CORAAL Corpus of Regional African American Language

Editors: Tyler Kendall & Charlie Farrington (University of Oregon)

CORAAL (Version 2018.10.06)

↳ CORAAL:DC

↳ CORAAL:DCA (Washington, DC 1968; Version 2018.10.06)

↳ CORAAL:DCB (Washington, DC 2016; Version 2018.10.06)

↳ CORAAL:PRV (Princeville, NC 2004; Version 2018.10.06)

↳ CORAAL:ROC (Rochester, NY 2016; Version 2018.10.06)

CORAAL User Guide (as of October 6, 2018)

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1358724.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

About the Corpus of Regional African American Language

The Corpus of Regional African American Language (CORAAL) is the first public corpus of African American Language (AAL) data. CORAAL features recorded speech from regional varieties of AAL and includes the audio recordings along with time-aligned orthographic transcription.

CORAAL is a long-term corpus-building project conceived of in terms of several components. 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 (e.g. Fasold 1972). CORAAL:DC, first released in January 2018, is comprised of over 100 sociolinguistic interviews with AAL speakers in DC born between 1890 and 2005. CORAAL:DC consists of two sub-components, CORAAL:DCA and CORAAL:DCB. In addition to CORAAL:DC, CORAAL is scheduled to increasingly include several smaller components to provide regional breadth. The two supplemental components included as of October 2018 include CORAAL:PRV, which includes 15 sociolinguistic interviews from a rural African American community in central North Carolina, as well as CORAAL:ROC, which includes 13 sociolinguistic interviews from Rochester, a city in Western Upstate New York.

All CORAAL recordings have been 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, ELAN (.eaf) files, and as plain text (.txt) files with tab-delimited fields.

The CORAAL team plans to release updates approximately quarterly, with the next release in Winter 2019. This update plans to include additional speakers for CORAAL:DCB and to release our third supplemental component, 13 sociolinguistic interviews from Atlanta, Georgia. Additionally, a syntactically parsed version for a portion of the CORAAL data is also under development, which will include disfluency coding, part-of-speech tagging, and syntactic annotation for approximately one million words. This is in progress and should be available by 2019.

Why AAL?

AAL (often referred to as African American English, AAE, or African American Vernacular English, AAVE) has been a central object of study in North American linguistics and especially sociolinguistics for over 50 years (e.g., Labov, Cohen, Robins, & Lewis 1968; Wolfram 1969; Labov 1969, 1972; Fasold 1972; Bailey, Maynor, & Cukor-Avila 1991; Mufwene, Rickford, Bailey, & Baugh 1998; Rickford 1999; Poplack & Tagliamonte 2001; Green 2002; Wolfram & Thomas 2002; Yaeger-Dror & Thomas 2010; Rickford, Sweetland, Rickford, & Grano 2012; Lanehart 2015). Already by the 1990s, AAL was described as having inspired more than five times as many sociolinguistic publications as any other ethnic or regional dialect (Schneider 1996:3).

From this extensive work, much is known about many structures of AAL varieties and a large body of research has investigated its origins (e.g., Kurath 1949; McDavid & McDavid 1951; Stewart 1968; Bailey et al. 1991; Poplack & Tagliamonte 2001; Wolfram & Thomas 2002) and current trajectories of change (e.g., Bailey & Maynor 1985, 1987; Labov 1987, 1998; Cukor-Avila 1995, 2001; Dayton 1996; Wolfram & Thomas 2002; Yaeger-Dror & Thomas 2010). Yet, there remain important questions about the origin of these varieties, their current and future development, and their relationship(s) to regional European American and other socioethnic varieties. There also continue to be a range of important social and educational applications of enhanced knowledge about the nature of AAL.

At the same time that AAL has been so extensively studied, it remains massively underrepresented in terms of publicly available datasets and in terms of its use in general linguistic theory building (Green 2002; Kendall, Bresnan, & Van Herk 2011). Sociolinguists (and the field of linguistics more generally; cf. Berez-Kroeker, Holton, Kung, & Pulsifer 2017) have increasingly adopted models of data compilation in recent years that include data sharing and promoting data re-use, but thus far almost all AAL data remain unavailable for wider, public sharing, due to ethical considerations or limitations from how the data were collected (e.g. participant consent; Warner 2014).

The availability of a public corpus of AAL is meant to enable new research and new uses. It provides access to primary data for a wider range of scholars, for example those who do not have access to field sites or to sociolinguistic data themselves (such as educational professionals and graduate students). It also seeks to support new "open science"-based approaches, where direct testing of competing theories or methodologies or reanalysis (cf. Rickford, Ball, Blake, Jackson, & Martin 1991; Kendall 2011) can be made on the same data.

The Online Resources for African American Language (ORAAL) Project

CORAAL is a publication of our larger, umbrella project, the Online Resources for African American Language (ORAAL) Project, housed at the Language Variation and Computation (LVC) Laboratory in the Department of Linguistics at the University of Oregon. The ORAAL website (**http://oraal.uoregon.edu/**; Kendall and McLarty 2018) seeks to act as a central web-based source for research and educational information about AAL. Please visit the ORAAL website for more general information about AAL, and to obtain CORAAL or find out about future updates.

Acknowledgments and Development Team

The Corpus of Regional African American Language (CORAAL) is part of the Online Resources for African American Language (ORAAL) Project at the University of Oregon, U.S.A. CORAAL and the larger ORAAL Project have been made possible by support from the U.S. National Science Foundation (Grant No. BCS-1358724 "Enhancing data and tools for research and education on African American English"), by the University of Oregon, and by the contributions of many people.

The ORAAL Project is centered in the Language Variation and Computation (LVC) Lab in the Department of Linguistics at the University of Oregon. The LVC Lab, directed by Tyler Kendall, is also home to other linguistic resources, including the NORM Vowel Normalization and Plotting Suite and Vowels.R Package (see http://lingtools.uoregon.edu/norm/; Thomas and Kendall 2007) and has developed the Sociolinguistic Archive and Analysis Project (SLAAP; Kendall 2007a, 2008), which houses sociolinguistic datasets through its website, https://slaap.chass.ncsu.edu/, hosted at North Carolina State University.

The main CORAAL development team over the years has consisted of Tyler Kendall, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler. Lucas Jensen, Emma Mullen, 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.

We are especially, and deeply, grateful to the very many individuals who participated in the sociolinguistic interviews for the corpus and the projects upon which it is based. Obviously, their generous contributions of their voices, their words, and their time, are the foundation of this project. Specifically, we thank:

Vernial Batts
Lamonte Belk
Devon Bennett
Darren Black
Jason Black
Nicholas Black
Niya Black
Michelle Broadus
Sheila Brockington
Robert Brown
Robin Brown
Marco Coleman
Yolanda Coppedge
Shanquette Dannah
Janet Davenport
Andrea Davis
Renee Edelin
R'Mani Fitchett
Devonte Gooding
Michelle Graham
Theressa Green
Ayanna Holmes
Domonique Inniss
Alonte Jenkins
Terri Johnson
Angelo Johnson
Jazmine Jones
LeVar Jones
Linda Jones
Dwayne Lawson-Brown
Gary Lewis
Barbara Ligon
LaTanya Malloy
Keisha Matos-Joaquin
Rhonice Miles
Melvin Moore
Curtis Robinson
Tavis Saunders
Zymiah Speller
Andrea Talbert
Mylz Taylor
Ted Thomas
Shirleen Thompson
Jamiris Tolbert
Devin Turner
Monique Van Buren
Tiffany Woodberry-Black

We also thank the very many other, anonymous participants who also contributed their speech and their stories to the corpus. (Participants were given the choice of whether they wished to be recognized by name or to remain anonymous – thus we only recognize by name those participants who asked not to remain anonymous; see discussion of redaction in this document.) A number of students and research assistants at the University of Oregon have contributed to the project. In addition to the research assistants mentioned above, we thank students in T. Kendall's Spring 2016 Seminar on African American English. Finally, we thank colleagues who have acted as beta-testers and consultants as we have developed the corpus, including: Tricia Cukor-Avila, Jon Forrest, Jessi Grieser, Chris Hall, Nicole Holliday, Taylor Jones, Sharese King, John Rickford, and Tracey Weldon.

Contact Information

Please contact the CORAAL development team via:
Email: corpusofregionalAAL@gmail.com
Twitter: @CorpusAAL

You can also write the editors:
Dr. Tyler Kendall: tsk _at_ uoregon _dot_ edu

Charlie Farrington: crf _at_ uoregon _dot_ edu

Department of Linguistics
1290 University of Oregon
Eugene, OR 97403-1290 USA

About Version 2018.10.06 (Change Log)

Version 2018.10.06 is the third release of CORAAL.

Through a round of close consistency checking and editing, changes were made to every TextGrid file previously available (in v. 2018.04.06), including removing spaces that sometimes occurred at the beginning and ends of utterance as well as fixing typos (e.g. "<unintellgible>" changed to "<unintelligible>" for DCA_se2_ag1_m_03_1). This also means that all text files and ELAN files have been changed as well. Some inconsistencies in transcripts were fixed to more closely align to the transcription conventions (see that section). This affects all files in DCA, DCB, and PRV. Additionally, audio file processing has been re-run on ten files listed in the errata below (redaction, amplitude normalization, and conversion). Due to the increased consistency across files, we recommend that all users replace prior versions of all CORAAL data with v. 2018.10.06 to have the most up-to-date transcriptions and audio.

CORAAL:DCB v. 2018.10.06 includes one additional speaker (five new files).

CORAAL:ROC, the second sub-component, has been added. We expect to add two speakers to this component in the next release.

Additionally, with this release comes the first version of a web-interface to CORAAL, the **CORAAL explorer website** (http://lingtools.uoregon.edu/coraal/explorer/) and R functions to work with CORAAL transcripts directly in R (http://lingtools.uoregon.edu/coraal/explorer/R/CORAAL_web.R).

