14 Managing Sociolinguistic Data with the Corpus of Regional African American Language (CORAAL)

Tyler Kendall and Charlie Farrington

1 Introduction

This data management use case describes our work building and managing the Corpus of Regional African American Language (CORAAL; Kendall & Farrington 2020a), the first corpus devoted to making publicly available spoken language data sets for research on and education about African American Language.[1] While CORAAL is specifically focused on African American Language, it also represents a rare case where sociolinguistic interview recordings have been collected and published as an open access, public corpus. As such, we hope our work can act as a model for the publication and management of other sociolinguistic data. Publicly oriented sociolinguistic data sets are rare and published treatments of their design criteria and decision processes are even more rare. Thus, we hope this chapter can add to the limited literature on public sociolinguistic corpora and can aid in the development of best practices for linguistic data management and for sociolinguistic corpus building in particular (for other discussions see, e.g., Beal, Corrigan, & Moisl 2007a, 2007b; Kendall 2007, 2008, 2011; Poplack 1989; Yaeger-Dror & Cieri 2014). Readers are also encouraged to read the CORAAL *User Guide* (Kendall & Farrington 2020b), which is available from the CORAAL website (https://oraal.uoregon.edu/coraal).

Before proceeding, we should ask: Why African American Language? And, why public corpora? African American Language (AAL) is an intentionally broad term meant to encompass all varieties of language use in African American communities reflecting "differences in age/generation, sex, gender, sexuality, social and socioeconomic class, region, education, religion, and other affiliations and identities that intersect with one's ethnicity/race and nationality" (Lanehart 2015:3). Within sociolinguistics, AAL has been one of the most studied varieties of English (or likely of any language). For over fifty years, researchers have turned to spoken language data from African American individuals and communities to investigate core, basic questions in sociolinguistics and in the history of American English. Researchers, since Labov et al. (1968) and Wolfram (1969), have also studied AAL to combat public myths and prejudice about language (see, e.g., Rickford 1999; Baugh 2000, 2005; or Wolfram 2008 for fuller discussions and Lanehart 2015 for a comprehensive treatment of AAL). Nonetheless, despite such an intense research tradition, almost no primary data are available for researchers or educators. As Kendall, Bresnan, and Van Herk (2011) discussed, research attempting to compare across AAL-speaking communities has been limited. As social science disciplines, and linguistics in particular, take more seriously issues of replicability, big data, and open access (see Berez-Kroeker et al. 2018; Gawne & Styles, chapter 2, this volume), it is, frankly, imperative, that publicly available data sets be developed and shared.[2]

2 CORAAL

CORAAL focuses on providing public access to recorded speech from regional varieties of AAL. CORAAL is a long-term corpus-building project conceived of in terms of several individual components. Each component includes audio recordings along with time-aligned orthographic transcription. The core components of CORAAL focus on AAL in Washington, DC, the nation's capital, a city with a long-standing African American majority, and the site of much early research on AAL (Farrington & Schilling 2019). CORAAL:DC, first released in January 2018, is composed of over one hundred sociolinguistic interviews with AAL speakers in DC born between 1890 and

2005. It consists of two sub-components, CORAAL:DCA, recorded around 1968 (Kendall, Fasold, et al. 2018), and CORAAL:DCB, recorded around 2016 (Kendall, Quartey, et al. 2018). In addition to CORAAL:DC, CORAAL is scheduled to increasingly include several smaller components to provide regional breadth. The three supplemental components available at the time of this writing are CORAAL:PRV (Rowe 2005; Rowe et al. 2018), which includes fifteen sociolinguistic interviews from a rural African American community in central North Carolina, CORAAL:ROC (King 2018; King et al. 2020), which includes thirteen sociolinguistic interviews from Rochester, a city in Western Upstate New York, and CORAAL:ATL (Farrington et al. 2020), which includes thirteen conversational interviews from Atlanta, Georgia. Additional supplemental components will be released in periodic updates as they become available. Updates typically include minor transcription revisions and new annotation versions, as well as new interviews and, when available, entire new supplemental components.

