centre for Industrial Language Training in association with the BBC.

Heritage, J. (1985). Analyzing news interviews: Aspects of the production of talk for an "overhearing" audience. In T. van Dijk (Ed.), *Handbook of discourse analysis*, Vol. 3: *Discourse and dialogue* (pp. 95–119). London: Academic Press.

Irvine, J., & Gal, S. (2000). Language ideology and linguistic differentiation. In P. V. Kroskrity (Ed.), *Regimes of language: Ideologies, politics, and identities* (pp. 35–84). Santa Fe, NM: School of American Research Press.

Kolker, A., & Alvarez, L. (Producers/Directors). (1987). *American tongues* [Motion picture]. USA: Center for New American Media.

Kramer, S., Miller, D., & Newberger, J. (Producers/Directors). (2009). *The linguists* [Motion picture]. USA: Ironbound Films.

Labov, W. (1972a). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1972b). *Language in the inner city: Studies in the black English vernacular.* Philadelphia: University of Pennsylvania Press.

Laforest, M. (1999). Can a sociolinguist venture outside the university? *Journal of Sociolinguistics, 3*, 276–282.

McIntosh, S. (2011, November 19). Words matter on immigration. *Atlanta Journal-Constitution*. Retrieved from http://www.ajc.com/opinion/words-matter-on-immigration-1233419.html

Pandey, A. (2000). Linguistic power in virtual communities: The Ebonics debate on the internet. *World Englishes, 19*, 21–38.

Perry, T., & Delpit, L. (1998). *The real Ebonics debate: Power, language, and the education of African American children*. Boston: Beacon Press.

Reaser, J., & Wolfram, W. (2007). *Dialect awareness curriculum*. Raleigh: North Carolina Language and Life Project. Retrieved from http://www.ncsu.edu/linguistics/dialectcurriculum.php

Rickford, J. R. (1999). The Ebonics controversy in my backyard: A sociolinguist's experiences and reflections. *Journal of Sociolinguistics, 3*, 267–275.

Ronkin, M., & Karn, H. E. (1999). Mock Ebonics: Linguistic racism in parodies of Ebonics on the Internet. *Journal of Sociolinguistics, 3*, 360–380.

Santa Ana, O. (2002). *Brown tide rising: Metaphors of Latinos in contemporary American public discourse*. Austin: University of Texas Press.

Sclafani, J. (2008). On the intertextual origins of public opinion: Constructing Ebonics in the *New York Times*. *Discourse and Society, 19*, 507–527.

Smitherman, G. (1977). *Talkin and testifyin: The language of Black America*. Boston: Houghton Mifflin.

Tannen, D. (1995). *Talking 9 to 5: Women and men in the workplace* [Video recording]. Burnsville, MN: Charthouse International Learning Corporation.

Thurlow, C., & Mroczek, K. (2011). *Digital discourse: Language in the new media*. New York: Oxford University Press.

Van Dijk, T. A. (1988). *News as discourse.* Hillsdale, NJ: Lawrence Erlbaum.

Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Washington, DC: Center for Applied Linguistics.

Wolfram, W. (Producer), & Cullinan, D., & Hutcheson, N. (Directors) (2011). *Spanish voices* [Video recording]. Raleigh: North Carolina Language and Life Project.

Wolfram, W., Dannenberg, C., Knick, S., & Oxendine, L. E. (2002). *Fine in the world: Lumbee language in time and place*. Raleigh: North Carolina State University.

Wolfram, W., & Hutcheson, N. (Producers), & Hutcheson, N. (Director). (2006). *The Queen family: Appalachian tradition and back porch music* [Video recording]. Raleigh: North Carolina Language and Life Project.

Wolfram, W., & Hutcheson, N. (Producers), & Hutcheson, N. (Director). (2008). *The Carolina brogue* [Video recording]. Raleigh: North Carolina Language and Life Project.

Wolfram, W. (Producer), & Rowe, R., & Grimes, D. (Directors).(2006). *This side of the river: Self-determination and survival in the oldest Black town in America* [Video recording]. Raleigh: North Carolina Language and Life Project.

Wolfram, W., & Schilling-Estes, N. (1997). *Hoi toide on the Outer Banks: The story of the Ocracoke brogue*. Chapel Hill: University of North Carolina Press.

Wolfram, W., & Schilling-Estes, N. (2006). *American English: Dialects and variation* (2nd ed.). Malden, MA: Blackwell.

Vignette 18a
Media Interest in Sociolinguistic Endeavors

Scott F. Kiesling

In 2004, I published an article about the word *dude* (Kiesling, 2004) and I was introduced to some of the ways the news media become interested in and talk about sociolinguistic endeavors. Mike Crissey (2004) of the Associated Press wrote an article on my study that was reprinted around the world, and in short order my voicemail and email were full. That was my most intense and widely publicized interaction with the world's media beast, but I've also been involved in talking to the news media because of my work with Barbara Johnstone on how Pittsburghers talk about "Pittsburghese." In fact, work on that project led to my first television and radio interviews. I've also had a number of interactions with reporters on other linguistic topics, including most recently the use of text initializations such as "OMG" and "WTF" in spoken conversation. In this vignette, I offer a few observations based on my experience and what I've noticed about other media coverage of things (socio)linguistic.

Feed the Beast

Crissey did not find out about my article by reading *American Speech*. He was told about it by Patricia White, one of the media relations people at my university, who in turn heard about the topic from me. She had contacted me after some publicity surrounding the Pittsburgh dialect two years earlier, and when the *dude* article was published I wrote and told her about that one too, since I knew that it was a term that generates interest (in fact, lots of the data in the article are from class projects over the years, projects that were successful partially because the topic held students' interest). My lesson from this experience was that journalists rarely find things themselves by reading academic journals (although we do have some linguist writers and journalists).