Errata (v. 2018.10.06)

- Audio re-processing
  - DCA_se1_ag2_m_01_1, DCA_se1_ag3_f_01_1, DCA_se1_ag3_f_02_1, DCA_se1_ag4_m_01_1, DCA_se1_ag4_m_02_1, DCA_se3_ag2_f_04_1, DCA_se3_ag4_m_01_1, DCB_se2_ag3_m_03_1, DCB_se3_ag4_f_02_1, PRV_se0_ag2_m_02_1
- Numerous minor changes throughout all CORAAL transcripts (see change log above).

About Version 2018.04.06 (Change Log)

Version 2018.04.06 is the second release of CORAAL.

CORAAL:DCA data have not changed. However, updates were made to the metadata adding new columns to parallel changes to CORAAL:DCB (see just below) and to correct word counts listed for individual speakers.

CORAAL:DCB v.2018.04.06 includes ten additional speakers who were not included in the opening release. New audio files are available in the download tar.gz parts 11-14. Additionally, one transcript file from CORAAL:DCB has been updated with corrections (see

Errata below). Users will need to download those new audio tar.gz bundles as well as the new metadata file and new transcript tar.gz files to obtain the new data. (Audio files in parts 01-10 are unchanged so the content of those tar.gz parts are the same as for v.2018.01.06.) CORAAL:DCB is not yet complete, but we anticipate completing the sample for the Summer 2018 release. The DCB metadata spreadsheet now includes a Primary Speaker column (Primary.Spkr) and columns indicating when files were added and last modified. The audio file is named with the primary speaker's code, but when there are other interviewees present, they are listed as non-primary speakers.

CORAAL:PRV, the first sub-component, has been added. This component is expected to be relatively stable, with no additional speakers or audio files planned, though we anticipate amending transcription files to improve accuracy if/when errors are discovered. For details about the sub-corpus, see CORAAL:PRV section below. The PRV metadata spreadsheet also includes a Primary Speaker column. The audio file is named with the primary speaker's code, but when there are other interviewees present, they are listed as non-primary speakers.

Errata (v.2018.04.06)

- DCB_se2_ag3_m_02_1 (.txt, .TextGrid, .eaf)
    - Changed four instances of "this that and a third" to "this that and the third"

About Version 2018.01.06 (Change Log)

Version 2018.01.06 is the opening release of CORAAL.

CORAAL:DCA is expected to be relatively stable, with no additional speakers or audio files planned (although future updates will presumably amend transcription files to improve their accuracy).

CORAAL:DCB is not quite complete. CORAAL:DCB will have additional interviews released in a future update, with the goal of including at least two speakers per demographic cell (see CORAAL:DCB section below). We do not anticipate making substantial changes to the existing interviews/transcripts but we hope to complete the sample for the next update to the corpus.

Obtaining and Using CORAAL

What comes with the corpus?

CORAAL contains audio files along with corresponding orthographic transcription, time-aligned at the utterance level. Metadata files are available for each component that provide extensive information about the speakers (see metadata section below). Metadata files are plain text in a tab-delimited format; these can be read by any text editing software, but can also be loaded into spreadsheet software like MS Excel or software like R. Audio files are available in uncompressed .wav format (generally 44.1 kHz, 16 bit, mono). Transcripts are available in three formats – Praat TextGrids (.TextGrid), ELAN files (.eaf), and plain, tab-delimited text (.txt). All three formats contain identical information. As described in the transcription section, transcripts are created by the CORAAL development team directly as TextGrids in Praat. The Praat TextGrid files are automatically processed (by script) into text files and (by ELAN) into ELAN format.

Obtaining the corpus

All components of CORAAL are available for download from its home at **http://oraal.uoregon.edu/coraal**. (See the following subsections for other ways to access the corpus.) The corpus is organized by sub-component and, due to their large sizes, each sub-component is broken down into several parts, which are then compressed using the standard "tar" and "gnu zip" compression. Audio files are contained in numerous separate tar.gz files. Transcripts are organized into tar.gz files by file type. Metadata files are available as uncompressed plain text files as these have smaller file sizes. You can download the corpus by saving the individual parts over your web-browser. Use your operating standard decompression software to decompress the downloaded files. We suggest you move all of the files from their individual parts' folders to a single folder, as the individual part folders, especially for the audio parts, are not meaningful other than as ways to organize the files for downloading.

Please note that CORAAL is quite large (≈34 GB for v.2018.10.06) and can take a long time to download. DCA, in its compressed format for download, is 7.5 GB; DCB, compressed, is 8.1 GB; PRV, compressed, is 3.3 GB; and ROC, compressed is 2.6 GB.

Quicker way: Automate the download

A text file containing a list of all files for all of the components of the current version of CORAAL is available here:

https://tinyurl.com/coraalfiles
(This is a shortcut to http://lingtools.uoregon.edu/coraal/coraal_download_list.txt)

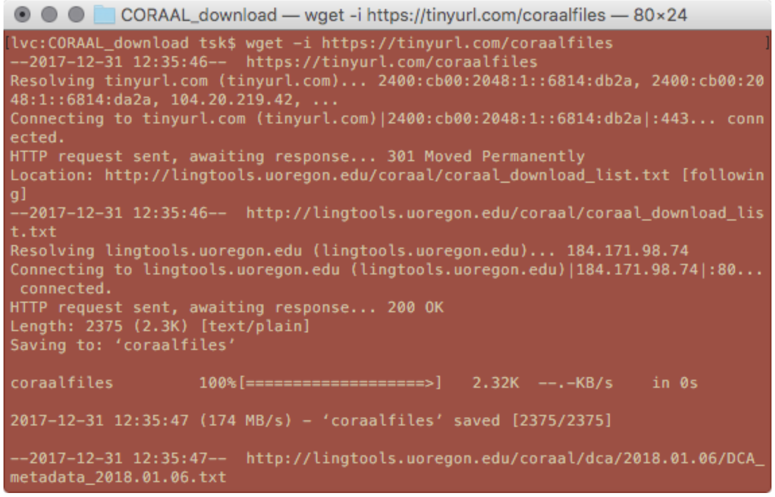
You can use this file to automate the downloading of the entire corpus, such as through one of the techniques described below. (It will probably still take a while but at least you can avoid manually downloading every file!) Here are some suggested ways you can automate this process:

Use wget software

An easy way to download all of the files is by using the popular open-source “wget” application. Download “wget” if you don’t already have it (see below). Then from your system’s shell (e.g. Mac Terminal application), navigate to the location you wish to save all of the files and run the following:

```
wget -i https://tinyurl.com/coraalfiles
```

This will download all of the files for the corpus. Here is a screenshot of a Mac Terminal after executing the wget command (the first line shows the wget command, everything below that is wget doing its magic):


CORAAL_download — wget -i https://tinyurl.com/coraalfiles — 80×24

```
[lvc:CORAAL_download tsk$ wget -i https://tinyurl.com/coraalfiles              ]
--2017-12-31 12:35:46--  https://tinyurl.com/coraalfiles
Resolving tinyurl.com (tinyurl.com)... 2400:cb00:2048:1::6814:db2a, 2400:cb00:20
48:1::6814:da2a, 104.20.219.42, ...
Connecting to tinyurl.com (tinyurl.com)|2400:cb00:2048:1::6814:db2a|:443... conn
ected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: http://lingtools.uoregon.edu/coraal/coraal_download_list.txt [followin
g]
--2017-12-31 12:35:46--  http://lingtools.uoregon.edu/coraal/coraal_download_lis
t.txt
Resolving lingtools.uoregon.edu (lingtools.uoregon.edu)... 184.171.98.74
Connecting to lingtools.uoregon.edu (lingtools.uoregon.edu)|184.171.98.74|:80...
 connected.
HTTP request sent, awaiting response... 200 OK
Length: 2375 (2.3K) [text/plain]
Saving to: ‘coraalfiles’

coraalfiles         100%[===================>]   2.32K  --.-KB/s    in 0s

2017-12-31 12:35:47 (174 MB/s) - ‘coraalfiles’ saved [2375/2375]

--2017-12-31 12:35:47--  http://lingtools.uoregon.edu/coraal/dca/2018.01.06/DCA_
metadata_2018.01.06.txt
```

You may find easy-to-install versions of “wget” for your operating system by searching the web for something like “download wget for mac”, but as of this writing the following links are helpful starters. For Mac and Linux, using a package manager (like “homebrew” on Mac, see https://brew.sh and note that wget is the example on the homebrew website!) is an easy way to install the software.

Mac OSX: https://coolestguidesontheplanet.com/install-and-configure-wget-on-os-x/ (or from source: http://osxdaily.com/2012/05/22/install-wget-mac-os-x/)

Windows: https://eternallybored.org/misc/wget/

Linux: e.g. https://www.tecmint.com/10-wget-command-examples-in-linux/

Use curl software

Many operating systems, like Mac OSX, have a program called “curl” already installed. “curl” can also be used to download the corpus quickly, although it cannot read a list of files as readily as “wget”. A fast way using “curl” involves also using a second program “xargs”. You can test to see if your system has “curl” and “xargs” by going to your system’s shell (e.g. Terminal on Mac) and executing:

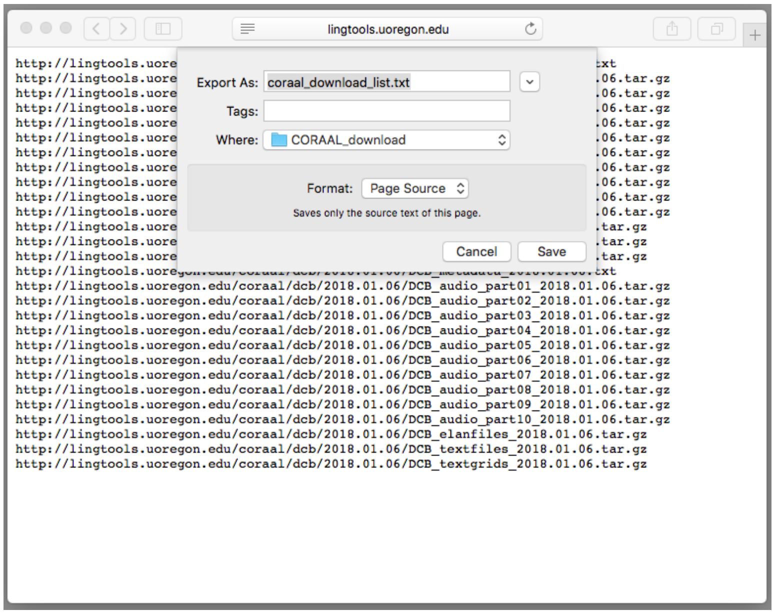
```
which curl
which xargs
```

If you have these programs, your system will respond by telling you where they are. If you don't, your system will simply show you another prompt.

