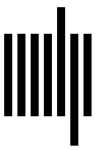
This is a section of doi:10.7551/mitpress/12200.001.0001

The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

19 Data Management Practices in an Ethnographic Study of Language and Migration

Lynnette Arnold

1 Introduction

In this chapter, I discuss the data collection and management processes utilized in an ethnographic study of language and migration. While much linguistic research is ultimately concerned with understanding language itself, this research took a sociocultural linguistic approach, asking "how the empirical study of language illuminates social and cultural processes" (Bucholtz & Hall 2008:405). In this case, I used ethnographic methods to explore everyday communication among Salvadoran migrants living in the United States and their families back home to gain new insights into experiences of migration and transnational life (Arnold 2012, 2015b, 2019). In this project, the ethical aspects of data practices (Holton, Leonard, & Pulsifer, chapter 4, this volume) were clear from the outset, given the substantial power imbalances between myself, as an English-speaking, middle-class US citizen who can travel the world with ease and the monolingual Spanish-speaking research participants, residents of an impoverished Salvadoran village, who can only access transnational mobility at great risk. Given the population that I worked with and the types of data I gathered, this study presented ethical and practical challenges at each stage of the data life cycle (Mattern, chapter 5, this volume), including data collection, processing, and management. In what follows, I describe how I worked to resolve these concerns while also laying out how ethnographic research such as this complicates the open data model. To contextualize this discussion, I begin with a brief discussion of my research project and methods.

2 Project and methods

This research involved an ethnography of communication (Hymes 1964, 1972) conducted from 2009 to 2014 with families impacted by migration, specifically undocumented Salvadoran migrants in the United States and their relatives back home in El Salvador. Despite being unable to travel to visit one another, these families remained in regular contact through a range of digital communication technologies that are becoming increasingly available around the world. Everyday cross-border conversations within these families provide an important window into the impacts of migration, as language becomes the primary means for sustaining familial relationships and navigating the everyday concerns of family life when migrants and their loved ones must live stretched across borders for years at a time (Arnold 2016).

To capture the complete circuit of transnational family life, I conducted a multisited ethnography (for a more complete discussion of this method in linguistic research, see Dick & Arnold 2017). My research sites included research in a rural Salvadoran village and three different urban locations in the United States where migrants from this village had settled. This phase of the project involved video and audio recording twenty-four interviews with members of twelve transnational families—both migrants and non-migrants—as well as participant observation. I joined in many aspects of the families' daily lives, using field notes to track what I was learning about the role of communication in cross-border kinship. I saw individuals receive phone calls and other forms of communication from distant relatives and observed how families living together in either country would talk about these transnational conversations, discussing what needed to be communicated and who would be responsible to communicate it. Through attending carefully to such conversations, I began to understand the crucial but complex role that such phone calls in particular played in these families' lives. I therefore conducted a final five-month stage of intensive data collection with two multigenerational families.

In this phase of the study, I gathered recordings of everyday conversations, both on the phone and face-to-face, within the families. This intimate domain of study was made possible by ongoing relationships with the participants that were first forged during the four years that I spent living in El Salvador (2001–2005), when I worked with gender development and youth engagement programs primarily in the rural village that later became the hub of my research. Throughout the course of this study, I sought to honor the families' trust in me through careful consideration of data management practices at each stage of the data life cycle. In the following sections, I discuss data collection, data processing and storage, and data sharing and citation, describing key considerations and strategies for each in turn.

3 Data collection

Because of the multisited nature of the research, my data collection methods needed to work both in urban settings in the United States and in rural El Salvador, where access to electricity was unpredictable and there was no regular Internet connection. In addition, I needed methods that would allow for the recording of both face-to-face and remote, technologically mediated conversations. At the same time, it was crucial to me that family members be able to record data on their own as much as possible. On the one hand, this was a practical consideration, both to minimize the observer's paradox that is an inevitable part of sociolinguistic research (Labov 1972) and also to facilitate simultaneous data collection across multiple sites. At the same time, my interest in conducting "research with" (Cameron et al. 1993:87) also sprang from my ongoing ethical concern to gather data with as much sensitivity as possible to the intimate nature of the conversations themselves.