In today's world of Twitter, blogs, and Facebook, the process of person-to-person contact may seem old-fashioned. But the "traditional" news media are still a powerful tool for communicating ideas, especially ideas about linguistic discrimination. Most universities have communications offices or media relations offices to help. I recommend finding this office, getting to know one person in the office, and explaining our field and your interests. When you have something that might be of interest, let them know, and if nothing else, they might just send it to a reporter as a story idea. You can even do this at the beginning of a project, as we did in Pittsburgh, and the ensuing publicity might even help with doing the research. In our Pittsburgh project, publicity has helped us understand more fully the importance of the Pittsburgh dialect to Pittsburghers and how they think about it, and it has also led to more opportunities for archiving. If you aren't at a university, find a local reporter who is interested in language topics and send them story ideas and comments on stories they publish.

It's Rarely Considered Serious Business

Most people don't think language is as serious a business as we linguists do. Of course, that's one big reason why we are linguists and they aren't. Often, stories on sociolinguistics will not be relayed as super-serious "science," or even "social science," pieces but as more of a fluffy feature piece; Msnbc.com has the *dude* piece in the "Weird News" section. I think it's important not to worry about these types of editorial decisions if you get involved with the media. Readers and listeners and viewers are looking to be entertained, even by the most serious organizations, such as National Public Radio. For most people, any language story will be a diversion, a curiosity to be looked at for a short while. These stories aren't out there to provide a full education to news media consumers. At the same time, you never know whether some high school student who likes language might just find out that the field of linguistics exists, and a linguist is born. And a story that makes people think about their language, even for a short while, might pave the way for further listening down the road. The more people hear stories of how everyone has a "funny accent," the more they might realize that they are normal, eventually leading to more tolerance of such "funny accents." One story is unlikely to do the trick, but the more times each person hears that language is a social object and not a window into a speaker's inherent intelligence, the more we make progress.

If you get involved with the media, especially if you end up doing a radio or TV interview, keep in mind that you are playing a role (remember to read your Goffman – e.g., Goffman, 1959 – for more on role and performance). Yes, you are ostensibly being featured in the media for your expertise, but you will be appearing in a particular format in which the interviewer determines the questions and agenda, and you have to fit into that. Try not to let it bother you, go with the flow, and be yourself.

Short and Sweet

You might talk to a reporter (especially a good one) for over an hour but then only have 10 seconds or one sentence reported in the resulting piece. I have experienced this situation at least three times. In one case, the story didn't even run! For those of us used to reading and writing longer articles (let alone books), a short media piece may seem incredibly superficial, but that's how it works, and that's all the space or time you get. My advice again is to go along with it, but try to think of fairly short ways of expressing your main points, especially if they might differ from the agenda. Be able to articulate your research in several versions: in one short sentence, in 20 seconds, and in around a minute. One minute is a really long time on radio and television. Also, have similar set pieces for why your research is important. Practice these in advance, and then you will at least not be caught speechless.

Reporters vary in their expertise, interests, and angles. If the interview is for a written piece, you might have to lay out some fairly basic sociolinguistic assumptions. Reporters will usually be patient with this instruction. Such background is unlikely to be quoted, but it is important for a reporter to know it in order to accurately present the material they write.

Interest Is Driven by Language Ideology, but You Can Make a Difference

As we know from lots of work in sociolinguistics and linguistic anthropology, ideologies about language are ones that the public rarely confronts. Because these ideologies are so naturalized, for the most part they are not even experienced as ideologies. Once the topic gets in the media, though, we have a brief chance, maybe only a sentence, in which to bend such ideologies even just a little bit, and maybe only for some people. But it is certainly worth the try. Here are a few points one might want to make in any media piece about sociolinguistics:

- There is no objectively good language or bad language. I think the best way to explain this position is to point out that language tells us about people's identities, so good and bad language can be read as (who we think are) good and bad people. This argument reverts a little bit to the idea that language is a reflection of the social world (rather than constitutive of it), but it's worth putting this way, I think.
- Language is not a static object. We need to point out that language changes and that a previous generation would almost inevitably be horrified by the language spoken by the current young people. It helps to note that even the Queen of England has changed the way she speaks.
- Language is "designed" for social information and not for giving instructions. Even linguists often conceive of language as a tool for unproblematically moving an idea in one person's mind to another person's mind. It is easy to point out that this process is rarely unproblematic (e.g., "just think of the last time you misunderstood someone"), but on the social front humans are quite sensitive to language differences. I also like to point out Robin Dunbar's argument that language evolved for gossip.

Reporters sometimes like to play language games that reflect particular language ideologies as well. The most common relate to defining and translating. In my case, they played different uses of *dude* and asked me to translate what each particular use was. This game is of course one that doesn't really take my point, but it was worth it to go along with the game and provide odd "translations." This sort of language game addresses all three ideologies: that there is some "correct" definition or definitions, that these meanings won't change, and that a word carries some denotational meaning, rather than indicating a stance or attitude that is contingent on the situation.

Prepare for Negativity from Media Consumers, but Have Fun

As we all know, language can incite serious passion among people, so prepare for some negative reactions, especially from those well steeped in prescriptive ideology. (Some of the comments on Crissey's story were pretty harsh, if predictable: "Duh. I could tell you that.") There are also going to be people who think sociolinguists are wasting our time (and their money, even if your work is like my *dude* paper, which wasn't supported by a government grant and didn't cost anything except my time). Most such reactions will not be mollified by rational discourse, but you never know. Just try not to take it personally; the best way to deal with it is humor, something that reporters appreciate as well. Given this view, I have tried to keep interviews fun and not too serious. The exception is when journalists actually ask why this research is important, to which you should have a good answer – maybe something like, "Because language tells us more about what it means to be human."

References

Crissey, M. (2004, December 10). So what's up with that "dude"? *Los Angeles Times*. Retrieved from http://articles.latimes.com/2004/dec/10/entertainment/et-dude10

Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NY: Anchor Doubleday.

Kiesling, S. F. (2004). Dude. *American Speech, 79*(3), 281–305.

Vignette 18b
Sociolinguistics on BBC Radio

Clive Upton

Although earlier dialectologists put a lot of effort into collecting information on regional vocabulary, since the 1960s "social" dialectologists have largely concentrated on phonology and grammar, at the expense of lexis. Anyone who doubts this imbalance of targeting need only consult the sociolinguistic literature: Milroy and Gordon (2003), for example, understandably and not untypically devote chapters to phonology and grammar and discourse but none to lexis. And it is not hard to see why lexis is comparatively little studied by variationists. While stimulating a flow of informal speech results in the gathering of pronunciations and grammatical structures for comparison across a range of speakers, only in carefully constructed, context-dependent speech will there be a likelihood of repeated lexemes or comparable variants occurring. Also, it has not generally proved possible to gather enough lexis to permit that quantification which objectifies observations on social variation. These issues were to the fore in the formulation of a fieldwork method for the British Broadcasting Corporation's *Voices* project, 2004–2007.