Here, you will need to manually download the download_list.txt files before you can proceed. Go to https://tinyurl.com/coraalfiles and save this file to your computer as a text file (e.g. coraal_download_list.txt). Then, from your system's shell (e.g. Terminal on Mac), navigate to the location you wish to save all of the files and execute:

```
xargs -n 1 curl -O < /path/to/coraal_download_list.txt
```

This will download each of the files in the list. Here are screenshots of saving the coraal_download_list.txt file (for CORAAL version 2018.01.06) from a web browser:



Export As: coraal_download_list.txt

Tags:

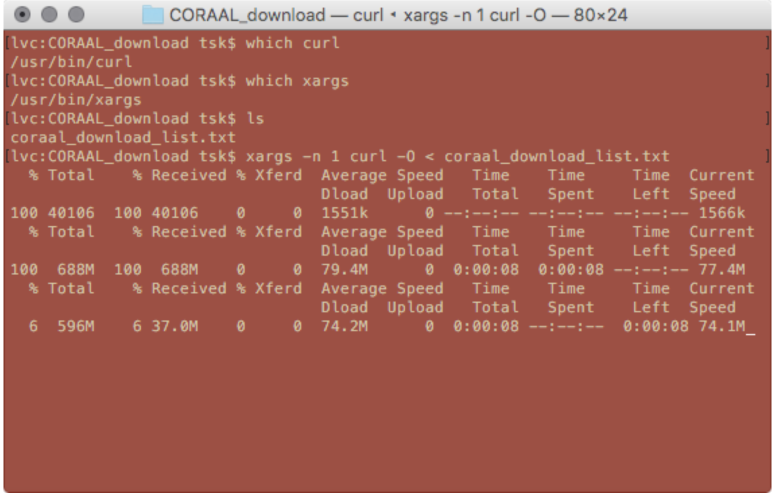
Where: CORAAL_download

Format: Page Source

Saves only the source text of this page.

```
http://lingtools.uore                                                      xt
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .06.tar.gz
http://lingtools.uore                                                      .tar.gz
http://lingtools.uore                                                      .tar.gz
http://lingtools.uore                                                      .tar.gz
http://lingtools.uore                                                      xt
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part01_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part02_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part03_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part04_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part05_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part06_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part07_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part08_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part09_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_audio_part10_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_elanfiles_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_textfiles_2018.01.06.tar.gz
http://lingtools.uoregon.edu/coraal/dcb/2018.01.06/DCB_textgrids_2018.01.06.tar.gz
```

And executing the xargs/curl commands in Mac Terminal (in this screenshot, the first two files have transferred and curl is 6% of the way through downloading the third file):



```
lvc:CORAAL_download tsk$ which curl
/usr/bin/curl
lvc:CORAAL_download tsk$ which xargs
/usr/bin/xargs
lvc:CORAAL_download tsk$ ls
coraal_download_list.txt
lvc:CORAAL_download tsk$ xargs -n 1 curl -O < coraal_download_list.txt
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 40106  100 40106    0     0  1551k      0 --:--:-- --:--:-- --:--:-- 1566k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  688M  100  688M    0     0  79.4M      0  0:00:08  0:00:08 --:--:-- 77.4M
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  6  596M    6 37.0M    0     0  74.2M      0  0:00:08 --:--:--  0:00:08 74.1M_
```

The CORAAL Explorer website and R functions

Beginning with CORAAL v. 2018.10.06, we have created a web-interface for “exploring” CORAAL. The CORAAL Explorer website is available at http://lingtools.uoregon.edu/coraal/explorer/. The online interface currently has two primary features, a browse page (http://lingtools.uoregon.edu/coraal/explorer/browse.php) and a search page (http://lingtools.uoregon.edu/coraal/explorer/search.php). The browse feature provides text and audio for each CORAAL file (and also allows users to download individual files). The search feature provides a web-based front end to a set of R functions for working with the corpus. The pages are (hopefully) relatively straightforward to use, but users can contact CORAAL developers with questions or problems. We anticipate that the web pages will continue to be developed and new versions will be released on a rolling basis (i.e. not at the same rate as the periodic CORAAL data updates).

An R script, which supports the direct downloading and creating of R data structures for CORAAL is available at http://lingtools.uoregon.edu/coraal/explorer/R/CORAAL_web.R. You can save that file to disk to use or you can simply load it into R over the internet. Execute the following to make CORAAL and some helper functions available in your R session:

```
source("http://lingtools.uoregon.edu/coraal/explorer/R/CORAAL_web.R")
```

This will show a short help message including instructions for how to download CORAAL directly into your R session (using the coraal.webget.data() function). Note that you need the RCurl library installed in R to use the web-based download feature in R.

Terms of use

The Corpus of Regional African American Language (CORAAL), its data, and websites are available for free, public use for research purposes. CORAAL is available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License. This means you are free to use and reuse the corpus for non-commercial purposes, but that you must cite the original corpus and any derivative versions of CORAAL you develop and wish to share with others must be distributed using the same license. A summary of the license is available on the Creative Commons website at https://creativecommons.org/licenses/by-nc-sa/4.0/ and the full license is available at https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.

For additional or other uses of the corpus, such as commercial purposes, please contact Tyler Kendall (tsk _at_ uoregon _dot_ edu) to discuss.

Citing the corpus

If you use the corpus, we ask that you cite the corpus. Below are suggested citations for CORAAL and its available subcomponents. More generally, we urge you to learn about the *Austin Principles of Data Citation*, which provide guidelines for citation and attribution of linguistic data. The Austin Principles are described at http://site.uit.no/linguisticsdatacitation/.

Recommended Citation and **Version Number** for the main CORAAL project:

- Kendall, Tyler and Charlie Farrington. 2018. *The Corpus of Regional African American Language.* Version 2018.10.06. Eugene, OR: The Online Resources for African American Language Project. [http://oraal.uoregon.edu/coraal]

Recommended Citations and **Version Number** for CORAAL:DCA (1968):

- 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.
- Fasold, Ralph. 1972. *Tense marking in Black English: A linguistic and social analysis.* Arlington, VA: Center for Applied Linguistics. [https://eric.ed.gov/?id=ED129065]

Recommended Citation and **Version Number** for CORAAL:DCB (2016):

- 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.

Recommended Citations and **Version Number** for CORAAL:PRV (2004):

- 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.
- 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. Raleigh: North Carolina State University. [http://repository.lib.ncsu.edu/handle/1840.16/711]

Recommended Citations and **Version Number** for CORAAL:ROC (2016):

- King, Sharese, Charlie Farrington, Tyler Kendall, Emma Mullen, Shelby Arnson, and Lucas Jenson. 2018. *The Corpus of Regional African American Language (Rochester, NY 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*. Ph.D. dissertation. Palo Alto, CA: Stanford University.

CORAAL Transcription

Transcription practices & conventions

This section outlines the transcription conventions used in the creation of CORAAL. Our practices, and thus parts of this section, have been adapted from those developed for the

Sociolinguistic Archive and Analysis Project (SLAAP), described in the SLAAP User Guide, Version 0.96, June 2009, available here: https://slaap.chass.ncsu.edu/userguide/.

Transcripts are created entirely in the Praat software (http://praat.org; Boersma and Weenink 2014) using TextGrid annotation objects. (Other transcription formats, like ELAN, are derived from these original TextGrid files.) In a TextGrid transcript, each speaker is represented in an interval tier. Occasionally an additional interval tier is used to transcribe speech by an interloper. In some cases, e.g. for some of the recordings in CORAAL:DCA the corpus development team had earlier transcript documents to work with, but even in these cases the entire transcript was re-keyed following the conventions outlined here. Earlier transcripts were referred to for spelling or for words that could not be determined by the transcription team from the audio.

Following SLAAP conventions, transcripts align text to speech at a per-utterance level, where utterances are defined as uninterrupted speech sounds by the same individual, with utterances delimited at pauses. For CORAAL, the criteria for how transcribers delimit utterances is set at a pause length of 60-70 milliseconds (see Kendall 2009, 2013).

For each audio file, three rounds of transcription and editing were completed. The first round of transcription was completed by one of two undergraduate research assistants, who were the primary transcribers between 2015 and 2017. The second round of transcription – a thorough editing of the TextGrid – was done by a graduate student in linguistics, who listened to the entire audio file while reviewing the transcript, making corrections where necessary. The third round created a Redaction (RD) tier which time-stamped portions of the transcript/audio that needed to be redacted, while also cleaning up any remaining inconsistencies. Despite our attempts to have maximally clean and accurate transcripts, we expect some degree of disagreement on transcript accuracy. Transcripts are always the (in process) product of individual analysis (Edwards 2001; Kendall 2008).

A primary goal of a time-aligned transcript is to act as a proxy to the original recording, to allow for easy searching and browsing of the recording. It is not to make an exact, textually accurate representation of the speech. Along these lines, orthographic transcription conventions use simple orthography and standard-like spelling. As a general rule, morphosyntactic variants (e.g. *was* for *were*) are transcribed, but phonological variants (such as *velar nasal fronting* (*in* for *ing*), *r*-lessness, and dialectal vowel qualities) are not.

At the same time, the transcript text attempts to accurately account for all the "noises" of speech, such as laughter, filled pauses (like "uh" or "um"), and restarts (e.g "I- I- I di- didn't mean to") as well as misspoken words (e.g. "/brack/ in the seventies"). Standard-like capitalization and punctuation is used, with the hyphen, -, used to indicated lexical and morphosyntactic restarts, as well as incomplete intonation. Silent pauses (of course) are not described or coded, as they are represented in all cases by empty intervals. Since a number of speech sounds (e.g. "Mm-hm") do not have codified spellings while some extremely common productions do have agreed upon non-standard written forms (e.g. "I'm'a"), transcription conventions attempt to standardize possible spellings.