For the final intensive stage of data collection, I hired and trained a family research assistant in each family to assist with making recordings; I asked young adults to fill these roles, because they tended to have the best literacy skills and were also more accustomed to using digital technologies.[1] Although these young adults had full-time jobs, they generally had less extensive family responsibilities than the older adults did, leaving them more free time for such work. This collaborative model of working had its challenges: some data were inevitably lost during the process of learning to operate the recording equipment. At the same time, working closely with these young people was crucial to the success of the project. Once the recordings were made, these research assistants provided vital background knowledge that helped to contextualize the conversations, providing feedback during regular face-to-face meetings or through digital correspondence. Family research assistants received a stipend of $100 per month and I also worked to provide mentoring. For instance, at their request, I took them to visit local colleges and libraries and helped them master the use of local public transportation systems (often with English-only signage), which gave them greater freedom of movement.

The family research assistants helped with the recording of both face-to-face and phone conversations. In-person conversations at each site were recorded using basic video cameras, tripods, and a sixteen-gigabyte microSDHC (secure digital high capacity) data card to store files. I passed all this equipment on to the families for their own use at the end of the study. Together, we gathered eighty-seven video recordings of spontaneous face-to-face interaction, totaling about fifty hours of video data. The recordings include everyday activities such as sharing meals, cooking, doing homework, and playing games, as well as special activities such as holiday and birthday celebrations. Phone conversations were recorded using a maximally flexible recording technology that would work regardless of the type of cell phone being used. The Olympus TP-8 consists of an earpiece with a microphone mounted on the back. Participants would wear this earpiece in their ear and connect it to an MP3 recorder (the Olympus VN-8100PC);[2] the phone was then held up to the ear with the earpiece so that phone calls could easily be recorded. These MP3 recorders were placed in carrying cases with carabiner clips, thus allowing the user to be as mobile as usual while making recordings of cell phone conversations. Although each recorder had two gigabytes of internal memory, I installed an eight-gigabyte microSDHC data card to be sure the recorders would not run out of space.

Using this methodology, sixty-seven transnational phone calls were recorded over the course of four months, ranging in length from two minutes to two hours, most averaging about twenty to thirty minutes, for a total of twenty-five hours of recordings. For both video and audio recorders, I showed the family research assistants how to delete recorded data, so that any conversations

that had been recorded could be removed after the fact if anyone in the family had any concerns. I also reviewed all recordings before beginning analysis to identify any sensitive segments—particularly those concerning legal matters involving undocumented migrants—which I deleted permanently. The types, formats, and amounts of data gathered over the course of the project are shown in table 19.1.

4 Data processing and storage

The quantity of recordings and the diversity of types of data gathered created challenges at the next stages of the data life cycle—data processing and storage—which I will consider jointly here because they were very much interwoven in this project. In addition to managing large file sizes, particularly with the video recordings, I also worked to maintain the confidentiality of the data in both its processing and storage. These considerations led me to avoid using cloud-based storage options and instead to rely on external drives. All data files, whether audio or video, were copied from the cards on which they had been originally recorded onto a one-terabyte external drive, after which they were deleted from the recording device or cards. To minimize the possibility of data loss, I copied the entire contents of this drive onto a second backup drive using rsync (https://rsync.samba.org/), an open source utility that efficiently synchronizes files across hard drives by comparing file modification times and file sizes.[3] While this process allowed me to easily create a backup of my data by plugging both the original drive and the backup into my laptop and running the program, it did require me to manually update the backup drive as I continued to process the data, revealing the challenges of an active storage process (see Mattern, chapter 5, this volume).

Table 19.1
Data gathered in this study

Type	Format	Quantity
Field notes	Word documents, non–optical character recognition PDF scans of handwritten notes	200 pages
Interviews	.mov and .wav files for each	24 interviews (45 hours of recordings)
Face-to-face interaction	.mov files	87 recordings (50 hours)
Phone conversations	.mp3 files	67 calls (25 hours)