BBC *Voices* (Elmes, 2005) grew from a plan of the Corporation to use its Nations and Regions journalistic resources to survey vernacular speech across the United Kingdom. In 2004, this plan crystallized as a project involving some 60 broadcast journalists interviewing groups of speakers across the country, using a "sense relation network" (SRN) interview technique pioneered at the Universities of Leeds and Sheffield (Kerswill, Llamas, & Upton, 1999; Upton & Llamas, 1999). The BBC's tool used to prompt discussion (the so-called Spidergram), a simplification of the SRN device, asked interviewees to give local and personal words for 38 different everyday concepts. It also formed the basis for a website, www.bbc.co.uk/voices (BBC, 2011), where the public were invited to offer their words and to enter into online discussion on matters of language use. The collection technique of *Voices* focused deliberately on lexis as a linguistic area in need of data and likely to stimulate debate. The technique, unsurprisingly, excited plenty of popular discussion while maintaining a determinedly linguistic focus.

Sound recordings and web-based input resulting from *Voices* inquiries provided material for much BBC radio and television broadcasting in 2005, with the website remaining "live" for more input until 2007. Subsequently it has led to academic analysis chiefly in two projects, *Voices of the UK* at the British

Library in London, describing more than 700 hours of sound recordings generated from journalistic fieldwork, and *Whose Voices?* at Leeds, concentrating on lexical analysis and on language-ideological issues emerging from the *Voices* exercise (Upton & Davies, forthcoming). Showcased here is just a fraction of the lexical variation uncovered for one of the variables investigated by *Whose Voices?*, which demonstrates some of the insights into lexical variation that come from the accumulation of media-driven data in the mass. It will be appreciated that, at 38, the number of variables is deliberately kept small so that the quantifiable element is maximized: beyond stimulating natural speech (and hence phonology and grammar), the aim of *Voices* was to *begin* to capture sociolinguistic data of this kind to see what insights might emerge from large data.

Simple un-lemmatized output relating to a variable, as extracted from the BBC material by *Whose Voices?*, and accessible via Excel pivot tables, is illustrated by Table 18b.1 (which relates to the (PLAY TRUANT) variable).

The raw scores show that women respond to this prompt more readily than do men; they generally predominate in the dataset anyway (itself a matter of some importance), but the proportion here is striking because it suggests greater concern with truancy. Adolescents and young adults (groups 2 and 3, 16- to 25-year-olds), for whom schooling is current or a recent memory, score high, as do young adults aged 26 to 35 (groups 4 and 5), some of whom might be expected to have children themselves. Unsurprisingly, scores tail off from here. Conversion of raw scores to percentage-of-column figures as in Table 18b.2 raises a less obvious issue.

Table 18b.1 (PLAY TRUANT), All Variants, Ages <16 to 85

word2	*(All)*		
Count of word	*Gender*		
Age	*Female*	*Male*	*Grand Total*
1	691	421	1,112
2	3,064	1,792	4,856
3	2,315	1,566	3,881
4	2,357	1,680	4,037
5	1,896	1,624	3,520
6	1,359	1,383	2,742
7	1,065	1,001	2,066
8	959	760	1,719
9	695	622	1,317
10	462	452	914
11	134	182	316
12	73	74	147
13	32	31	63
14	17	13	30
15	7	11	18
Grand total	15,126	11,612	26,738

Table 18b.2 (PLAY TRUANT), All Variants by Age, Gender as Percentage of Column

Word	*(All)*		
Count of word2	*Gender*		
Age	*Female*	*Male*	*Grand total*
1	4.57%	3.63%	4.16%
2	20.26%	15.43%	18.16%
3	15.39%	13.49%	14.51%
4	15.58%	14.47%	15.10%
5	12.53%	13.99%	13.16%
6	8.98%	11.91%	10.26%
7	7.04%	8.62%	7.73%
8	6.34%	6.54%	6.43%
9	4.59%	5.36%	4.93%
10	3.05%	3.89%	3.42%
11	0.89%	1.57%	1.18%
12	0.48%	0.64%	0.55%
13	0.21%	0.27%	0.24%
14	0.11%	0.11%	0.11%
15	0.05%	0.09%	0.07%
Grand total	100.00%	100.00%	100.00%

Here we see, unsurprisingly, that 20% of female answers and 15% of male answers are volunteered by 16- to 20-year-olds (group 2). But while there is a general decline in the proportion of respondents answering on truancy thereafter, from ages 31 to 35 (group 5) onward, men respond in larger proportion than women. It is not necessary to speculate on the reason here; rather, the point is that data in large amounts, readily able to be processed, prompt questions that would otherwise go unasked.

We can, of course, drill down more finely into such data, to the level of grouped or single lexical items. Figure 18b.1 shows details of support in the data for the 10 most frequently offered (PLAY TRUANT) items, following lemmatization. It shows a pattern typical across all the lexical variables studied, in which one very strongly supported variant – in this case, 'skive (off)' – dominates, with a rapid falling away beyond. Such quantification addresses speakers' colloquial and non-standard lexicon: 'skive (off)' is widely supported throughout the population, giving rise to its high numerical scoring, while with the likely exceptions of the dated phrase 'play hookey' and the semantically broad 'skip,' the other variants are non-standard regional. We have a population that, regardless of age or gender, has access to major colloquial variants, which some use exclusively and others use in conjunction with more localized forms. The BBC-collected data allow for this finding to be explored in detail.

Going still more finely into (PLAY TRUANT) data, it is possible to isolate a variant such as 'mitch (off),' again allowing us to perform exercises relating to usage by age and gender. As variants are picked out, it also becomes increasingly rewarding to investigate them by postcode-related geographical distribution.