The following subsections outline the specific transcription conventions. Part A provides information on symbols and punctuation used in the corpus. Part B provides orthographic conventions and examples for commonly encountered non-standard words and constructions. Part C gives conventions for disfluent speech, and Part D gives examples of other features encountered in CORAAL (including miscellaneous topics that came up for our transcription team).

A. Symbols and punctuation

Special symbols

- The basic symbols used in transcription follow the conventions used by SLAAP.

  […] contains overlapping speech, e.g.,

  ```
  Speaker A: So [I went-]
  Speaker B: [You did] what?
  ```

  / … / represents several categories, including
    - Inaudible or unintelligible speech
    - Redacted speech
    - Misspoken words

  < … > contains non-linguistic sounds, such as **`<cough>`**

  (…) contains line-level notes, such as **`(laughing)`**

- Standard punctuation is used for ease of transcription and readability. Punctuation includes periods (.), commas (,), question marks (?), exclamation points (!), all of which are used as normal in English (prosody). Apostrophes (') are used as they are generally used in English. Dashes (-) are limited to disfluent speech and compounds.
    - `I considered myself- Oh! Horseshoes! I played quite a bit of horseshoes at that age. (DCA_se3_ag3_m_01)`

- In initial versions of the corpus, quoted speech is not represented orthographically (e.g. by use of quotation marks). This may change in future versions, but determining what speech is direct vs. indirectly reported speech vs. other "quoted speech" is not trivial and highly interpretive.

B. Orthographic conventions

General Notes

- Spelling and capitalization follow standard English writing practices.
- All numbers are written out as complete words.
    - `I've been in Washington for twenty-two years. (DCA_se3_ag3_m_04)`
    - `When aught years (e.g. 2008) are referred to as '08, o eight is transcribed.`
- Transcribers write compounds as two words, or a hyphenated word.
- Abbreviations are avoided except for personal titles (e.g. Mr., Mrs., Dr.). Junior (e.g. Thomas Junior) is not abbreviated.
- Acronyms are written without punctuation (e.g. TV, DC). Letters that are pronounced are separated by dashes. ASAP is always transcribed in caps. An apostrophe is included for plural acronyms (e.g. TV's, A's).
    - `And I think his name is spelled B-O-O-N-K. (DCB_se3_ag1_f_01)`

- Question marks (?) are used in transcripts as normal in English, based on a combination of prosodic and syntactic cues.

Reduced forms

- In CORAAL, several reduced constructions have orthographic representations. Transcribers used this list when fuller (more conventionally standard forms) are *reduced*. Otherwise full forms are transcribed.
- In the case of *have* reduction, conventional contractions (e.g. must've, would've, etc.) are also transcribed when necessary.
- Some commonly reduced constructions (e.g. used to, kind of, sort of, out of) are never orthographically represented as reduced. In the following table, reduced forms are listed as transcribed in CORAAL. Rows without "reduced form" entries should always be transcribed in their full form, regardless of pronunciation.

Category	CORAAL Representation (Full Form)	CORAAL Representation (Reduced Form)	Notes
have reduction	`must have`	`musta`	
	`would have`	`woulda`	for wouldn't'ta, transcribe `wouldn't have`
	`should have`	`shoulda`	
	`could have`	`coulda`	
	`might have`	`mighta`	
to reduction	`going to`	`gonna ~ I'm'a`	`gonna` includes [gon]/[gõ] variants
	`have to`	`hafta`	
	`used to`		Always transcribed as `used to`
	`trying to`	`tryna`	
	`supposed to`	`sposta`	`sposta` includes [po] variants
	`fixing to`	`finna`	
	`got to`	`gotta`	
	`want to`	`wanna`	
	`ought to`	`oughta`	
syllable reduction	`because`	`cause`	
	`until`	`til`	
	`about`		
	`remember`		
	`around`		
Other reduction	`talking about`		e.g. quotative talmbout (Vaughn-Cooke 1976; Jones 2016) transcribed as `talking about`
	`them`	`'em`	
	`let me`	`lemme`	
	`what do you/what are you`	`whatchu/whatcha`	The difference between `whatchu` and `whatcha` is the final vowel (whatchu is more common in CORAAL)
	`got you`	`gotcha`	

C. Disfluent speech

Restarts

- Speaker restarts are indicated with a single dash.
  - ▪ `That was- that was some good times. (DCB_se2_ag2_m_03)`
- When a filled pause precedes a restart, the restart should be indicated before the filled pause.
  - ○ `And so, she went to- um, she was born in Georgia. (ROC_se0_ag2_f_04)`

Filled Pauses

- Only *uh, um,* and clear cases of *ah* are used. More than one filled pause in a row is not treated as a restart. A comma should precede and follow each filled pause.
  - ▪ `It'd be at the, um, other campus, in Largo. (DCB_se1_ag1_f_01)`
  - ▪ `So, when I moved to East Capitol, uh, mm, you know, I had to wait for a bunch of stuff when I first got there, had to set up my room. (DCB_se3_ag1_f_01)`
  - ▪ `Ah, I think the former Grammar teacher that I was just telling you about, um, was very good. (DCA_se3_ag3_m_01)`

Mispronounced Words

- Mispronounced words are transcribed as they are pronounced, in slashes.
  - ○ `Christopher /Folumbus/ I mean Columbus. (DCA_se2_ag1_f_03)`

D. Other features

Discourse markers

- Utterance final *so.* is transcribed with a following period.
  - ○ `And there's sort of like a slump there, you know, so. (DCA_se2_ag2_f_03)`
- Lip Smacking/Teeth Sucking, *if* relevant to the present discourse, are transcribed as <ts>.
  - ○ `Like, try to be funny, so. <ts> I gave him my number. But a long story short, we started talking. (DCB_se1_ag1_f_03)`

Overlap

- Speaker overlap is noted by the use of square brackets, for all parties to the overlap. The overlap markers are always placed at word boundaries.
  - ▪ `DCB_int_01: Or is he [retired? No?]`
  - ▪ `DCB_se1_ag2_f_02: [No, he was] in the reserve but he just went back to [active.] Yeah.`
  - ▪ `DCB_int_01: [Oh.]`

Unintelligible/Inaudible Speech:

- Slashes are used to enclose sections of unsure transcription. Transcribers often place "best guesses" within the slashes, or write /unintelligible/ for unintelligible talk or /inaudible/ for inaudible talk. For unintelligible talk of less than three syllables, transcribers can also use question marks, /??/, within the slashes to indicate each syllable of unintelligible speech.

    - `DCA_se3_ag3_m_01: And here again now, I can appreciate that /unintelligible/, [but at] that time, I thought he was being very unreasonable. I was only at the circus.`
    - `DCA_int_01: [Yeah.]`
    - `DCA_se3_ag3_m_01: [<laugh>]`
    - `DCA_int_01: [/inaudible/.]`

Non-linguistic/meta-linguistic noises:

- Noises like laughter, hand clapping, and throat clears are often indicated by short descriptions enclosed within angle brackets. These are only used to describe actual noises, not features like voice quality.
    - `DCA_int_04: What about the kids that go there? Can you tell me anything about them?`
    - `DCA_se3_ag1_f_06: Well we're all different. <laugh>`
    - `DCA_int_04: Mm-hm.`
- Other examples used in CORAAL
    - `<cough>, <clears throat>, <laugh>, <yawns>, <snap>, <sound effect>, <grumbles>, <inhale>, <exhale>, <microphone feedback>, <clap>, <ts>, <pp>, <imitates music>,`
        - `<ts>` represents teeth suck. Transcribers were asked to transcribe `<ts>` if it appeared to be relevant to the present discourse, and related to communication. `<ts>` covers a range of sounds.
        - <pp> represents a bilabial trill interjection.
    - *speech in slashes as well as angle brackets, if overlapping, can also be within square brackets
        - `[I mean /unintelligible/] you know, you- you gonna have your trickery, <clears throat> in the government as far as that goes. (DCB_se1_ag3_m_02)`

Line-level notes:

- Notes can be included by the use of parentheses in a transcribed utterance. Features like voice quality are noted this way. These do not have to be within square brackets.
    - `(whispered) (while chewing) (laughing) (atypical pronunciation) (singing) (rapping) (coughing) (breathy)`
        - `(atypical pronunciation)` (e.g. in DCB_se1_ag3_m_02) is used when the speaker is describing an accent specific pronunciation.
        - `Baltimore, they be like, what up dog (atypical pronunciation) (DCB_se1_ag3_m_02)`

Redaction

- Slashes and a redaction code are used to obscure real names, addresses, places of work, and schools.
- Redaction codes are as follows, with the # being the number of syllables obscured:
- `RD-WORK, RD-SCHOOL, RD-ADDRESS, RD-NAME`
    - e.g. Winston High School → `/RD-SCHOOL-4/`
- In CORAAL, all redacted codes have been replaced with tones, which were generated based on the mean pitch and amplitude of the speech being redacted.

Morphophonological vs. morphosyntactic differences

- When it's not clear that the process is phonological or morphosyntactic (e.g. a result of consonant cluster reduction vs. an unmarked verb), transcribers were instructed to err on the side of using standard orthographical conventions (i.e. transcribed the *d* in "The boy named Bill").
- But, for clearly morphological/syntactic features, transcribers were instructed to transcribe as close to the audio as possible.
- Examples of common AAL morphosyntactic features included in transcripts are:

Category	*Example*	*Notes*
Third person singular –*s* absence	`So she run over to her bag (DCA_se2_ag3_m_01)`	
Possessive –*s* absence	`We was over my uh, father mother house. (DCA_se1_ag1_f_01)`	
Possessive *they*	`The young girls today wanna be friends with they kids. (DCB_se1_ag3_f_02)`	Not r-less *their*

Common interjections/filler words

- Since several speech sounds (e.g., "mm-hm") don't have codified spellings. The table below gives orthographic guidelines for common interjections and filler words.