Data processing proceeded in the same way for all of my recordings, regardless of type, with the goal of tracing discourse patterns across the corpus. As a fluent second-language speaker of Salvadoran Spanish, I conducted all of my analysis with the original Spanish and only translated transcription excerpts for presentation and publication. Each audio or video file was first indexed using Excel, creating a time-stamped summary for each recording that included aspects of the content of the conversation as well as salient linguistic forms (see figure 19.1). The indexes averaged a new line of notation for every thirty seconds of conversation.[4] Based on an inductive coding of the indexes, I decided which parts of the recordings to transcribe. Time-aligned transcripts were made using ELAN, segmenting the data into intonation units. To obtain the waveforms necessary to help with this segmentation, I had to transform some of the data (see Han, chapter 6, this volume), converting all MP3 files to WAV format.[5]

Each audio or video recording thus had several derivative data files associated with it (an Excel spreadsheet index and three ELAN transcription files), while some audio files had both an MP3 and a WAV format. To keep all of the related files associated with one another, I followed a strict file-naming convention linked to a metadata spreadsheet that I used to process the data (see Mattern, chapter 5, this volume for more discussion of file naming).[6] I entered each separate recording on its own line of the spreadsheet, filling out each of the fields in order (see figure 19.2). Using the *concatenate* function in Excel, I then set up the spreadsheet to automatically compile the file name that I then copied and used for recordings, indexes, and ELAN files, simply changing the file format to the appropriate type.

I designed this file-naming system to explore the questions that had emerged from my ethnographic research, with each piece of the file name being tied to some crucial aspect of the data. The first indicator for each file was family name, which was crucial as I was concerned with how communication functioned within the family; similarly, the location where the data had been recorded was crucial for attending to the transnational dimensions of this communication. Tracking dates next allowed me to trace particular topics or themes of conversation over time. Finally, the event portion of the file name encoded

Time	Index
0:00:00	brief greeting
0:00:10	OP issues complaint about her eye
0:00:25	OP asks F what he's doing - soccer game - minimal responses from F.
0:00:48	OP returns to eye issue
0:01:00	matching story by F - basura in his eye at work - OP matches with her pain
0:01:15	F continues story - OP asks about eye drops - tells her story
0:01:45	OP reports going to doctor - F asks about operation - technical terms - 'la carnosidad'
0:02:33	F asks about his grandfather - OP reports - lots of QS: he can't get up etc.
0:03:00	OP - reports on grandfather's health problems - QS: 'no voy a durar mucho'
0:03:35	OP & F talk about whether grandfather will be OK
0:03:50	OP reports what grandfather eats in great detail - more QS
0:04:35	F makes complaint about not having beans cooked for him to eat
0:04:45	OP continues with elaboration of grandfather's diet
0:05:05	OP: grandfather is like a child - more report about his diet
0:05:45	F continues/elaborates complaint about tortillas frias
0:06:00	OP reports on her diet - not allowed to eat tortillas but sends corn to molino
0:06:20	F asks who OP means (a woman who has something to do with molino) - he can't figure it out (OP reports who her parents are, where she lives)
0:06:40	F reports his day's activities: going to work and dentist
0:06:46	OP asks about more details of F's dental problems - sympathetic response
0:07:20	OP & F talk about dental details - pain and procedure
0:07:42	matching story by OP about her eye operation - 'todo duele'
0:08:14	OP tells F to be careful - reminds him of previous time he got sick with something she thinks is similar - although not clear with what
0:08:25	F denies similarity to previous case - 'los otros nervios no usted'
0:08:47	OP - well-wishing re: F's health

Figure 19.1
Sample index of phone conversation.

Family	Country	Year	Month	Day	Event	Format	Filename
Portillo	US	2013	11	09	Breakfast	mov	Portillo_US_2013-11-09_Breakfast.mov
Portillo	US	2013	11	10	Call-F-O	mp3	Portillo_US_2013-11-10_Call-F-O.mp3
Portillo	US	2013	11	10	Tamales1	mov	Portillo_US_2013-11-10_Tamales1.mov
Portillo	US	2013	11	10	Tamales2	mov	Portillo_US_2013-11-10_Tamales2.mov

Figure 19.2
Filename portion of metadata spreadsheet.

a shorthand description of the type of recording, for instance "Breakfast," "Tamales1," or "EveningGames"; phone calls were described as "call" and then appended with the pseudonym initial of all the participants in the order they spoke.[7] Multiple files with the same name were disambiguated numerically in sequential order (e.g., "Tamales1," and "Tamales2"). In addition to highlighting important aspects of each communicative event, utilizing consistent file-naming conventions across the complete data set was essential for keeping this large corpus organized and therefore as easy to work with as possible.