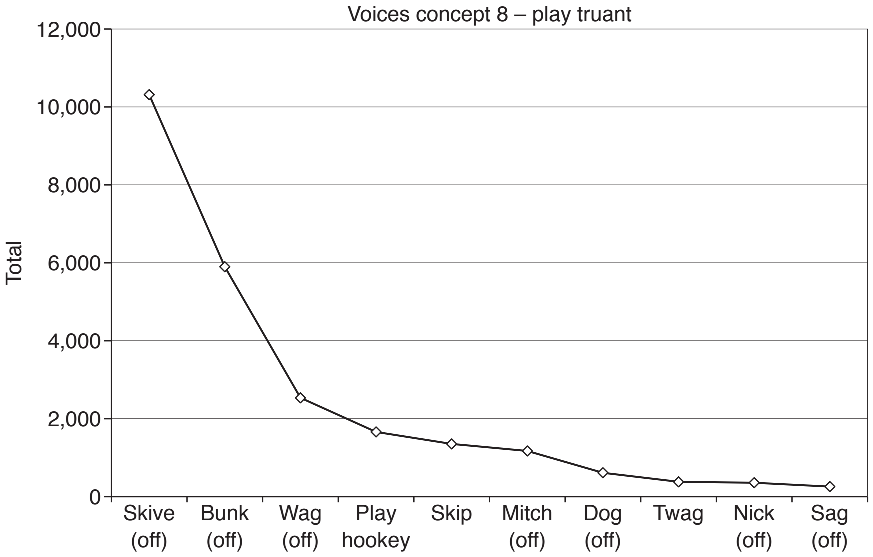


Figure 18b.1 (PLAY TRUANT) Top Ten Variants.

Table 18b.3 details those areas where 'mitch (off)' is found 10 or more times in the database. It can be seen that heaviest use of 'mitch (off)' is in Wales, the west of England, and Northern Ireland, with support especially coming from speakers in Cardiff, Belfast, and Swansea postcode areas.

The choropleth map of Figure 18b.2 indicates with darker shading that speakers in southwest Wales, the English West Country, and Northern Ireland are

Table 18b.3 (PLAY TRUANT), Variant 'Mitch (Off),' Distribution by Postcode Area, $n = >10$

Word	*(Multiple item)*
Count of word2 by area	*Total*
Bristol	17
Belfast	194
Cardiff	206
Exeter	38
Llandudno	10
Manchester	12
Newport	19
Plymouth	30
Swansea	152
London SE	19
London SW	19
Torquay	20
Grand total	736

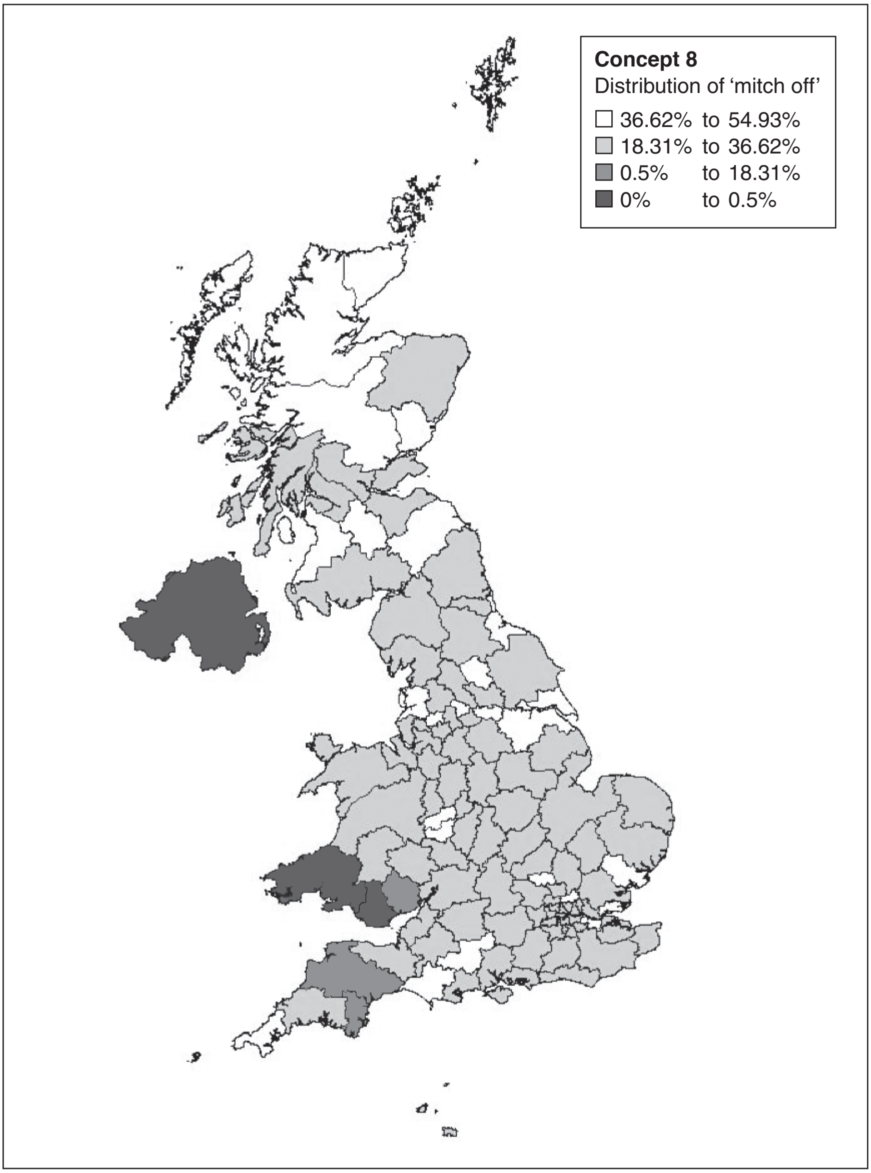


Figure 18b.2 (PLAY TRUANT), Variant 'Mitch (Off),' by Postcode Area (Map by Ann Thompson).

decidedly linked in their usage. This finding is no accident, but rather speaks to social connections – and disconnections – of considerable antiquity. Norman south Wales was home to colonists who, with a hostile Welsh-speaking hinterland, looked southward across the Bristol Channel for their identity and material support. And it was from southwest Wales that the Normans launched their invasion of Ireland and maintained their Irish colony. Such are these bonds, and so much easier have communications traditionally been across water than through the mountainous Welsh interior, that an England-focused orientation remains today, controversial, as it was historically.