Transcription	*Notes*
`Uh-huh`	Positive/neutral
`Uh-uh`	Negative
`Nuh-uh`	Negative (nasalized uh-uh)
`Mm-hm`	Positive/neutral
`Mm`	Positive/neutral
`Mm-mm`	Negative
`Okay`	
`Mkay`	
`Yep/Yup`	Both are transcribed
`Nah/Naw`	Both are transcribed
`Oh`	[oʊ]
`Ooh`	[u]
`Ayo`	[ejo]
`Hoo`	[hu]

CORAAL lexical conventions and dialect specific items

- The following are several common, dialect specific lexical items that appear in the corpus that (1) don't have a standardized spelling or (2) might be unfamiliar to some CORAAL users. In some cases, these are features of African American Language.

Transcription	*Notes*
`aight`	Reduction of *alright*
`aks`	Metathesis of *ask*

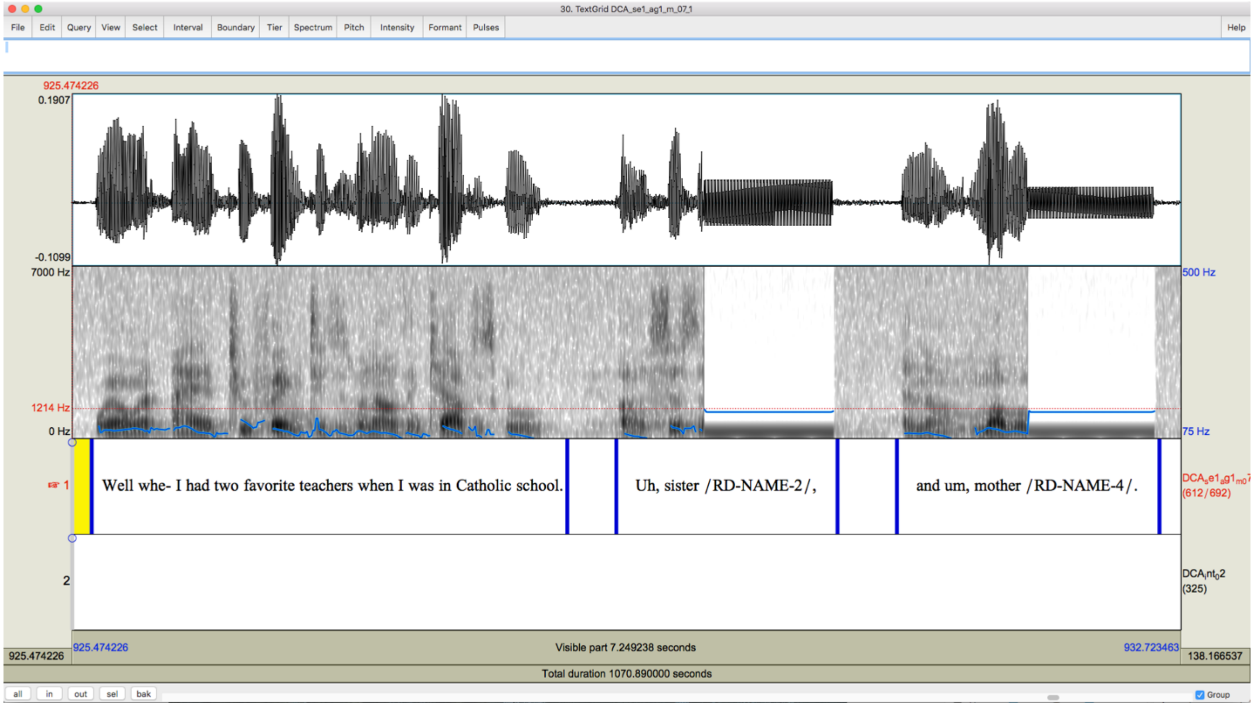
ay!	Common exclamation ([eɪ])
‘bacca	Reduced *tobacco* in Princeville
bih	Reduced *bitch*
bougie	Perceived as upscale
brazy	Means crazy
bruh	Variant of *bro*
cuz	Reduced *cousin*. Not reduced *because*
‘em	Reduced *them*
fella	Variant of *fellow*
go go	Popular style of music in DC
hisself	Regularized *hisself*
jai	Means really/very in DC. [ʤaɪ]
effed up	Not *F-ed up*
mama	Not spelled *momma*
mumbo sauce	Popular condiment found in DC
murk	Means murder
ratchet	Negative evaluative term
shorty	Female companion
wilding	To act wild ([waɪlɪn])
wont	Past tense *wont* in Princeville
youngin	Transcribed as –in

Redaction and participant anonymity

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. A large number of participants (the majority) did ask to be recognized by name and we acknowledge them by name in the acknowledgments section above. (We are incredibly grateful to all of our participants, named and unnamed.) 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. *(If you are a participant in the corpus and did not get recognized by name, but wish to, please contact the project team and we will be happy to send you a new consent form, where you can give us this permission.)*

Our redaction process involved several steps. During the first round of transcription, transcribers were asked to mark different categories of sensitive information, such as names, street addresses, places of work, and other kinds of personally identifying information. Additionally, transcribers marked the numbers of syllables of the item(s) to be redacted. Redaction codes are described in Part D of the Transcription practices and conventions section, above. 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 based on the mean pitch and amplitude of the speech being redacted, replaced the sensitive information. An example of two redacted utterances in an interview from CORAAL:DCA is shown below in a screenshot from Praat.



Well whe- I had two favorite teachers when I was in Catholic school.

Uh, sister /RD-NAME-2/,

and um, mother /RD-NAME-4/.

Note that 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. This is marked in the Notes section of the metadata spreadsheet. Other excised content includes reading passages and word lists. *Although they are not included in the published corpus, many of the excised portions may still be available for bona fide research use upon request.*

CORAAL transcription formats

Transcripts are available in three formats:

- Praat TextGrids (.TextGrid): Each speaker is on a separate interval tier. Occasionally transcripts have a “misc” (miscellaneous) tier with additional information or transcription of substantial interlopers.
- Plain text (.txt): These files are automatically generated from Praat TextGrids, using a script in Praat. They are formatted as tab-delimited text so they can be opened as text, as Excel spreadsheets, or in R, etc.; they contain the same timestamp information as TextGrids.
- ELAN files (.eaf): ELAN versions of the transcripts were automatically generated from the Praat TextGrids, using ELAN’s “Import Multiple Files As…” feature. (Note that

ELAN files are unmodified output from ELAN's batch conversion tool; the files are not linked to their corresponding audio files and may need to be processed in other ways to maximize their usability in ELAN.)

While they are not published as a part of the official corpus (at this time), the CORAAL development team has also generated phone- and word-level alignments for the transcripts using the Montreal Forced Aligner (McAuliffe et al. 2017) and these may be available upon request, although please note that these are a work-in-progress at this time. In addition to releasing the corpus in the formats available here, we are also working on a syntactically parsed version for a large portion of the corpus (similar to Tortora et al.'s ongoing work on AAPCAppE and CUNY-CoNYCE). This work is in progress, and we will release those annotated versions of the data once they are completed.

The R functions in http://lingtools.uoregon.edu/coraal/explorer/R/CORAAL_web.R can create a few different versions of the CORAAL transcripts as R (data.frame) objects. These include both utterance-level transcripts (as are provided in the main data files) and turn-level transcripts, where speaker utterances are collapsed on a turn-by-turn basis.

CORAAL:DCA (Washington, DC 1968)

Authors: Tyler Kendall, Ralph Fasold, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler
Release Date: January 6, 2018 (v. 2018.01.06)
Update Date(s): April 6, 2018 (v. 2018.04.06) [see errata]
October 6, 2018 (v. 2018.10.06) [see errata]
Interviewers: Ralph Fasold, Walt Wolfram, Carolyn Cunningham, Virginia Lundstrom, Veronica Johnson, Roger Shuy, James Goines, and Gail Marble

About CORAAL:DCA

CORAAL:DCA consists of 68 speakers across 74 recordings, originally collected as part of Ralph Fasold's foundational study of African American Language in Washington, DC (Fasold 1972). The speakers were recorded between March 1968 and August 1969, with dates of birth ranging from 1891 to 1958. The 68 speakers selected for CORAAL are not the exact same set of speakers analyzed by Fasold (1972). We have selected speakers from Fasold's interviews to best represent four age groups and three social class groups, although a balanced demographic matrix is not possible given the emphasis of the original project on young speakers. The youngest age group has additional speakers for two reasons: there are lots of these speakers in Fasold's data and their interviews tend to be shorter, so extra speakers were included to increase the amount of total data available for the demographic group. The social class groups are not completely analogous to Fasold's groups, which are based on the Index of Status Characteristics, but are meant to capture broad social strata.

CORAAL:DCA data

The 74 audio files are 44.1 kHz, 16 bit, mono in WAV format, totaling 34.0 hours and 333.5K words. The data reflects a sociolinguistic interview style in the Labovian tradition, with interview topics including games, school, and favorite movies, among others.

Speaker numbers are listed in each cell.

	Socio-Economic Group 1 (≈LWC)		**Socio-Economic Group 2 (≈UWC)**		**Socio-Economic Group 3 (≈MC)**	
	Female	*Male*	*Female*	*Male*	*Female*	*Male*
Age Group 1 (under 19)	5	8	7	6	6	6
Age Group 2 (20 to 29)	1	1	0	3	5	3
Age Group 3 (30 to 50)	2	1	0	3	1	4
Age Group 4 (51 and over)	0	2	1	1	0	2

CORAAL:DCB (Washington, DC 2016)

Authors: Tyler Kendall, Minnie Quartey, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler
Release Date: January 6, 2018 (v. 2018.01.06)
Update Date(s): April 6, 2018 (v. 2018.04.06) [several speakers added]
October 6, 2018 (v. 2018.10.06) [one speaker added; see errata]
Interviewers: Minnie Quartey and Carlos Huff

About CORAAL:DCB

CORAAL:DCB currently consists of 48 primary speakers across 63 audio files, collected specifically for CORAAL. The speakers were recorded between July 2015 and December 2017. Speakers were collected through a friend of a friend network to fill a 4 x 3 demographic matrix, as was done for DCA. The socioeconomic groups here are meant to capture broad social strata; the qualitative labels are simple descriptors to help orient users around the ordering. These are not meant to represent theoretically motivated socioeconomic assessments of individuals. They are also not intended to be perfectly analogous to Fasold's classifications. There are theoretical and practical issues comparing socioeconomic indices in the DC community 50 years apart. We have tried to capture and include in the metadata broad information about speakers' demographic backgrounds, but leave questions of interpretation up to end-users.