All CORAAL recordings are anonymized and orthographically transcribed with time alignment at the utterance level. Audio is available in high-quality uncompressed (.wav) format, and transcripts are available in three formats: Praat TextGrid (.TextGrid) files (Boersma & Weenink 2018), ELAN (.eaf) files (Wittenburg et al. 2006), and as plain text (.txt) files with tab-delimited fields. The corpus is intended to be downloaded by users for direct use, but CORAAL is also accessible through a website, CORAAL Explorer (see section 3.4), which provides browsing (audio and transcripts) and search capabilities directly in a web browser. A syntactically parsed version of much of the data is also scheduled for upcoming release.

3 Building CORAAL

Building CORAAL involved, and continues to involve, a combination of advanced planning and the development of a consistent workflow for data collection, processing, and sharing. In this section, we discuss each of these aspects of the project.

3.1 Collection

CORAAL's data come from two broad sources, legacy materials and new recordings collected specifically for the public corpus. By *legacy materials*, we mean sociolinguistic recordings that were collected not intentionally for inclusion in CORAAL. This includes a subset of recordings from Ralph Fasold's foundational fieldwork in Washington, DC (Fasold 1972), which form the basis for CORAAL:DCA and were in many ways a motivating factor behind the entire CORAAL project. However, these legacy materials also include recordings from fieldwork projects contemporary with the development of CORAAL, for instance work by Sharese King for her dissertation (King 2018) on AAL in Rochester, New York (CORAAL:ROC).

For these legacy collections, the CORAAL development team makes arrangements with the collections' primary investigators to obtain the data. In the case of CORAAL:DCA, the reel-to-reel recordings were first digitized at the North Carolina State University Sociolinguistics Lab, and then uploaded to the Sociolinguistic Archive and Analysis Project (SLAAP; Kendall 2007; https://slaap.chass.ncsu.edu/) for storage. The development team then organized the recordings according to a 4×3 demographic table (see table 14.1), covering four age groups and three social classes. The social class groups, listed simply as classes 1, 2, and 3, are meant to capture broad social class differences and are not meant to represent theoretically motivated socioeconomic classes. Qualitative labels are included in the CORAAL *User Guide*, ranging from lower working class to upper middle class, with differences depending on the specific component (see tables 14.2 and 14.3). Our goal was to have two male and two female speakers for each demographic cell (48 speakers), but this was not always possible and when more data have been available, we have opted to include more than this many speakers per cell.

One of the goals of Fasold's (1972) study was a focus on teenagers, which resulted in a sample that was not balanced according to our designs for CORAAL. Furthermore, some of Fasold's recordings with the youngest speakers were quite short. Because of these factors, we

Table 14.1
Demographic matrix targeted for CORAAL:DC components

	Social class 1		Social class 2		Social class 3	
Age group	Male	Female	Male	Female	Male	Female
12–19	2	2	2	2	2	2
20–29	2	2	2	2	2	2
30–50	2	2	2	2	2	2
51+	2	2	2	2	2	2

Table 14.2
Demographic matrix for CORAAL:DCA component

	Social class 1 (≈LWC)		Social class 2 (≈UWC)		Social class 3 (≈MC)	
Age group	Male	Female	Male	Female	Male	Female
12–19	8	5	6	7	6	6
20–29	1	1	3	0	3	5
30–50	1	2	3	0	4	1
51+	2	0	1	1	2	0

LWC = lower working class; UWC = upper working class; MC = middle class (see *User Guide*).

Table 14.3
Demographic matrix for CORAAL:DCB component

	Social class 1 (≈WC)		Social class 2 (≈LMC)		Social class 3 (≈UMC)	
Age group	Male	Female	Male	Female	Male	Female
12–19	3	3	1	1	1	1
20–29	3	3	1	2	0	1
30–50	3	3	3	2	2	2
51+	2	1	1	5	2	2

WC = working class; LMC = lower middle class; UMC = upper middle class (see *User Guide*).

included at least twelve speakers in each of the social class groups for the youngest age group, for a total of sixty-eight speakers for CORAAL:DCA (see table 14.2 for the full demographic matrix of CORAAL:DCA).

For CORAAL:DCB (table 14.3), some demographic groups have been harder to recruit than others and so there too our actual sample of speakers diverges somewhat from our initial target. More information about speaker selection for CORAAL:DCA and CORAAL:DCB can be found in the "CORAAL:DCA (Washington, DC 1968)" and "CORAAL:DCB (Washington, DC 2016)" sections in the *User Guide*).