I did not complete the data processing alone, but rather with a group of undergraduate research assistants who helped primarily with the transcription in exchange for course credit and mentoring. Each semester, I found one or two students with the necessary linguistic background and trained them in the basics of discourse transcription using ELAN.[8] I maintained confidentiality of

the data by consistently using participant pseudonyms with all of the research assistants throughout the data processing. Working with these students certainly accelerated the transcription, and as members themselves of extended transnational families, they shared keen insights into the nature of the data that have been fundamental to my ongoing research. Working with these students also created space for ongoing mentoring relationships, one of the ethical interventions whereby I sought to support the community that had helped me in my research. At the same time, this collaboration—like the one with the family research assistants during data collection—produced challenges in processing the data. Working with different assistants over a short period of time, each of whom was transcribing on their own, sometimes resulted in multiple ELAN files for a single recording, each with different parts of the recording transcribed. For instance, file 1 could have minutes 0–5 and 15–20 transcribed, while file 2 would have minutes 8–12 and 24–30 transcribed. This overlapping meant that syncing these separate files is not a straightforward, automatable undertaking. Moreover, I needed to check students' transcriptions before adding them to my main database of transcription. I created an ad hoc Excel sheet to track student transcriber assignments and files, but the lack of planning in data processing at this stage created a bottleneck in the workflow that future collaborative research would do well to avoid.

5 Sharing data

In the final stage of the data life cycle, data sharing, my ethnographic research diverges significantly from the open data model advanced in this volume, because I ultimately decided not to deposit my data with an archive nor to share it with other researchers. Nevertheless, considering research like mine that opts to maintain closed data can illuminate our thinking about the open data model, in particular revealing its limitations and challenges. For that reason, in this section I lay out the ethical and practical reasons that motivated my decision not to share data, while also outlining a non-academic form of public diffusion I engaged in at the request of my participants.

Maintaining closed data has long been, and still remains, a largely unquestioned norm within ethnographic research. This is reflected, for instance, in the fact that data citation practices in ethnographic journals almost never allow for particular segments to be traced back to their position in the larger corpus. In fact, I am often asked to remove reference to specific files and time stamps from examples in manuscripts submitted for publication and have thus opted to track this information for myself as part of my metadata spreadsheet. The pervasiveness of closed data is clearly tied to ethnographic epistemologies, in which knowledge emerges through the interaction between researcher and participants (Clifford & Marcus 1986; James, Hockey, & Dawson 1997). As such, ethnographic data do not easily lend themselves to interpretation and analysis by those not involved in the research process. In my research, for instance, the data I gathered were deeply embedded in the everyday lives of families, such that making sense of them at times required more background knowledge even than I had gathered in fifteen years of working with the participants. Such practical considerations were certainly part of my decision not to share my data.

It is possible that ethnographic data may be of interest to researchers pursuing other questions; for instance, my data could allow for a study of the lexical and morphosyntactic properties of Salvadoran Spanish. Questions of reproducibility alone are thus not sufficient to justify closed data. Beyond practicality, however, the question of whether to share data is fundamentally an ethical one, particularly when research involves working with minority language communities who have long been subject to extraction and misappropriation of their knowledge (Holton, Leonard, & Pulsifer, chapter 4, this volume). Although my research involved speakers of a majority language (Spanish), the political, economic, and social marginalization of undocumented immigrants in the United States meant that sharing data could potentially put some of my participants in danger of deportation. Other considerations, such as the intimate nature of the data gathered and the fact that children were often involved in the recordings, also factored into my decision-making process. For these reasons, it was clear that sharing any data in their raw form would be an ethical violation of the trust the participants had placed in me in allowing me to conduct my research. While the ethical implications of shared data were perhaps more stark in my project than in most linguistic scholarship, ultimately all our research involves people in one way or another. It is thus crucial for the field to take seriously the ethical implications of data sharing for all of the varied types of linguistic research.