CORAAL:DCB data

The 63 audio files are 44.1 kHz, 16 bit, mono in WAV format, totaling 46.0 hours and 515K words. Most interviews were recorded at a higher audio quality (48 kHz, 32 bit, stereo) but down-sampled for distribution. Interviews are sociolinguistic styled interviews on topics such as life in Washington DC, and the interviewee's neighborhood, schooling, and work history. (In most cases, interviewers also collected two reading passages, a word list, and metalinguistic commentary although these are not transcribed, redacted, or otherwise included in CORAAL; these may be made available at a later date.)

Speaker numbers are listed in each cell. We seek to eventually include, at minimum, two speakers per demographic cell. In a few cases, more than two speakers are available for a cell, due to the availability of additional persons in these demographic groups.

	Socio-Economic Group 1 (≈WC)		**Socio-Economic Group 2 (≈LMC)**		**Socio-Economic Group 3 (≈UMC)**	
	Female	*Male*	*Female*	*Male*	*Female*	*Male*
Age Group 1 (under 19)	3	3	1	1	1	1
Age Group 2 (20 to 29)	3	3	2	1	1	0
Age Group 3 (30 to 50)	3	3	2	3	2	2
Age Group 4 (51 and over)	1	2	5	1	2	2

CORAAL:PRV (Princeville, NC 2004)

Authors: Ryan Rowe, Walt Wolfram, Tyler Kendall, Charlie Farrington, and Brooke Josler
Release Date: April 6, 2018 (v. 2018.04.06)
Update Date(s): October 6, 2018 (v. 2018.10.06) [see errata]
Interviewers: Ryan Rowe, Jeanine Carpenter, Drew Grimes, Kristy D'Andrea

About CORAAL:PRV

CORAAL:PRV consists of 16 primary speakers across 32 audio files, collected by Ryan Rowe, Walt Wolfram, and colleagues for the North Carolina Language and Life Project. Princeville, NC is the oldest town incorporated by African Americans in the U.S. Many community members can trace their families back to the original founders of the town. The speakers here were recorded between August 2003 and June 2004. As of the 2000 census, African Americans composed 97% of the population (Rowe 2005, see also Kendall 2007b and Kendall & Wolfram 2009).

Speakers were selected for CORAAL from the larger dataset to fill a 2 x 3 demographic matrix. For PRV, as well as additional upcoming corpus sub-components, less comprehensive samples are targeted than for CORAAL:DC. We do not focus on socioeconomic strata, but focus on providing a distribution across gender and age groups. In file naming, the socioeconomic group is listed as "0" (e.g., PRV_se0_ag1_m_01) to denote no focus on socioeconomic groups (not to indicate a group lower than 1). This is meant simply to maintain the file naming structure of CORAAL:DC. We have attempted to capture and include in the metadata broad information about speakers' demographic backgrounds, but leave questions of interpretation up to end users.

CORAAL:PRV data

The 32 audio files are 44.1 kHz, 16 bit, mono in WAV format, totaling 13.9 hours and 156.1K words. Interviews were recorded on cassette tape and transferred to digital formats in 2007 and 2008 at NC State University. Interviews are sociolinguistic styled interviews on topics such as life in Princeville, schooling, and Hurricane Floyd, which left much of the town underwater in 1999.

Speaker numbers are listed in each cell. This CORAAL sub-component includes only a selection of the total interviews collected in Princeville. There are currently no plans to transcribe more speakers. In a few cases, more than two speakers are available for a cell, due to the availability of additional persons in these demographic groups.

	Socio-Economic Group 0	
	Female	*Male*
Age Group 1 (under 29)	2	2
Age Group 2 (30 to 50)	3	2
Age Group 3 (51 and over)	4	3

CORAAL:ROC (Rochester, NY 2016)

Authors: Sharese King, Charlie Farrington, Tyler Kendall, Emma Mullen, Lucas Jenson, and Shelby Arnson
Release Date: October 6, 2018 (v. 2018.10.06)
Interviewer: Sharese King

About CORAAL:ROC

CORAAL:ROC consists of 14 primary speakers across 13 audio files, collected in 2016 and 2017 by Sharese King as part of her dissertation research in Rochester, New York. Rochester is city on Lake Ontario, in Monroe County in western New York state. Since the early twentieth century, Rochester has been home to a large African American population (see King 2018).

Speakers were provided by King for CORAAL from a larger dataset to fill a 2 x 3 demographic matrix. For ROC, less comprehensive samples are targeted than for CORAAL:DC. We do not focus on socioeconomic strata, but focus on providing a distribution across gender and age groups. In file naming, like with CORAAL:PRV, the socioeconomic group is listed as "0" (e.g., ROC_se0_ag1_m_01_1) to denote no focus on socioeconomic groups (not to indicate a group lower than 1). We have attempted to capture and include in the metadata broad information about speakers' demographic backgrounds, but leave questions of interpretation up to end users.

CORAAL:ROC data

The 13 audio files are 44.1 kHz, 16 bit, mono in WAV format, totaling 11.8 hours and 126.1K words. Interviews were recorded on a Zoom H2N recorder, with an AudioTechnica AT831b microphone, and a MiniJack to XLR audio cable between 2016 and 2017, often at the homes of the interviewees. Interviews are sociolinguistic styled interviews on topics such as life in Rochester, schooling, as well as metalinguistic questions about perception of Rochester accents.

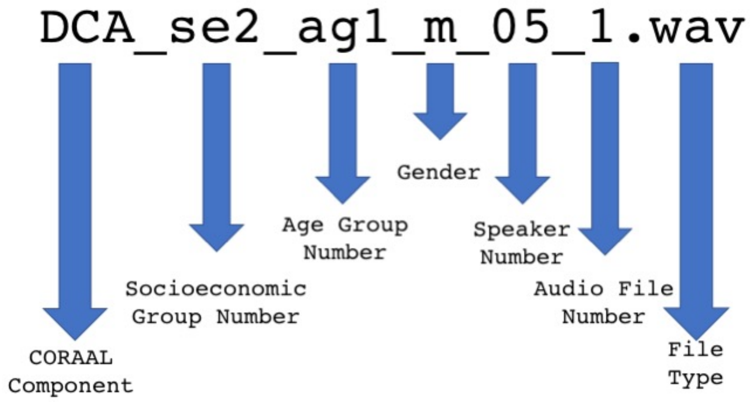
Speaker numbers are listed in each cell. We anticipate adding two more speakers in the next release of CORAAL (two males in Age Group 2). In two cases, more than two speakers are available for a cell, due to the availability of additional persons in these demographic groups.

	Socio-Economic Group 0	
	Female	*Male*
Age Group 1 (under 29)	3	3
Age Group 2 (30 to 50)	4	0
Age Group 3 (51 and over)	2	2

CORAAL Metadata

Speaker and file labeling conventions

Speaker and file names in the corpus are labeled systematically.



For example, file *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 metadata information below), male (m) number 5 (i.e. the fifth speaker in the cell of the demographic matrix). The final 1 indicates the audio file number. As noted in the description of CORAAL:PRV and CORAAL:ROC above, 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.

Most speakers are contained in just one audio file but occasionally, such as for DCA_se2_ag1_m_05, there are multiple files (this speaker has two files: DCA_se2_ag1_m_05_1.wav & DCA_se2_ag1_m_05_2.wav). As discussed above, for each WAV file there is a corresponding Praat TextGrid, ELAN file, and plain text file with identical file labels. (Again, all three of these transcript versions contain identical information and are derived from the Praat Textgrid.)

Metadata files and notes

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, like MS Excel, or in R. The metadata files are .txt files downloadable with the rest of each corpus component's files. For example, metadata for DCA are in the file labeled DCA_metadata_2018.10.06.txt (for version 2018.10.06).

All files in CORAAL have been anonymized and have been trimmed only to contain conversation portions of the sociolinguistic interviews. All of the files are stored (in original formats) on the Sociolinguistic Archive and Analysis Project (SLAAP; https://slaap.chass.ncsu.edu/), and SLAAP often contains more files from a sociolinguistic fieldwork project than just those included in CORAAL. The SLAAP codes for the files (e.g. *SLAAP.Spkr*) are provided in the metadata file when appropriate. For DCA, SLAAP codes are reflective of the original codes used in Fasold (1972), included here for backward compatibility. For DCB, SLAAP codes were given to recordings as they were completed. For PRV, SLAAP codes are reflective of the codes used when the audio tapes were digitized and uploaded to SLAAP in 2007 and 2008.

Several categories in the metadata spreadsheets apply to all CORAAL components, while others apply only to specific components. Metadata notes are provided below, organized by those that apply to all components first, with information specific to sub-components in the sections following. Most speaker files obtained from Ralph Fasold (DCA) came with an Informant Data Sheet (IDS), where much of the metadata information comes from. The IDS collected basic demographic data, such as sex, age, address, birthplace, parents' birthplace, as well as a social class section (see below for discussion). For DCB, interviewers were asked to complete a similar Interview Report Form (IRF) for each speaker, which collected the same kinds of demographic information as Fasold's IDS. In addition to general demographic information, the IRF contains additional interview notes (e.g., interruptions, background noise, etc.) as well as topics covered over the course of the interview. For PRV, some speaker information was gathered from the metadata on SLAAP, while other information was obtained from the content of the interviews themselves.

Metadata notes (all CORAAL components)

- In the follow table, each column provided in the tab-delimited text file is outlined. The majority of the categories apply to all components, but some categories require further information about how the categories were coded in each corpus component.