The non-core components within CORAAL target a smaller number of speakers, with a goal of two to three speakers per demographic cell (table 14.4) within a single social class. In the case of legacy collections, such as recordings from Princeville, North Carolina (Rowe 2005), recordings were selected from the larger collection (N = 37) to fit this smaller demographic matrix.

Many additional resources, including the ones in SLAAP, are not fully publicly available and so will not become a part of CORAAL proper, but can be made available to bona fide researchers. In addition to CORAAL components,

Table 14.4
Demographic matrix targeted for other CORAAL components

Age group	Male	Female
18–29	2	2
30–50	2	2
51+	2	2

CORAAL also includes "Supplements", which are prepared and curated by the CORAAL team to highlight recordings and selections from larger datasets important to the field of sociolinguistics (https://oraal.uoregon.edu/coraal/supplements). CORAAL Supplements acts as a vehicle to help preserve and share other unique materials that do not fit the matrix criteria for CORAAL components, with the aim to promote greater public availability of diverse AAL data. We hope that CORAAL can help to promote the wider use of these resources and believe they can make valuable contributions to the overall coverage available in CORAAL.

For collections designed specifically for CORAAL, including the major fieldwork project led by Minnie Quartey in Washington, DC, for CORAAL:DCB but also data collected for a supplemental component for an urban Southern friend network in Atlanta (CORAAL:ATL), the corpus development team worked with fieldworkers to define demographic categories of targeted interest and interview protocols. For CORAAL:DCB, Quartey based her fieldwork recruiting on the matrix in table 14.1 and the available data for CORAAL:DCA. With several connections to the local community, interviews were done through friend of a friend networks. The interview schedule for CORAAL:DCB was designed to touch on several of the same general themes and topics as the CORAAL:DCA interviews did, while placing greater emphasis on engaging the participants in topics of interest to them that might promote conversational interactions. The six general categories of the interview schedule include general demographic information, neighborhood, family information, school days, friendship group, and work/occupation. A final category focused on DC-specific questions developed in consultation with our fieldworker and her primary interests.

Fieldworkers completed interview report forms for each sociolinguistic interview and administered informed consent documents for each participant. (All fieldwork

and consent documents were designed in consultation with and approved by the overseeing university's—the University of Oregon's—human subjects research office.) Because the goal was intentionally to collect data that would be shared as widely as possible, consent forms explicitly asked the participants if they would allow us to use the recording for a public corpus on language diversity in the United States. Consent forms have four levels of permission from whether we can use the recording as researchers to the explicit recognition of the participant by name in the corpus. Once recordings were collected, the fieldworkers transmitted the files to the development team, at the Language Variation and Computation Laboratory (LVC Lab) in the Department of Linguistics at the University of Oregon.

3.2 Processing

Once the data were on hand in the LVC Lab, files were stored on a networked drive in the LVC Lab hosted by a Macintosh Pro desktop computer and backups were kept locally on an external hard drive in the LVC Lab. Each audio file was resaved with a unique code, following the naming conventions for SLAAP. Audio files and basic metadata were then uploaded to SLAAP, so the files could be accessed remotely by project members and to act as an off-site archival version for the original recordings.

3.2.1 Transcription Each audio file was initially transcribed in Praat using Praat's TextGrid annotation features by an undergraduate research assistant. Our transcription process was designed to recognize that transcription is a process and one that does not yield a single, "correct" outcome (Bucholtz 2007; Du Bois et al. 1993; Edwards 2001; Kendall 2008, 2011; Mishler 1991; Ochs 1979). To quote Edwards (2001:321), "transcripts are not unbiased representations of the data. Far from being exhaustive and objective, they are inherently selective and interpretive. The researcher chooses what types of information to preserve, which descriptive categories to use, and how to display the information in the written and spatial medium of a transcript." The challenge for public corpora, especially for public "unconventional corpora" (see Kendall 2011), is to make choices about the transcription process that support diverse use cases and are not overly focused around a single research framework or perspective. At some level, this is an impossible task. For instance, our convention of delimiting the speech in transcripts into utterance-based units (with pauses of a certain length as the boundaries; based on Kendall 2007) supports a number of different conceptions of spoken language data, but leads to different chunking than a conception based on syntactic units (e.g., clauses) or intonational groups (see, for instance, Chafe 1993) and therefore could hinder certain analytic uses. No single transcription convention can support every possible use. To this end, transcript conventions for CORAAL are laid out explicitly in the *User Guide*, giving users access to basically the same information used by the transcribers so that users have the ability to see our explicit decisions. We also chose not to implement certain conventions whatsoever in the transcripts. For example, we did not include quotation marks in the current versions of transcripts. While some cases of direct quotation are easy and uncontroversial to identify, other cases are highly subjective or come down to an analyst's conception of quotative marking (see Romaine & Lange 1991; Buchstaller 2006). Rather than risk arbitrary quotation marking we opted to forego all use of quotation marks to mark quotations.