Although the ethical imperative to maintain closed data was clear to me, at the same time, many of my participants emphasized that part of their motivation for participating in my research was that they wanted me to share their stories. They felt that US citizens did not sufficiently understand the realities of migrants' lives and their ties to family across borders. Because many of them had known me first as a community worker before I became a researcher, they saw me as an ally in the struggle for justice for immigrant communities, a positioning that I affirmed as well. They thus wanted me to share their stories with the broader public as a means of raising awareness about the full impact of US immigration policy on individuals living around the world. I thus felt the imperative to share particular aspects of my data with the broader public, beyond the narrow scope of academic publications and conference presentations, and have done so as best I can (Arnold 2015a, 2018, 2020; Hallett & Arnold 2016, 2018).

To facilitate the kind of public sharing that my participants were requesting, in my consent forms I incorporated questions about data sharing in three contexts: in presentations, in academic publications, and in broader public forums.[9] For each context, participants could select what level of sharing they were comfortable with, ranging from no sharing, sharing of anonymized recordings, to sharing of original recordings. Unsurprisingly, out of thirty participants who agreed to have their data shared publicly, only four were comfortable sharing materials that were not anonymized. Because most of my recordings involve several participants, this has effectively meant that any data I shared publicly must first be anonymized. This introduces a challenge, because full anonymization, particularly of video data, is quite labor intensive. Moreover, given the ethnographic nature of the data, it must be contextualized for the impact and breadth of the stories to be adequately conveyed to a broader audience.

I therefore experimented with using a blogging platform to publicly share segments of my data as part of larger stories about Salvadoran migration to the United States (for an example, see https://alasmigratorias.wordpress.com/2013/10/28/poverty/). To produce this shareable data, I combined anonymized audio with open access artwork to create short videos that tell individual migration stories. These videos are then embedded within a written narrative that recounts the history and current realities of unauthorized migration from El Salvador to the United States. This format works well, but it requires a great deal of time and effort that generally go unrecognized as valid academic labor for purposes of the job market or tenure and promotion. Nevertheless, in an era of virulent anti-immigrant sentiment, which is increasingly directed toward Central Americans, this form of data sharing is an ethical necessity.[10]

6 Conclusions

In this chapter, I have traced the data life cycle throughout an ethnographic study of language and migration, describing my data collection, processing, and sharing practices. Ultimately, although I chose not to archive or share my corpus as a whole, writing about data management nevertheless remains a key component of the openness that is a central principle of the social science data movement (see Gawne & Styles, chapter 2, this volume). Although ethnographic research ultimately does not strive toward replicability of particular studies as a means of asserting validity, I have shared my data practices here as a means of supporting future ethnographers of language and communication to think through their data management practices at each stage of the research. Such discussions are a crucial component of recent calls to incorporate ethnography as a more central methodological approach within linguistics (Snell, Shaw, & Copland 2015). More broadly, my project illustrates the benefits of collaboration with community members and students at multiple stages of the research process. In particular, I have pointed to unforeseen challenges that arose in the course of these collaborations, considerations that will be of use to those planning future collaborative research. Finally, it is my hope that consideration of research such as mine can be of help to proponents of more open data within linguistics. By highlighting the ethical complications and limitations of such an approach, I seek to contribute to an open data movement that does not let lofty goals override the absolute necessity of carefully implementing this approach in ways that are as variable as our research.

Acknowledgments

The research reported on in this chapter was made possible through the Jacob K. Javits Fellowship Program, the Chicano Studies Institute of the University of California at Santa Barbara, and the University of California

Institute for Mexico and the United States. I wish to thank two anonymous reviewers and Patrick Hall for their comments on this chapter, which have made it much stronger. Any remaining errors are my own.

Notes

1. Older adults in the families had often had little to no formal schooling, and my participants also included preliterate and school-aged children.

2. Given the degraded audio quality of phone calls, and in order to purchase several versions of the complete recording setup, I decided to opt out of the more expensive WAV recorders usually utilized for linguistic research.

3. I owe a debt of gratitude to Patrick Hall for introducing me to rsync and helping me set it up.

4. Many thanks to Mary Bucholtz for teaching me this method of processing sociocultural linguistic data.

5. Phone conversations were recorded in compressed format due to their degraded audio quality that made higher-quality WAV recorders useless.

6. Thanks to Laura Robinson and her graduate seminar on research methods for introducing me early on to the importance of file-naming conventions.