Field	*Description and Examples*
CORAAL.Sub	CORAAL component three-character code (DCA; DCB; PRV)
Version.Created	Version number of the recording's initial release (e.g. v. 2018.01.06, v. 2018.04.06, v. 2018.10.06).
Version.Modified	When modifications are made to the transcripts, the most up-to-date version number is listed here. Any modifications will also be listed in the Change Log found earlier in this guide.
CORAAL.Spkr	Speaker code for CORAAL (e.g. DCB_se1_ag1_f_01; PRV_se0_ag3_m_01).
CORAAL.File	File name used on audio file and transcription files (e.g. DCB_se1_ag1_f_01_1; PRV_se0_ag3_m_01_1).
Audio.Folder	Name of folder that houses the specific audio file (e.g. DCA_audio_part01_2018.10.06)
Tarball	Name of compressed folder with extension (tar.gz) (e.g. DCA_audio_part01_2018.10.06.tar.gz)
Primary.Spkr	In DCB & PRV there are several interviews where there are other interviewees present in addition to a primary interviewee. In the metadata, there is a row for each interviewee present in the interview. *Yes* means the speaker was the primary interviewee, *no* means the speaker was secondary.
SLAAP.Collection	Alternate codes were used in SLAAP. DCA==wds, DCB=wdc, PRV==prv.

SLAAP.Spkr	SLAAP speaker code (e.g. wdc003).
SLAAP.Interview	SLAAP interview file code (e.g. wdc0030d).
Gender	Speaker's gender: Male, Female, Non-binary.
Age	Speaker's actual age in years (currently ranging from 12 to 83).

Age.Group	CORAAL:DC (four age groups):
	-19; 20 to 29; 30 to 50; 51+.
	
	CORAAL:PRV (three age groups):
	-29; 30 to 50; 51+.

Year.of.Birth	Speaker's actual year of birth (currently ranging from 1891 to 2005).
Year.of.Interview	DCA interviews took place in either 1968 or 1969. DCB interviews took place between 2015 and 2017. PRV interviews took place in either 2003 or 2004.
CORAAL.SEC.Group	See DCA and DCB Metadata notes for specific information on how CORAAL.SEC was coded. For sub-components without coding for socioeconomic group, such as PRV, SEC.Group is always listed as 0 (e.g., PRV_se0_ag2_f_02).
Education	For DCA, this is what was provided on the IDS, variability in what was reported is a result of different fieldworkers. For DCB, this is what was provided on the IRF. For PRV, this was primarily determined through the information found in the sociolinguistic interview.
Edu.Group	Categories are collapsed from the Education field into groupings: College (completed college) Elementary School (completed elementary school) Graduate School (completed graduate school) High School (completed high school, or equivalent) Some College (attended college but did not graduate) Some High School (attended high school but did not graduate) Student_college (current college student) Student_hs (current high school student) Student_ms (current middle school student).
Occupation	For DCA, this is provided on the IDS (Student; Postal worker; Engineering). For DCB, this is provided on the IRF. For PRV, this was primarily determined through the information provided in the sociolinguistic interview, but occasionally available in the metadata available on SLAAP.
Region.in.City	CORAAL:DC only. Quadrant in DC where the participant resided at the time of the interview (NW; NE; SE; SW). "DC" is used if participant resided in multiple quadrants for an unknown amount of time.
LOR	Length of Residence in current location. CORAAL:DC as listed on the IDS/IRF (e.g. 20 years). Not included in PRV metadata because information is not consistent across speakers. Most speakers have lived in Edgecombe County area their whole life.

LOR.Percent	LOR/Age (e.g. 20 (LOR)/25(Age) = 80% LOR.Percent).
Other.Places.Lived	LOR is not available for these other places. If the LOR.Percent is 100 and there is something listed in this category, assume this is under a year of residence.
Relationship.To.Others.In.Corpus	Occasionally, relationships are made clear in the interview, or is written on the IDS/IRF. The format lists the CORAAL.Code of the speaker, with the relationship in parentheses: e.g. `DCB_se1_ag3_f_02 (mother)` indicates that DCB_se1_ag3_f_02 is this speaker's mother.
Guardian.1.Birthplace	Mother birthplace.
Guardian.1.Birthplace.State	Mother birthplace by state.
Guardian.1.Education	Mother education level; only available for DCA.
Guardian.1.Occupation	Mother occupation.
Guardian.2.Birthplace	Father birthplace.
Guardian.2.Birthplace.State	Father birthplace by state.
Guardian.2.Education	Father education level; only available for DCA.
Guardian.2.Occupation	Father occupation.
Guardian.Notes	Occasionally, Guardian.1 and Guardian.2 do not correspond to mother and father. This column describes any other relevant information.
Interviewer.Code	CORAAL code for interviewers, (e.g., DCA_int_01). In DCB, speaker DCB_se1_ag3_m_01 is also an interviewer. His CORAAL.Code is also used as his Interviewer.Code. In PRV, there is one primary interviewer (PRV_int_01), and three secondary interviewers.
Interviewer.Initials	For DCA & PRV, initials are included (e.g., RF; WW; RWS; RR). For DCB, SLAAP codes are included here (e.g., wdc000; wdc001).
Interviewer.Gender	Female, Male.
Interviewer.Ethnicity	Ethnicity of interviewer.
Interviewer.Age	Age is approximated into five year ranges.
Interviewer.Relationship	In DCB, information provided by interviewer. Collapsed into three categories: Acquaintance; Close relationship; No previous relationship . In DCA & PRV, no previous relationship is assumed, but this is not definitive.
Recording.Equipment	Equipment used during the recording (if known).
Microphone	Microphone(s) used during the recording (if known).
Stereo.Mono	Number of channels used during the recording (if known).
Bit.Rate	Bit rate used during the recording (if known).
Sampling.Rate	Sampling rate used during the recording (if known).
Source.Device	Source device used during the recording (if known).
Dig.Capture.System	System used in digitization (if used).
Dig.Capture.Device	Device used in digitization (if used).
Dig.Capture.Software	Software used in digitization (if used).
Dig.Sampling.Rate	44.1kHz.
Dig.Bit.Rate	16 bit.
Dig.Channels	Mono.
CORAAL.Length.of.Transcript	Length of transcripts in seconds (listed for Primary.Spkr only).
CORAAL.Word.Count	Number of words in each transcript (listed for Primary.Spkr only).
Is.Misc.Tier	Lists whether a miscellaneous tier was included in the CORAAL release.
Notes	Describes the activity and time (in seconds) where a misc tier is relevant, as well as any general notes regarding recording.

Metadata notes (DCA:1968)

Recordings were made on a reel-to-reel recorder of unknown make and model in a variety of settings (at the Center for Applied Linguistics, the participant's home, a local church, etc.). Digitization occurred at North Carolina State University in 2013, using an AEC A-6300 Reel-to-Reel player, captured on Audacity with a Sound Devices USBPre 2 preamp, in an effort supervised by Michael J. Fox. Most files were digitized with sampling rates primarily at 48kHz, but a few were digitized at 44.1kHz. For consistency across CORAAL, all audio files were converted to mono with a sampling rate of 44.1kHz and a bit rate of 16.

Field	*Description and Examples*
In.Fasold.1972	This column reports whether the speaker is examined in Fasold (1972). The 68 speakers selected for CORAAL are *not* the exact same set of speakers analyzed by Fasold (1972), which only used a selection of the entire dataset of 90 speakers. We have selected speakers from Fasold's interviews to best represent four age groups and three social class groups.
Fasold.ISC.Group	The categories are based on Warner et al.'s (1960) Index of Status Characteristics (ISC). The actual number value is provided in *ISC.Fasold*. More information about this rating system can be found in Fasold (1972:17-21). An example ISC is shown below.
Fasold.SEC.Group	This category is Charlie Farrington's attempt to collapse the detailed ISC groupings into workable class categories, paying attention to number of speakers per group. The categories are: ▪ Lower Working Class (CORAAL.SEC.Group 1) ▪ Upper Working Class (CORAAL.SEC.Group 2) ▪ Lower Middle Class (CORAAL.SEC.Group 3) ▪ Upper Middle Class (CORAAL.SEC.Group 3) ▪ Upper Class (CORAAL.SEC.Group 3).
Occ.Score.Fasold	ISC Occupation Score. See Fasold.ISC.Group above, and Fasold (1972: 17-21) for more information.
House.Score.Fasold	ISC House Type Score. See Fasold.ISC.Group above, and Fasold (1972: 17-21) for more information.
Dwelling.Area.Score.Fasold	ISC Dwelling Area Score. See Fasold.ISC.Group above, and Fasold (1972: 17-21) for more information.
ISC.Fasold	ISC Score calculated by Fasold. See rating example below.

- *Fasold ISC Example*
    - The categories are based on Warner et al.'s (1960) Index of Status Characteristics (ISC). The actual number value is provided in *ISC.Fasold*. More information about this rating system can be found in Fasold (1972:17-21).

Sample ISC (Speaker # DCA_se1_ag1_m_06)

	Rating	*Weight*	*Total*
Occupation	7	5	35
House Type	6	4	24
Dwelling Area	5	3	15
		I.S.C. 74 (Lower Working)	

Metadata notes (DCB:2016)

Almost all interviews for DCB were recorded on a Marantz PMD661 MK II digital recorder, using SHURE SM93 and Microflex MX 100 microphones. These recordings were made in stereo, with a bit rate of 24 (PCM-24), and a sampling rate of 48kHz. The last few interviews were recorded using a SHURE SM93 microphone alone, in mono, with a bit rate of 24 (PCM-24), and a sampling rate of 44.1kHz. There is one interview, recorded in early 2015, using an Olympus LS11 digital recorder and an Audio Technica AT8010 Omnidirectional Condenser microphone. Stereo files were converted to mono and down sampled to 44.1kHz/16 bit for the CORAAL release (this was done using the software “sox”).

Field	*Description and Examples*
CORAAL.SEC.Group	These SEC groups are estimations based on the fieldworker’s knowledge of the individuals and the African American community in Washington D.C. We’ve done our best to be careful in making group assignments, but these group assignments are meant for balancing the corpus and as a heuristic. They are not meant to be taken as the outcome of a sociological/socio-economic analysis of the speakers. ▪ Group 1 – roughly correlates to Working Class ▪ Group 2 – roughly correlates to Lower Middle Class ▪ Group 3 – roughly correlates to Upper Middle Class.