On the *Voices* website, there is a large and lively discussion board on the language situation of Wales, which provides insights into both the language and language ideologies of the Principality. There, online contributor "Jeff of Abergavenny" writes of Welsh that a "kind of street language, disliked by purists, is used by … youngsters, absorbing many English words and influenced by popular culture. At the same time they are capable of using 'standard' Welsh. Surely evidence of a living language?" Bringing together this ideological remark with *Voices* (PLAY TRUANT) submissions, we find Welsh-influenced variants of English 'mitch (off)' are 'mitcho' in Swansea and 'mitsio' in Llandudno and Shrewsbury (the latter is in England but close to the English–Welsh border). The relation between English and Welsh is, especially for many Welsh speakers, a highly charged identity-bound area of debate, here extending into the non-standard.

It is understandable why, after an initial blossoming of interest in dialect words, lexicographical dialectology took something of a back seat once sociolinguistically oriented studies got under way from the middle of the 20th century onward. Large datasets are needed for quantification to be employed to provoke sociolinguistic questions and to allow appropriate analysis, and it is not immediately apparent how large-scale comparable records of lexical data might be assembled. BBC *Voices*, however, goes some useful way toward filling a gap in the modern dialectological record.

References

BBC. (2011). *Voices*. Retrieved from http://www.bbc.co.uk/voices/

Elmes, S. (2005). *Talking for Britain: A journey through the nation's dialects*. London: Penguin.

Kerswill, P., Llamas, C., & Upton, C. (1999). The first SuRE moves: Early steps towards a large dialect database. In C. Upton & K. Wales (Eds.), *Dialectal Variation in English: Proceedings of the Harold Orton Centenary Conference 1998* [Special edition]. *Leeds Studies in English, 30*, 257–269.

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

Upton, C., & Llamas, C. (1999). Two large-scale and long-term language variation surveys: A retrospective and a plan. *Cuadernos de Filología Inglesa*, *Variation and Linguistic Change in English: Diachronic and Synchronic Studies, 8*(1), 291–304.

Upton, C., & Davies B., (Eds.). (Forthcoming). *Analysing twenty-first century British English*. London: Routledge.

Vignette 18c
Media, Politics, and Semantic Change

Andrew D. Wong

I became fascinated with the semantic change of the Chinese label *tongzhi* because it captured my interest in both sociolinguistic variation and the study of language and sexuality. *Tongzhi* (often glossed as "comrade") was first adopted by Nationalist revolutionaries in Republican China at the beginning of the 20th century. During the Communist Revolution (1921–1949), it acquired stronger political connotations, and its use as an address term among revolutionaries became more popular. The reciprocal use of the term indexed solidarity, equality, respect, and intimacy. After the founding of the People's Republic of China in 1949, the Communist Party made great efforts to promote the use of *tongzhi* as a general address term among the masses. Since the opening up of the market economy of China in 1978, it has become disfavored because of its political connotations. In the late 1980s, it was appropriated by gay rights activists in Hong Kong as a term of reference for those of non-normative sexual orientations (i.e., lesbians, gay men, bisexuals, and transgender people).

I first became aware of the use of *tongzhi* to refer to members of sexual minorities in 1989, when the label made its debut in the Chinese title of the "First Hong Kong Gay and Lesbian Film Festival" (*Heung-Gong tongzhi din-ying gwai*). Like many linguistic changes, the semantic change of *tongzhi* was started on a whim. Probably, few people expected that it would catch on. Ten years later, I was surprised to learn that this semantic innovation had spread from Hong Kong to Taiwan. I noticed its prevalent use in *G&L*, a now-defunct magazine published in Taiwan that catered to lesbians and gay men in Taiwan, Hong Kong, and overseas Chinese communities. With Qing Zhang, I examined the use of *tongzhi* in *G&L* and found that the magazine used the term, together with a host of lexical and discourse features, to build an imagined Chinese gay community and to underscore the cultural distinctiveness of same-sex desire in Chinese societies (Wong & Zhang, 2001).

Encouraged by what we found, I decided to pursue this topic further and return to the place of origin of this semantic innovation. *G&L* was, after all, a niche magazine published in Taiwan. To what extent had this semantic innovation been adopted by mass-circulation print media in Hong Kong? The ongoing semantic change of *tongzhi* from "comrade" to "sexual minorities" offered me a unique opportunity to study semantic variation and change, as well as the role of sexuality in sociolinguistic variation. At that time, neither topic had received much attention in variationist sociolinguistics. Reflecting on my experience in conducting this research, I discuss in this vignette (1) the methodological and analytical challenges that I encountered; (2) how I coped with them; (3) the limitations of using print media data to study sociolinguistic variation and change; and (4) how I tried to understand the implications of my work beyond variationist sociolinguistics.

Methodological and Analytical Challenges

The most significant challenge that I encountered was the lack of models and precedents that I could draw upon for inspiration. Over the years, variationists have developed sophisticated models and a standard methodology for data collection and analysis; however, these methods are primarily designed for studying phonological and morphosyntactic variation. To develop my research project, I adopted an integrated quantitative and qualitative approach, using insights from several sources in addition to variationist sociolinguistics – for example, Jane Hill's (1993) research on appropriation, Sally McConnell-Ginet's (2001) work on meaning contestation, and a sizable literature on media discourse in critical discourse analysis (e.g., Fairclough, 1992).

First, I needed to create a corpus of newspaper articles that would be big enough for identifying patterns of label use but also small enough to allow for a close examination of label use in discourse. It is easy to feel overwhelmed with the amount of data that print media offer. Nowadays, many newspapers maintain easily searchable online archives. I found it important to develop a corpus specifically for my project, as well as clear criteria for inclusion and exclusion. I decided to focus on *Oriental Daily News* (*ODN*) because it was, at that time, the most widely circulated newspaper in Hong Kong. I used *tongzhi* and other labels with similar meanings (e.g., *tung-sing-lyun* "homosexual") as search terms and included in the corpus all the articles about lesbians, gay men, and/or other sexual minorities published between November 1998 and December 2000. These articles were found in three sections of the newspaper: local news, international news, and news from Taiwan and mainland China. It is in these three sections that the putative objectivity of news reporting is often underscored. In contrast, it is more acceptable for journalists to express their personal opinions in other sections (e.g., the entertainment section) and in other types of articles (e.g., editorials).