7. The calls generally involved a series of dyadic conversations that often involved several different individuals on both ends of the line, resulting in the sometimes long string of initials of pseudonyms.

8. I was fortunate to be able to find six students of Salvadoran descent who were able to understand my data.

9. There is a long-standing critique of informed consent and its mismatch with ethnography (Bell 2014) that I certainly experienced in my research. But my institutional affiliations and funding nevertheless made it necessary to follow these procedures.

10. Using social media and other online platforms also raises questions about reach, and I certainly don't think my blog posts were able to reach a very wide audience. A possible resolution to this consideration comes in the form of curated online venues (e.g., Medium https://medium.com/ and The Conversation https://theconversation.com/us) that help scholars present their research to a broader public. (See for instance, Arnold 2020, which has reached almost 6,000 unique readers from the time of publication to November 2020).

References

Arnold, Lynnette. 2012. "Como que era Mexicano": Cross-dialectal passing in transnational migration. *Texas Linguistics Forum* 55:1–9.

Arnold, Lynnette. 2015a. Deleting the I-word in Santa Barbara. *Center for California Languages and Cultures* (blog). January 28, 2015. http://www.ccalc.ucsb.edu/deleting-i-word-santa-barbara-guest-post-lynnette-arnold.

Arnold, Lynnette. 2015b. The reconceptualization of agency through ambiguity and contradiction: Salvadoran women narrating unauthorized migration. *Women's Studies International Forum* 52 (September): 10–19. https://doi.org/10.1016/j.wsif.2015.07.004.

Arnold, Lynnette. 2016. Communicative care across borders: Language, materiality, and affect in transnational family life. PhD dissertation, University of California, Santa Barbara.

Arnold, Lynnette. 2018. Imprisoning families is not the solution. *Youth Circulations* (blog). June 20, 2018. http://www.youthcirculations.com/blog/2018/6/20/imprisoning-families-is-not-the-solution.

Arnold, Lynnette. 2019. Language socialization across borders: Producing scalar subjectivities through material-affective semiosis. *Pragmatics* 29 (3): 332–356. https://doi.org/10.1075/prag.18013.arn.

Arnold, Lynnette. 2020. Four tips for staying connected during coronavirus, from migrants who live far from family. *The Conversation*. March 30. https://theconversation.com/4-tips-for-staying-connected-during-coronavirus-from-migrants-who-live-far-from-family-134362.

Bell, Kirsten. 2014. Resisting commensurability: Against informed consent as an anthropological virtue. *American Anthropologist* 116 (3): 511–522. https://doi.org/10.1111/aman.12122.

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

Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Ben Rampton, and Kay Richardson. 1993. Ethics, advocacy, and empowerment: Issues of method in researching language. *Language and Communication* 13 (2): 81–94.

Clifford, James, and George E. Marcus, eds. 1986. *Writing Culture: The Poetics and Politics of Ethnography*. Berkeley: University of California Press.

Dick, Hilary Parsons, and Lynnette Arnold. 2017. Multisited ethnography and language in the study of migration. In *The Routledge Handbook of Migration and Language*, ed. Suresh Canagarajah, 397–412. London: Routledge, Taylor and Francis.

Hallett, Miranda Cady, and Lynnette Arnold. 2016. Detention, disappearance, and the power of language. *Anthropology News* 57 (11): e250–253. https://doi.org/10.1111/AN.241.

Hallett, Miranda Cady, and Lynnette Arnold. 2018. Compounding the crisis. *North American Congress on Latin America (NACLA)* (blog). July 24, 2018. https://nacla.org/news/2018/07/24/compounding-crisis.

Hymes, Dell. 1964. Introduction: Toward ethnographies of communication. *American Anthropologist* 66 (6): 1–34. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00010.

Hymes, Dell. 1972. Models of the interaction of language and social life. In *Directions in Sociolinguistics: The Ethnography of Communication*, ed. John Gumperz and Dell Hymes, 35–71. New York: Blackwell.

James, Allison, Jenny Hockey, and Andrew Dawson, eds. 1997. *After Writing Culture: Epistemology and Praxis in Contemporary Anthropology*. New York: Routledge.

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

Snell, Julia, Sara Shaw, and Fiona Copland, eds. 2015. *Linguistic Ethnography*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137035035.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363