Metadata notes (PRV:2004)

Recordings were made on cassette recorders (most likely Marantz PMD222 devices, with a Sony ECM-44B lavalier microphone) in a variety of settings in the Princeville community (the participant’s home; fire house, etc.). Digitization occurred at North Carolina State University in 2006/2007 by graduate student employees using a Tascam CC-222 MKIII to convert the tape to compact disc format, which was then captured on a Mac OS using Audacity or, in a few cases, Praat. Princeville files were digitized with sampling rate of 44.1kHz and a bit rate of 16. All PRV audio files are mono with a sampling rate of 44.1kHz and a bit rate of 16.

Metadata notes (ROC:2016)

Recordings were made on a Zoom H2N digital recorder, with an AudioTechnica AT831b microphone, connected with a MiniJack to XLR audio cable in a variety of settings in the Rochester area. Rochester files were recorded with a sampling rate of 44.1kHz and a bit rate of 16.

Projects Using CORAAL

We may not be able to keep an accurate list of project and publications that have used CORAAL indefinitely, but for now we attempt to provide a list of current and published studies that have used CORAAL here. *Please let us know if you are using CORAAL and want to be added to this list!*

Publications

Arnson, Shelby & Charlie Farrington. 2017. Twentieth century sound change in Washington DC African American English. *Penn Working Papers in Linguistics* 23(2): Article 2. https://repository.upenn.edu/pwpl/vol23/iss2/2/

Cukor-Avila, Patricia & Ashley Balcazar. Forthcoming. Exploring grammatical variation in the Corpus of Regional African American Language. *American Speech* 94(1).

Farrington, Charlie. Forthcoming. Incomplete neutralization in African American English: The case of final consonant voicing. *Language Variation and Change*.

Farrington, Charlie & Natalie Schilling. Forthcoming. Contextualizing the Corpus of Regional African American Language:DC AAL in the Nation's Capital. *American Speech* 94(1).

Forrest, Jon & Walt Wolfram. Forthcoming. The status of (ING) in African American Language: A quantitative analysis of social factors and internal constraints. *American Speech* 94(1).

Grieser, Jessica A. Forthcoming. Investigating topic-based style shifting in the classic sociolinguistic interview. *American Speech* 94(1).

Holliday, Nicole R. Forthcoming. Variation in question intonation in the Corpus of Regional African American Language. *American Speech* 94(1).

McLarty, Jason, Taylor Jones, & Christopher Hall. Forthcoming. Corpus-based sociophonetic approaches to post-vocalic R-lessness in African American Language. *American Speech* 94(1).

Quartey, Minnie & Natalie Schilling. Forthcoming. Shaping 'connected' vs. 'disconnected' identities in narrative discourse in DC African American Language. *American Speech* 94(1).

Presentations

Cukor-Avila, Patricia. 2018. Exploring Grammatical Variation in the Corpus of Regional African American Language. Paper presented at the 92nd Annual Meeting of the Linguistic Society of America: Salt Lake City.

Farrington, Charlie. 2015. Word final stop weakening in African American English. Poster presented at New Ways of Analyzing Variation 44: Toronto.

Farrington, Charlie. 2017. Incomplete neutralization in African American English: the role of vowel duration. Paper presented at New Ways of Analyzing Variation 46: Madison, WI.

Farrington, Charlie. 2018. Regionality and final fricative deletion in African American Language. Paper presented at New Ways of Analyzing Variation 47: New York, NY.

Farrington, Charlie, Shelby Arnson, & Tyler Kendall. 2018. Back vowel changes in Washington DC African American Language over the twentieth century. Paper presented at the 92nd Annual Meeting of the Linguistic Society of America: Salt Lake City.

Farrington, Charlie, Jason McLarty, & Tyler Kendall. 2016. Corpus and sociophonetic approaches to possessive *they* in African American English. Poster presented at American Dialect Society Annual Meeting: Washington DC.

Forrest, Jon & Walt Wolfram. 2018. A Quantitative Analysis of Social Factors and Internal Constraints on (ING) in African American English. Paper presented at the 92nd Annual Meeting of the Linguistic Society of America: Salt Lake City.

Jones, Taylor, Jason McLarty, & Chris Hall. 2018. Corpus-based sociophonetic approaches to gradient post-vocalic R-lessness in African American Language. Paper presented at the 92nd Annual Meeting of the Linguistic Society of America: Salt Lake City.

Quartey, Minnie & Natalie Schilling. 2018. Shaping 'connected' vs. 'disconnected' identities in discourse: Narratives, position and stance in DC AAL. Paper presented at the 92nd Annual Meeting of the Linguistic Society of America: Salt Lake City.

References

Bailey, Guy & Natalie Maynor. 1985. The present tense of be in Southern Black folk speech. *American Speech* 60:195-213.

Bailey, Guy & Natalie Maynor. 1987. Decreolization? *Language in Society* 16:449-74.

Bailey, Guy, Natalie Maynor, & Patricia Cukor-Avila. eds. 1991. *The Emergence of Black English: Text and Commentary*. Amsterdam and Philadelphia: John Benjamins.

Berez-Kroeker, Andrea, Gary Holton, Susan Smyth Kung, & Peter Puslifier. Reproducible research in linguistics: toward a data-driven science of language. Paper presented at the 2017 Annual Meeting of the Linguistics Society of America. Austin, TX.

Boersma, Paul & David Weenink. 2014. Praat: Doing phonetics by computer. [Software]

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

Green, Lisa. 2002. *African American English: A Linguistic Introduction*. Cambridge, U.K.: Cambridge University Press.

Jones, Taylor. 2016. AAE talmbout: An overlooked verb of quotation. *Penn Working Papers in Linguistics* 22: Article 11.

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

Kendall, Tyler. 2007b. "The people what makes the town": The semiotics of home and town spaces in Princeville, NC. *The North Carolina Folklore Journal* 54(1): 33-53.

Kendall, Tyler. 2008. On the history and future of sociolinguistic data. *Linguistic and Language Compass* 2:332-351.

Kendall, Tyler. 2011. Corpora and from a sociolinguistic perspective (Corpora sob uma perspectiva sociolinguística). In Stefan Th. Gries (ed.), Corpus studies: Future directions, special issue of *Revista Brasileira de Linguística Aplicada* 11(2):361-389.

Kendall, Tyler & Jason McLarty. 2018. *ORAAL: Online Resources for African American Language.* Eugene, OR: Online Resources for African American Language Project. http://oraal.uoregon.edu/

Kendall, Tyler & Walt Wolfram. 2009. Local and external language standards in African American English. *Journal of English Linguistics*, 37(4): 305-330.

Kendall, Tyler, Joan Bresnan, & Gerard Van Herk 2011. The dative alternation in African American English: Researching syntactic variation and change in a conglomerated corpus. *Corpus Linguistics and Linguistic Theory*, 7(2):229-244.

Kurath, Hans. 1949. *A Word Geography of the Eastern United States*. Ann Arbor: University of Michigan Press.

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

Labov, William 1969. *The Logic of Nonstandard English*. In James Alatis (ed.), Monograph Series on Languages and Linguistics, No. 22 [20th Annual Round Table: Linguistics and the Teaching of Standard English to Speakers of Other Languages or Dialects], Washington, DC: Georgetown University Press. 1-43.

Labov, William. 1972. *Language in the Inner City: The Black English Vernacular*. Philadelphia, PA: University of Pennsylvania Press.

Labov, William. 1987. Contribution to: Are black and white vernaculars diverging? Papers from the NWAV XIV panel discussion. *American Speech* 68(2):5-12.

Lanehart, Sonja (ed.). 2015. *The Oxford Handbook of African American Language*. Oxford, UK: Oxford University Press.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner. Version 0.9.0. Computer program. http://montrealcorpustools.github.io/Montreal-Forced-Aligner/.

McDavid, R. I. & V. G. McDavid. 1951. The relationship of the speech of negroes to the speech of whites. *American Speech* 26:3-17.

Mufwene, Salikoko, John Rickford, Guy Bailey, & John Baugh (eds.). 1998. *African-American English: Structure, History and Use*. London: Routledge.

Poplack, Shana & Sali Tagliamonte. 2001. *African American English in the Diaspora*. Malden/Oxford: Blackwell.

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

Rickford, John R., Arnetha Ball, Renee Blake, Raina Jackson, & Nomi Martin. 1991. Rappin on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English. *Language Variation and Change* 3: 103-132.

Rickford, John R., Julie Sweetland, Angela Rickford, & Thomas Grano. 2012. *African American, Creole, and Other Vernacular Englishes in Education: A Bibliographic Resource.* New York: NCTE-Routledge Research Series.

Rowe, Ryan. 2005. The development of African American English in the Oldest Black Town in America: Plural -s Absence in Princeville, NC. Master's Thesis. Raleigh: North Carolina State University.

Schneider, Edgar W. (ed.). 1996. *Focus on the USA.* Philadelphia/Amsterdam: John Benjamins.

Stewart, William A. 1968. Continuity and change in American Negro dialects. *The Florida FL Reporter* 6:3-4,14-16, 18.

Thomas, Erik R. & Tyler Kendall. 2007. NORM: The Vowel Normalization and Plotting Suite. Online resource. http://lingtools.uoregon.edu/norm/

Tortora, Christina, Beatrice Santorini, Michael Montgomery, & Frances Blanchette. 2012. A hands-on introduction to the Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE). Workshop at Southeastern Conference on Linguistics 79. Lexington, KY.

Warner, Natasha. 2004. Sharing of data as it relates to human subjects issues and data management plans. *Language and Linguistics Compass* 8: 512-518.

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

Wolfram, Walt & Erik R. Thomas. 2002. *The Development of African American English.* Malden/Oxford: Blackwell.

Yaeger-Dror, Malcah & Erik R. Thomas (eds). 2010. *African American Speakers and their Participation in Local Sound Changes: A Comparative Study*. Durham, NC: Duke University Press.