After this initial round of transcription, subsequently each file went through at least two rounds of checking and editing. The first round involved editing primarily by the second author and was focused largely on the accuracy of the text and boundaries, and the second round of editing was to check for spelling convention consistency.

Transcription did not cover the entirety of each recording but focused on providing a forty-five to sixty minute sample of the conversational portion of the interview when possible. Audio for each file was trimmed to include only the transcribed portions. In a few cases, we have excised portions of interviews. This is occasionally by the request of the participant, and it is occasionally done as a decision of the project team based on the content of the interview. There are a few cases where we have excised content even though the participant gave us permission to include it. For example, in DCB_se1_ag4_f_01, a passage from 537.8 to 671.8 is excised because of a graphic personal story. Other excised content includes reading passages and word lists. In CORAAL:DCA, a majority of the recordings have other tasks (e.g., language games, described in Fasold 1972), while CORAAL:DCB includes reading passage data (including the "Please Call Stella" passage; Weinberger 2015). The CORAAL:DCB "Please Call Stella"

passages are scheduled to be published in an upcoming release.

3.2.2 File naming For the corpus publication, files were renamed from their SLAAP names for greater transparency to end users (for a discussion of data management practices, including file naming, see Mattern, chapter 5, this volume). Speaker and file names are labeled systematically based on speaker demographics. For example, *DCA_se2_ag1_m_05_1.wav* is an audio (WAV) file for DCA, the Washington, DC 1968 component of CORAAL. The file's primary speaker is in socioeconomic group 2 (se2), age group 1 (ag1; this is the youngest age group, see section 3.1), male (m) number 5 (i.e., the fifth speaker in the cell of the demographic matrix). The final 1 indicates the audio file number. For supplements that do not stratify the sample by socioeconomics (including CORAAL:PRV and CORAAL:ROC), se0 is used to notate that a speaker is uncategorized for socioeconomic group (not that the speaker is in group 0). For gender, three codes are used: f, for female; m, for male; and n, for non-binary.

3.2.3 Redaction A guiding principle in the development of CORAAL is to protect the anonymity of its participants. Participants who were interviewed specifically for the corpus project (e.g., for CORAAL:DCB) were given the choice in the consent process of whether they wished to be recognized by name, with the default being that they will not be named. The majority of participants did ask to be recognized by name and we acknowledge them by name in the *User Guide*. For participants not interviewed by the project team (e.g., DCA, PRV, ROC, and some upcoming supplements), we do not disclose any names, unless there is an explicit (i.e., documented) permission given by the participant.

Our redaction process involved several steps. During the first round of transcription, transcribers marked different categories of sensitive information, such as names, street addresses, places of work, and other kinds of personally identifiable information. Additionally, transcribers noted the numbers of syllables of the item(s) to be redacted. The third round of transcription involved the creation of a redaction tier in Praat, where boundaries were placed directly around the portion of the interview to be redacted. The amount of material redacted varies widely by interview. Some interviews have only one or two redacted utterances while others have a great many. Once completed, redaction "bleeps," which were generated to match the mean pitch and amplitude of the speech being redacted, replaced the sensitive information.

After the completion of all processing (transcription, file naming, redaction), the Praat TextGrids were automatically processed (by script) into tab-delimited text files and (by ELAN) into ELAN format files. All three formats are available for download.

3.3 Metadata files

Each component of CORAAL has its own metadata file that contains a range of information about the recordings and their speakers. These files are tab-delimited text files that can be readily opened in a spreadsheet program, such Microsoft Excel, or in R. The metadata files are downloadable with the rest of each corpus component's files. For example, metadata for CORAAL:DCA are in the file labeled DCA_metadata_2018.10.06.txt (for version 2018.10.06). This file can be accessed from http://lingtools.uoregon.edu/coraal/dca/2018.10.06/DCA_metadata_2018.10.06.txt.