To get a full picture of how and why *tongzhi* is used in *ODN*, I found it necessary to use both quantitative and qualitative methods to analyze the articles in the corpus. The quantitative analysis confirms the widespread use of *tongzhi* to refer to sexual minorities in *ODN*. It also shows that while *tung-sing-lyun je* "homosexual" is primarily found in medical and legal news, *tongzhi* is mostly used to refer to lesbians and gay men in highly sensationalized crime reports. Keeping in mind McConnell-Ginet's (2001) insight that words are endowed with meaning through their use in discourse, I performed a qualitative analysis to further examine the kinds of articles and headlines in which *tongzhi* is used, as well as the linguistic and paralinguistic features that co-occur with *tongzhi* and the effects that their combined use creates in context. The qualitative analysis reveals that the parodic use of *tongzhi* is one of the strategies adopted by *ODN* editors and journalists to make fun of gay rights activists and others with same-sex desire, so as to increase the entertainment value of the news story. At the same time, it mocks activists' demand for respect and equality and sows the seeds for the pejoration of the term. One might argue that, at least in *ODN*, *tongzhi* does not denote "sexual minorities" in general, but lesbians and gay men who engage in illegal or indecent behavior (Wong, 2005).

Through this project, I became more aware of issues surrounding the representativeness of print media data. These issues do not make print media data any less valuable, but they remind us to be cautious about the claims we make. Print media, despite the rich data that they offer, represent only particular kinds of writing produced by certain people for specific audiences. We need to understand the nature of print media if we are to use them as a source of data to study sociolinguistic variation and change – for example, who the target audience is, whose voices are represented (and not represented), and who is involved in the production of news language. Different newspapers espouse different ideologies and cater to different readerships. My research findings only apply to *ODN*; they have nothing to say about label use in other newspapers or in spoken discourse. In fact, I was surprised by the widespread use of *tongzhi* in *ODN*, which was at odds with my casual observations of label use in everyday interaction. This led me to investigate labeling practices in spoken discourse using data collected through directed interviewing and systematic observation (Wong, 2008).

Implications beyond Variationist Sociolinguistics

During my field research in Hong Kong, I shared with *tongzhi* activists my findings on the representation of sexual minorities in *ODN* and engaged in productive discussions with them about the political implications of labeling practices. Although my starting point was the semantic variation and change of *tongzhi*, I came to realize that this research has implications beyond variationist sociolinguistics. *Tongzhi* and other social category labels such as *gay*, *nigger*, and *queer* serve as an ideal terrain for investigating how power is exercised and contested through language. The right to make meaning is an important form of symbolic capital. Never solely about language per se, meaning contestation often operates at the level of use rather than through explicit discussion, and it serves to legitimize one's own interests and to naturalize one's viewpoint as the "truth." Through this process, individuals occupying different social positions attempt to inscribe their own ideologies and sometimes competing meanings to a linguistic form. Meaning contestation is a form of power struggle, and semantic variation and change is, in a sense, a product of that power struggle.

References

Fairclough, N. (1992). *Discourse and social change*. Cambridge: Polity Press.

Hill, J. (1993). Hasta la vista, baby: Anglo Spanish in the American Southwest. *Critique of Anthropology, 13*, 145–176.

McConnell-Ginet, S. (2001). "Queering" semantics: Definitional struggles. In K. Campbell-Kibler, R. J. Podesva, S. Roberts, & A. Wong (Eds.), *Language and sexuality: Contesting meaning in theory and practice* (pp. 137–160). Stanford, CA: CSLI Publications.

Wong, A. (2005). The reappropriation of *tongzhi*. *Language in Society, 34*, 763–793.

Wong, A. (2008). The trouble with *tongzhi*: The politics of labeling among gay and lesbian Hongkongers. *Pragmatics, 18*, 277–301.

Wong, A., & Zhang, Q. (2000). The linguistic construction of the *tongzhi* community. *Journal of Linguistic Anthropology, 10*, 248–278.

19 Conclusion

Christine Mallinson, Becky Childs, and Gerard Van Herk

In *Data Collection in Sociolinguistics: Methods and Applications*, we have explored a primary aim of sociolinguistics: to create and refine methods for the collection of data that reflect spoken and written language in use. As the contributors to this volume have suggested, research is only as solid as the data on which it is built, as sociolinguistic analysis and application depend on the valid and reliable collection of sociolinguistic data. The chapters and vignettes give important texture and insight into the many processes that are often involved in sociolinguistic data collection, from research design and ethical considerations to selecting appropriate methods, establishing archives, and sharing data with communities and other publics.

Several key themes have emerged in this volume. From the outset, a clear research design must be the solid foundation for data collection, as it will provide researchers with the ability to make informed methodological choices when gathering data. Throughout the research process, as several scholars in this volume assert, questions of ethics and how we should represent our research participants must be considered. We must also confront the increasingly pertinent question of how we as sociolinguists can preserve our data and how to provide access to it, whether to other scholars or to public groups. Several contributors further assert that methods of data gathering and data sharing are best seen as interrelated aspects of the same research process, which speaks to the growing concern among sociolinguists that we "give back" to those we study. Finally, the theme of communication is paramount, as scholars in this volume call upon sociolinguists to publicize our research and to share data and findings in ways that maximize our abilities to address research questions of academic and lay interest and to benefit those from whom we obtain our data.

The topic of methods of sociolinguistics data collection is one that is ever pertinent and relevant, and that changes as new methodologies are developed and explored. We continue the conversation about data collection in sociolinguistics on our website, http://sociolinguisticdatacollection.com, where we provide additional resources that accompany, complement, and expand upon the information in this volume and are of interest to those who continue to learn or teach about data collection in sociolinguistics.