All of the pretrimmed files are stored (in WAV format) on SLAAP. SLAAP often contains more files from a sociolinguistic fieldwork project than just those included in CORAAL. The SLAAP codes for the files are provided in a column the metadata file when appropriate and described in the CORAAL *User Guide*. For example, CORAAL:DCA SLAAP codes are reflective of the original codes used in Fasold (1972). In CORAAL:DCB and CORAAL:ROC, SLAAP codes were given to recordings by the date of completion. For CORAAL:PRV, SLAAP codes are reflective of the codes used when the audio tapes were digitized and uploaded to SLAAP in 2007 and 2008. While SLAAP access to the files is limited, this can facilitate the possibility of tracing a CORAAL file back to its original recording.

Several categories in the metadata spreadsheets apply to all CORAAL components, while others apply only to specific components. Metadata explanations and notes are available in the CORAAL *User Guide*. Metadata can vary across the components, but in many cases extremely rich information is available about each speaker. Most speaker files obtained from Ralph Fasold (DCA) came with an informant data sheet, where much of the CORAAL metadata information comes from. The informant data sheet collected basic demographic data, such as sex, age, address,

birthplace, parents' birthplace, as well as a detailed socioeconomic status determination. For CORAAL:DCB, interviewers were asked to complete a similar interview report form for each speaker, which collected similar kinds of demographic information as Fasold's informant data sheet. In addition to general demographic information, the interview report form contains additional interview notes (e.g., explanations about interruptions and background noise) as well as topics covered over the course of the interview. For CORAAL:PRV, some speaker information was gathered from the metadata on SLAAP, while other information was obtained from the content of the interviews themselves.

3.4 Sharing

The main CORAAL web page (https://oraal.uoregon.edu/coraal/) is housed on the Online Resources for African American Language (https://oraal.uoregon.edu) website, which is devoted to providing information and resources about language in the African American community, targeting educators, researchers, and the wider public. (ORAAL is a Drupal site, hosted as a part of the University of Oregon's Drupal services.) CORAAL data are hosted on and provided to the public via a virtual server (http://lingtools.uoregon.edu) managed by the University of Oregon's College of Arts and Sciences information technology services group. Originally (i.e., when CORAAL was first publicly released in January 2018), the data were provided in downloadable form only. That is, users could access the data by downloading a series of compressed file bundles (tar.gz format). Making the data available by download, but not including a web-based interface initially, was a design decision, based on our belief that truly public data means that the data themselves are public, and not simply that there are open interfaces to the data. Our first priority was ensuring that potential users had access to the entirety of the data. Following this initial release, we developed a set of webpages, the CORAAL Explorer site (http://lingtools.uoregon.edu/coraal/explorer/), which provide access to the individual interviews and to a search interface for the entire collection. We anticipate further developing the online tools over time, although we intend for the data to *always* be the main emphasis of the CORAAL project. Through the CORAAL Explorer site, users can also download and use CORAAL transcripts directly in the R programming environment.

3.5 Citation and attribution for CORAAL

CORAAL seeks to follow and support the *Austin Principles of Data Citation in Linguistics* (http://site.uit.no/linguisticsdatacitation/; see also Conzett & De Smedt, chapter 11, this volume). As such, the CORAAL *User Guide* provides suggested citations and version numbering for each individual component of CORAAL, as well for the main CORAAL project (Kendall & Farrington 2020a) and the umbrella ORAAL project (Kendall, McLarty, & Farrington 2020), which houses the corpus. Furthermore, individual files and speakers are given persistent and relatively transparent labels that can be used for uniquely referring to individual files (and further with resolution down to individual transcript line numbers or time stamps). This can aid in using the corpus to provide specific passages or examples from the discourse or of AAL features. For instance, (1) provides an example from CORAAL of negative inversion, a morphosyntactic feature of AAL.

(1) Negative Inversion in AAL: "Cause didn't nobody hardly stay home then."
DCA_se1_ag1_m_03_1 (2084.34–2086.31)

We can further provide a direct link to the transcript line and audio in the CORAAL Explorer website for the corpus: http://lingtools.uoregon.edu/coraal/explorer/browse.php?what=DCA_se1_ag1_m_03_1.txt&line=2006&settime=2084.34. One issue with this current approach is that the URLs are not persistent identifiers (that is, future updates could impact the integrity of the links; see Conzett & De Smedt, chapter 11, this volume).

To protect the privacy of the speakers in the corpus, individual speakers and files are only labeled with their CORAAL identifier, and no reference is made back to the speakers' actual names (see redaction discussion in section 3.2.3). However, we also wanted to acknowledge the contributions of the individual participants and many participants did not wish to remain anonymous. Thus, the *User Guide* explicitly acknowledges by name all of the participants who indicated in their informed consent process that they wished to *not* remain anonymous.

4 CORAAL now and in the future

As mentioned earlier, the first components of CORAAL were released publicly in January 2018. Now that there are several published components included in CORAAL, the

development team's efforts center on adding to current components and revising what has already been published when errors or inconsistencies are discovered. Despite our close attention in the transcription creation and editing process, we regularly discover cases where our transcription practices have not been implemented as consistently as they could have been. (For instance, in v.2018.10.06, all instances of 'outta' and 'kinda' were changed to "out of" and "kind of" for internal consistency across each of the components.) The fact that we continue to find ways to improve CORAAL's transcripts does not surprise us, and, we hope, it will not disappoint CORAAL's users. We are also happy to receive from users corrections to mistranscribed elements. Sometimes users identify errors although at other times transcribers/listeners will just disagree about what they hear (Bucholtz 2000). Altogether, we anticipate periodically releasing corrections and we publish these in errata sections of the *User Guide*.

The decision to publish changes to transcriptions with periodic updates to the corpus means that different versions are in circulation. With the CORAAL Explorer website, updates to the corpus are also updated immediately on the Explorer website. In the metadata for each component, information about when each file was first added to CORAAL is included as well as the date of the most recent update for each file. We hope that whenever there is an update to the corpus, users will download the new update, although we cannot ensure that users do this. How to best manage versioning in a living, public corpus remains a question that we suspect we will be tackling for some time. We hope to improve our practices in future versions, for instance, by implementing a more formalized versioning system for the corpus, such as through Zenodo, GitHub, and/or other platforms.

5 Conclusion

CORAAL seeks to fill a major gap in linguistic research infrastructure by providing freely available sociolinguistic data sets for regionally situated AAL samples. The corpus is the first of its kind. Sociolinguistically oriented spoken language data raise a number of issues for corpus building, including sampling, metadata, and annotation (Beal, Corrigan, & Moisl 2007a, 2007b; Kendall 2007, 2008, 2011; Poplack 1989; Yaeger-Dror & Cieri 2014). The CORAAL development team has attempted to build as widely useful a resource as possible and hope to have in place a distribution and data archiving plan that will prove robust over both the long and short terms. We hope the resources, as well as the lessons we have learned, are useful to the wider linguistics research and outreach community.

Notes

1. CORAAL, pronounced [ˈkoɹəl], is part of the Online Resources for African American Language (ORAAL) project at the University of Oregon. CORAAL and the larger ORAAL Project have been made possible by support from the U.S. National Science Foundation (grant no. BCS-1358724), by the University of Oregon, and by the contributions of many people. In addition to the authors, the main CORAAL development team has included Jason McLarty, Shelby Arnson, and Brooke Josler. Lucas Jensen, Emma Mullen, Chloe Tacata, Jaidan McLean, Deepika Viswanath, Savanah Ray, and Matthew Bauer have also contributed to the corpus transcription, annotation, and redaction. Fieldwork would not have been possible without the major contributions of Minnie Quartey, Carlos Huff, Patrick Slay Brooks, Sharese King, and Ryan Rowe. We also thank Ralph Fasold, Natalie Schilling, Charlotte Vaughn, Walt Wolfram, and Danica Cullinan for their many contributions to the project. In the CORAAL *User Guide* (Kendall & Farrington 2020b), we express our deep gratitude to the main individuals who contributed their speech to the corpus, as well as many additional colleagues who have supported the project in various ways. Please see that document for complete acknowledgments.