Contributors

Michael Adams is Associate Professor of English at Indiana University. He is editor of *From Elvish to Klingon: Exploring Invented Languages* (2011), co-editor, with Anne Curzan, of *Contours of English and English Language Studies* (2011), and co-author, with Anne Curzan, of *How English Works: A Linguistic Introduction*, 3rd ed. (2012).

Jannis Androutsopoulos is Professor of German and Media Linguistics at the University of Hamburg. His research interests are in sociolinguistics and media discourse studies.

Philipp Sebastian Angermeyer is Associate Professor of Linguistics at York University, Toronto. His research interests include language contact and codeswitching, language and law, and interpreting. His recent publications appear in journals including the *International Journal of Bilingualism* and the *Journal of Sociolinguistics*.

Naomi S. Baron is Professor of Linguistics and Executive Director of the Center for Teaching, Research, and Learning at American University in Washington, DC. A Guggenheim and Fulbright Fellow, she is the author of *Alphabet to Email* (2000) and *Always On* (2008).

Joan C. Beal is Professor of English Language at the University of Sheffield. Her research interests include the history of English and the relationship between dialect, place, and identity.

Kara Becker is Assistant Professor of Linguistics at Reed College. Her research interests include American English regional and social variation, social practice and social meaning, and linguistic ideology.

Niko Besnier is Professor of Cultural Anthropology at Universiteit van Amsterdam. His most recent books are *Gossip and the Everyday Production of Politics* (2009) and *On the Edge of the Global: Modern Anxieties in a Pacific Island Nation* (2011).

Charles Boberg is Associate Professor of Linguistics at McGill University. He is the author of *The English Language in Canada* (2010) and co-author (with William Labov and Sharon Ash) of the *Atlas of North American English* (2006).

Kathryn Campbell-Kibler is Assistant Professor of Linguistics at The Ohio State University. Her research investigates the social meanings of sociolinguistic variation, focusing on how listeners process individual variables and incorporate them into their social perceptions of speakers.

Anne H. Charity Hudley is Associate Professor of Education, English, Linguistics, and Africana Studies and the William & Mary Professor of Community Studies at the College of William & Mary, where she also directs the William & Mary Scholars Program.

Becky Childs is Associate Professor of English at Coastal Carolina University. Her research focuses on language variation, language and identity, and language and gender in varieties of American and Canadian English, as well as the creation, encoding, and ethics of publicly accessible linguistic corpora.

Cynthia G. Clopper is Associate Professor of Linguistics at The Ohio State University. She received a PhD in Linguistics and Cognitive Science from Indiana University. Her major areas of expertise are phonetics, speech perception, sociophonetics, and laboratory phonology.

Karen P. Corrigan is Professor of Linguistics and English Language at the School of English Literature, Language and Linguistics at Newcastle University. Her research interests include Celtic Englishes, corpus linguistics, dialectology, discourse analysis, the history of English, sociolinguistics, and the sociology of language.

Alexandra D'Arcy is Associate Professor and the Director of the Sociolinguistics Research Lab at the University of Victoria. Her research interests include grammaticalization and variation and change in the morphosyntax and discourse-pragmatics of English, both diachronic and synchronic.

Mark Davies is Professor of Corpus Linguistics at Brigham Young University, where he specializes in research on language change and variation. He is the creator of several large corpora (http://corpus.byu.edu) used by tens of thousands of researchers from throughout the world.

Boyd Davis is Professor of Applied Linguistics at the University of North Carolina at Charlotte. Her research includes sociohistorical linguistics; Alzheimer's speech; narrative, pragmatics, and stance; and digital corpora. Recent edited collections are *Alzheimer Talk, Text and Context* (2005) and *Fillers, Pauses and Placeholders* (2010).

Paul De Decker is Assistant Professor of Linguistics at Memorial University, Newfoundland. His research examines the role of migration on sociophonetic variation and change as well as issues concerning sociolinguistic methodology.

Susan Ehrlich is Professor of Linguistics at York University, Toronto. She has published in the areas of discourse analysis, language and gender, and language and the law.

Marcia Farr is Professor Emerita of Language, Literacy, and Culture and English at The Ohio State University. A sociolinguist and linguistic anthropologist who studies language and cultural diversity, she publishes on language and identity, multilingualism, and literacy practices and ideologies.

Lisa Green is Professor of Linguistics and Director of the Center for the Study of African American Language at the University of Massachusetts Amherst. She is the author of *African American English: A Linguistic Introduction* (2002) and *Language and the African American Child* (2010).

Rania Habib is Assistant Professor of Linguistics and Coordinator of the Arabic Program in the Department of Languages, Literatures, and Linguistics at Syracuse University. She obtained her MA and PhD in Linguistics from the University of Florida.

Lauren Hall-Lew is Lecturer in Sociolinguistics in the Department of Linguistics and English Language at the University of Edinburgh. She primarily studies sociophonetic variation in Western US English and is particularly interested in the relationship between methodology and theory.

Joseph Hill is Assistant Professor in the Specialized Educational Services Department at the University of North Carolina at Greensboro. He is a co-author of *The Hidden Treasure of Black ASL: Its History and Structure* (2011).

Michol F. Hoffman is Associate Professor in the Department of Languages, Literature and Linguistics at York University, Toronto. Her interests include sociolinguistics, ethnicity, identity, language and dialect contact, language attitudes, historical linguistics, phonetics, and phonology. She works primarily in variation and change in Spanish and English in Toronto.

Barbara M. Horvath is an Honorary Associate in Linguistics at the University of Sydney. Her primary areas of interest are Australian English and Cajun English (with Sylvie Dubois) and the role of place in dialect studies (with Ronald Horvath).

Tyler Kendall is Assistant Professor of Linguistics at the University of Oregon. He works on the corpus-based and sociophonetic study of language variation and change, and has developed several software programs for archiving and analyzing sociolinguistic data.

Scott F. Kiesling is Associate Professor of Linguistics at the University of Pittsburgh. His work centers on language and masculinities, sociolinguistic variation, discourse analysis, ethnicity in Australian English, and Pittsburgh English. His publications include *Linguistic Variation and Change* (2011) and "Dude" (2004).