2. Note that the authors are not arguing that all data sets need to be shared or that data should never be kept private. There are many reasons for sociolinguistic data sets not to be shared (see Warner 2014; Holton, Leonard, & Pulsifer, chapter 4, this volume), and a half century of productive work has demonstrated that vast discoveries can be made from relatively small, private data collections. Nonetheless, our point is that new advances and better science can be done if there are *more* public and larger data sets.

References

Baugh, John. 2000. *Black Street Speech: Its History, Structure, and Survival*. Austin: University of Texas Press.

Baugh, John. 2005. Linguistic profiling. In *Black Linguistics: Language, Society, and Politics in Africa and the Americas*, ed. Arnetha Ball, Geneva Smitherman, and Arthur K. Spears, 155–168. London: Routledge.

Beal, Joan, Karen Corrigan, and Hermann Moisl, eds. 2007a. *Creating and Digitizing Language Corpora*. Vol. 1: *Synchronic Databases*. New York: Palgrave-Macmillan. http://doi.org/10.1057/9780230223936.

Beal, Joan, Karen Corrigan, and Hermann Moisl. 2007b. Taming digital voices and texts: Models and methods for handling unconventional synchronic corpora. In *Creating and Digitizing Language Corpora*. Vol. 1: *Synchronic Databases*, ed. Joan Beal, Karen Corrigan, and Hermann Moisl, 1–16. New York: Palgrave-Macmillan. https://doi.org/10.1057/9780230223936_1.

Berez-Kroeker, Andrea, Lauren Gawne, Susan Smythe Kung, Barbara Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Boersma, Paul, and David Weenink. 2018. *Praat: Doing Phonetics by Computer* (computer program). Version 6.0.43. http://www.praat.org.

Bucholtz, Mary. 2000. The politics of transcription. *Journal of Pragmatics* 32: 1439–1465. https://doi.org/10.1016/S0378-2166(99)00094-6.

Bucholtz, Mary. 2007. Variation in transcription. *Discourse Studies* 9 (6): 784–808. https://doi.org/10.1177/1461445607082580.

Buchstaller, Isabelle. 2006. Diagnostics of age-graded linguistic behaviour: The case of the quotative system. *Journal of Sociolinguistics* 10:3–30. https://doi.org/10.1111/j.1360-6441.2006.00315.x.

Chafe, Wallace. 1993. Prosodic and functional units of language. In *Talking Data: Transcription and Coding in Discourse Research*, ed. Jane Edwards and Martin Lampert, 33–43. Hillsdale, NJ: Lawrence Erlbaum.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In *Talking Data: Transcription and Coding in Discourse Research*, ed. Jane Edwards and Martin Lampert, 45–89. Hillsdale, NJ: Lawrence Erlbaum.

Edwards, Jane. 2001. The transcription of discourse. In *Handbook of Discourse Analysis*, ed. Deborah Tannen, Deborah Schiffrin, and Heidi Hamilton, 321–348. Malden, MA: Blackwell. https://doi.org/10.1002/9780470753460.ch18.

Farrington, Charlie, Tyler Kendall, Patrick Slay Brooks, Lucas Jenson, Chloe Tacata, and Jaidan McLean. 2020. *The Corpus of Regional African American Language: ATL (Atlanta, GA 2017)*. Version 2020.05. Eugene, OR: Online Resources for African American Language Project.

Farrington, Charlie, and Natalie Schilling. 2019. Contextualizing the Corpus of Regional African American Language, D.C.: AAL in the nation's capital. *American Speech* 94 (1): 21–35. https://doi.org/10.1215/00031283-7308060.

Fasold, Ralph W. 1972. *Tense Marking in Black English*. Arlington, VA: Center for Applied Linguistics.

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

Kendall, Tyler. 2008. On the history and future of sociolinguistic data. *Linguistic and Language Compass* 2:332–351. https://doi.org/10.1111/j.1749-818X.2008.00051.x.

Kendall, Tyler. 2011. Corpora from a sociolinguistic perspective (Corpora sob uma perspectiva sociolinguística). In *Corpus Studies: Future Directions*, ed. Stefan Th. Gries, special issue of *Revista Brasileira de Linguística Aplicada* 11 (2): 361–389. http://dx.doi.org/10.1590/S1984-63982011000200005.

Kendall, Tyler, Joan Bresnan, and Gerard Van Herk. 2011. The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic data sets. *Corpus Linguistics and Linguistic Theory*, 7 (2): 229–244. https://doi.org/10.1515/cllt.2011.011.

Kendall, Tyler, and Charlie Farrington. 2020a. *The Corpus of Regional African American Language*. Version 2020.05. Eugene, OR: The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal.

Kendall, Tyler, and Charlie Farrington. 2020b. *The Corpus of Regional African American Language User Guide*. Version 2020.05. Eugene, OR: The Online Resources for African American Language Project. http://lingtools.uoregon.edu/coraal/userguide.

Kendall, Tyler, Ralph Fasold, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler. 2018. *The Corpus of Regional African American Language: DCA (Washington DC 1968)*. Version 2018.10.06. Eugene, OR: The Online Resources for African American Language Project.

Kendall, Tyler, Jason McLarty, and Charlie Farrington. 2020. *ORAAL: Online Resources for African American Language*. Eugene, OR: Online Resources for African American Language Project. https://oraal.uoregon.edu/.

Kendall, Tyler, Minnie Quartey, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler. 2018. *The Corpus of Regional African American Language: DCB (Washington DC 2016)*. Version 2018.10.06. Eugene, OR: The Online Resources for African American Language Project.

King, Sharese. 2018. Exploring social and linguistic diversity across African Americans from Rochester, New York. PhD dissertation, Stanford University.

King, Sharese, Charlie Farrington, Tyler Kendall, Emma Mullen, Shelby Arnson, and Lucas Jenson. 2020. *The Corpus of Regional African American Language: ROC (Rochester, NY 2016)*. Version 2020.05. Eugene, OR: The Online Resources for African American Language Project.

Labov, William, Paul Cohen, Clarence Robins, and John Lewis. 1968. *A Study of the Non-Standard English of Negro and Puerto Rican Speakers in New York City*. Final Report, Research Project 3288. Washington, DC: United States Office of Education.

Lanehart, Sonja. 2015. Language use in African American communities: An introduction. In *The Oxford Handbook of*

African American Language, ed. Sonja Lanehart, 1–19. Oxford: Oxford University Press. https://dx.doi.org/10.1093/oxfordhb/9780199795390.001.0001.

Mishler, Elliot. 1991. Representing discourse: The rhetoric of transcription. *Journal of Narrative and Life History* 1 (4): 255–280. https://doi.org/10.1075/jnlh.1.4.01rep.

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

Poplack, Shana. 1989. The care and handling of a megacorpus: The Ottawa-Hull French Project. In *Language Change and Variation*, ed. Ralph W. Fasold and Deborah Schiffrin, 411–444. Amsterdam: John Benjamins.

Rickford, John R. 1999. *African American English: Features, Evolution, and Educational Implications*. Malden, MA: Blackwell.

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

Rowe, Ryan. 2005. The development of African American English in the oldest Black town in America: Plural *-s* absence in Princeville, North Carolina. MA thesis, North Carolina State University.

Rowe, Ryan, Walt Wolfram, Tyler Kendall, Charlie Farrington, and Brooke Josler. 2018. *The Corpus of Regional African American Language: PRV (Princeville, NC 2004)*. Version 2018.10.06. Eugene, OR: The Online Resources for African American Language.

Warner, Natasha. 2014. Sharing of data as it relates to human subjects issues and data management plans. *Language and Linguistics Compass* 8 (11): 512–518. https://doi.org/10.1111/lnc3.12107.

Weinberger, Steven. 2015. *The Speech Accent Archive* (website). George Mason University. http://accent.gmu.edu/.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559. http://hdl.handle.net/11858/00-001M-0000-0013-1E7E-4.

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

Wolfram, Walt. 2008. Language diversity and the public interest. In *Sustaining Linguistic Diversity: Endangered and Minority Languages and Language Varieties*, ed. Kendall A. King, Natalie Schilling-Estes, Lyn Fogle, Jia Jackie Lou, and Barbara Soukup, 187–204. Washington, DC: Georgetown University Press.

Yaeger-Dror, Malcah, and Chris Cieri, eds. 2014. *Special Issue on Archiving Sociolinguistic Data*. *Language and Linguistics Compass* 8 (3).

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

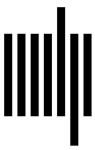
The Open Handbook of Linguistic Data Management

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 open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

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