William A. Kretzschmar, Jr. is the Harry and Jane Willson Professor in Humanities at the University of Georgia. He is the Editor of the American Linguistic Atlas Project, the oldest and largest national research project to survey how people speak differently in different parts of the United States.

Erez Levon is Lecturer in Linguistics at Queen Mary University of London. Using variationist and ethnographic methods, his work examines language, gender, and sexuality in Israel, Britain, and the United States. He is the author of *Language and the Politics of Sexuality: Lesbians and Gays in Israel* (2010).

Ceil Lucas is Professor of Linguistics at Gallaudet University, where she has taught since 1982. She has broad interests in the structure and use of sign languages, and her works include the co-authored book *The Linguistics of American Sign Language*, 5th ed. (2011).

Christine Mallinson is Associate Professor of Language, Literacy, and Culture and Affiliate Associate Professor of Gender and Women's Studies at the University of Maryland, Baltimore County. Her research explores variation in American English, and she is the co-author of *Understanding English Language Variation in U.S. Schools* (2011).

Stephen L. Mann is Assistant Professor of English at the University of Wisconsin-La Crosse. He completed his doctoral work at the University of South Carolina and currently researches factors shaping gay men's attitudes toward gay male varieties of American English.

France Martineau is Professor in the French Department at the University of Ottawa, where she holds the Research Chair in Language, Identity, and Migration in French America. Her research is on the history of French, particularly varieties of French found in the Americas, and minority language communities.

Rajend Mesthrie is Professor of Linguistics at the University of Cape Town, where he holds a research chair on language, migration, and social change.

Arapera Ngaha is Senior Lecturer in the Department of Māori Studies at the University of Auckland, New Zealand. She lectures in contemporary issues for Māori that include Māori language and Māori sociolinguistics.

Patricia Causey Nichols is Professor Emerita of Linguistics at San José State University. Born in South Carolina and educated in its public schools, she studied sociolinguistics at Stanford University with Charles Ferguson.

Jennifer Nycz is Assistant Professor of Linguistics at Georgetown University. Her research focus is the quantitative analysis of phonological variation and the implications of this variation for phonological theory.

Bartlomiej Plichta is a doctoral candidate in hearing science at the University of Minnesota. He is the author of Akustyk, an online resource for reviews, tutorials, and recommendations for audio recording, processing, and analysis tools, at http://bartus.org.

Robin Queen is Associate Professor of Linguistics, Germanic Languages and Literatures, and English Language and Literatures and Arthur F. Thurnau Professor at the University of Michigan.

D. Victoria Rau is Professor and Director of the Institute of Linguistics at National Chung Cheng University in Taiwan. Her research expertise includes sociolinguistics, Austronesian linguistics, and applied linguistics. She is the author of *Yami Texts with Reference Grammar and Dictionary* (2006).

Randall Sadler is Associate Professor of Linguistics at the University of Illinois at Urbana-Champaign. He is also the co-owner of the EduNation Islands, located in the Virtual World Second Life, where he may often be found in the guise of his avatar Randall Renoir.

Edgar W. Schneider is Chair Professor of English Linguistics at the University of Regensburg, Germany. He has written and edited about 20 books and edits the journal *English World-Wide*. He has published and lectured on all continents on the dialectology, sociolinguistics, history, and semantics of English and its varieties.

Jennifer Sclafani is Assistant Professor of Applied Linguistics at Hellenic American University. She has published on the discursive construction of language ideologies in public media discourse and is currently examining sociolinguistic aspects of political identity in American presidential campaigns.

Robert Serpell is Professor of Psychology at the University of Zambia. His publications touching on sociolinguistics include *Culture's Influence on Behaviour* (1976), *The Significance of Schooling* (1993), and "Learning to Say It Better" in *The New Englishes* (1983).

James N. Stanford is Assistant Professor of Linguistics and Cognitive Science at Dartmouth College. His research focuses on language variation and change in less commonly studied indigenous minority communities. Before graduate school, he spent about seven years in mainland China where he learned to speak Mandarin Chinese and Sui, a Tai–Kadai language of southwest China.

Donna Starks is Senior Lecturer in the Faculty of Education at La Trobe University, Melbourne, where she teaches Applied Linguistics and Language Education. Her research interests focus on the interface of language and identity and the development of ethnolects.

Sara Trechter is the former Associate Dean of Graduate Studies and Professor of Linguistics at California State University, Chico. Her fieldwork and research on Lakhota and Mandan have focused on gender pragmatics, language endangerment, and maintenance.

Clive Upton is Emeritus Professor of Modern English Language at the University of Leeds. Involved with the Survey of English Dialects for almost 40 years, he advised the British Broadcasting Corporation on its "Voices" project, 2004–2005. He is also the editor of the journal *English Today*.

Gerard Van Herk is Associate Professor and Canada Research Chair in Regional Language and Oral Text at Memorial University of Newfoundland. His research focuses on historical and contemporary varieties of English in

Canada, the United States, and the Caribbean. He is the author of *What Is Sociolinguistics?* (2012).

Cécile B. Vigouroux is Associate Professor in the Department of French at Simon Fraser University in Canada. Her research interests include sociolinguistics, ethnography, linguistic anthropology, ideology of language, discourse analysis, Africa, Francophonie, and migration.

James A. Walker is Associate Professor in Linguistics at York University, Toronto. His interests include sociolinguistics, multilingualism, ethnicity, language contact, phonology, and morphosyntax. He has worked on sociolinguistic variation and change in African American, Canadian and Caribbean English.

Tracey L. Weldon is a quantitative sociolinguist specializing in African American English and Gullah. She is Associate Professor in the English Department and the Linguistics Program at the University of South Carolina.

Walt Wolfram is the William C. Friday Distinguished University Professor at North Carolina State University, where he directs the North Carolina Language and Life Project. He has pioneered research on social and ethnic dialects since the 1960s and has published more than 20 books and over 300 articles.

Andrew D. Wong is Associate Professor of Anthropology at California State University, East Bay. He has studied the role of ideology in semantic variation and change, the relationship between genre and social change, and the social meanings of unconventional spelling.

Index

Page numbers in *italics* denote tables, those in **bold** denote figures.