Augmented Datasheet for Speech Datasets

Dataset	Paper	Link
CORAAL	https://direct.mit.edu/books/book/5244/chapter/3537382/Managing-Sociolinguistic-Data-with-the-Corpus-of	https://oraal.uoregon.edu/coraal

1. Motivation

What is the speech dataset name, and does the name accurately describe the contents of the dataset?

The full title “Corpus of Regional African American Language (CORAAL)” shows that the dataset features recorded speech from regional varieties of African American Language.

Can the dataset be used to draw conclusions on read speech, spontaneous speech, or both?

Spontaneous speech only.

Describe the process used to determine which linguistic subpopulations are the focus of the dataset.

While AAL has been extensively studied, it has remained massively underrepresented in terms of publicly available datasets and in terms of its use in general linguistic theory building, and thus far almost all AAL data have remained unavailable for wider, public sharing, due to ethical considerations or limitations from how the data were collected. The CORAAL website details more about AAL and its cultural impact.

2. Composition

How many hours of speech were collected in total (of each type, if appropriate), including speech that is not in the dataset? If there was a difference between collected and included, why?

The creators have transcribed only a portion of the speech and have excluded any parts of the audio the participants requested or that the creators deemed unnecessary, but they do not mention how much the deleted content account for.

How many hours of speech and number of tokens are in the dataset (by each type, if appropriate)?

CORAAL:ATL - 14 files, 8.6 hours and 93.5K words
CORAAL:DCA - 74 files, 34.0 hours and 333.5K words
CORAAL:DCB - 63 files, 46.0 hours and 515K words.
CORAAL:LES - 15 files, 8.4 hours and 102.2K words
CORAAL:PRV - 32 files, 13.9 hours and 156.1K words
CORAAL:ROC - 19 files, 13.2 hours and 138.9K words
CORAAL:VLD - 14 files, 11.5 hours and 112K words

Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified in the dataset and described in the metadata?

The term ‘African American Language’ is defined and well cited, but the term ‘regional dialect/accent’ is not explicitly defined.

For any linguistic subpopulations identified in the dataset, please provide a description of their respective distributions within the dataset.

For each regional group, the table containing the specific numbers of hours and files for each gender (binary: female & male) as well as for each socio-economic background is included in the paper.

How much of the speech data have corresponding transcriptions in the dataset?

All speech data have a corresponding transcription. Any speech data without corresponding transcription was discarded.

Does the dataset contain non-speech mediums (e.g. images or video)?

No.

Do speakers code switch or speak multiple languages, and if so, how is this identified in the data?

There is no mention of speakers code-switching, though there is a possibility of speakers code-switching between dialects.

Does the speech dataset focus on a specific topic or set of topics?

Each group has a set of topics the interviewers bring up often, such as their neighborhood, school life, sports, etc.

Does the dataset include sensitive content that can induce different emotions (e.g., anger, sadness) that can cause the speakers to produce unusual pitch or tone deviating from plain speech?

From the set of topics the interviewers use, there are mentions of natural disasters that occurred in their regions, which could be traumatic for participants as well as bringing up emotional responses.

Does the dataset contain content that complies to the users' needs, or does it result in symbolic violence (the imposition of religious, political, cultural values etc.)?

From the set of topics the interviewers use, there are mentions of metalinguistic questions regarding their regional accents and their perceptions. While this could generate feelings of symbolic violence, we expect this is mitigated in conversation due to the interviewers also being AAL speakers. Otherwise no other sets of topics explicitly indicate symbolic violence.

3. Collection Process

What mechanisms or procedures were used to collect the speech data, e.g.: is the data a new recording of read speech or an interview? Or is it downloaded speech data from public speeches, lectures, YouTube videos or movies, etc.?

Participants were interviewed.

Were all the data collected using the same technical methodology or setting, including the recording environment (e.g., lab, microphone) and recording information (e.g., sampling rate, number of channels)?

Each regional group had different methods of recording the participants, but they manipulated the audio files to be of the same configuration when being released.

Is there presence of background noise?

Not mentioned, assumed to have technical noise especially for older data since the audio quality is low.

For interviewer/interviewee speech data: during the interview process, did interviewers consistently ask questions that are ``fair and neutral''?

The interview questions are not explicitly listed, but the topics brought up are fairly related to everyday lives and can be considered fair and neutral.

Have data subjects consented to the disclosure of the metadata in the dataset? Also, does the metadata include sensitive personal information such as disability status?

Subjects consented to the release of the metadata.

4. Preprocessing/cleaning/labeling

When generating the dataset, was any background noise deleted or adjusted to make all recording qualities similar?

All audio files were adjusted to be of the same configuration when being distributed.

Did the data collectors hire human annotators to transcribe the data? If so, how trained were the annotators in speech transcription for this context? How familiar were they with the corpus material, the vocabulary used, and the linguistic characteristics of different dialects and accents?

The creators built their own transcription convention to follow and annotators were trained on the convention, which included AAL-specific details.

If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?

The creators built their own transcription convention to follow and each audio file went through at least two rounds of validation to ensure consistency.

If the speech data include transcriptions, what software was used to generate the transcriptions (including, e.g., software used by human transcribers)? Are timestamps included in transcriptions? Are the alignments provided with the transcripts?

Praat TextGrid was the main tool utilized for transcribing, but the files were distributed in three different forms of transcription (TextGrid, ELAN, text). Time-alignments are also provided.

Were transcription conventions (such as tagging scheme, treatment of hate speech or swear words, etc.) disclosed along with the corpus?

The transcription conventions and redaction methods were released in the user guide.

Is additional coding performed, separate to transcriptions and tagging?

No additional coding is performed.

5. Uses / Distribution / Maintenance

How are redactions performed on the dataset? Are personally identifiable information or sensitive information removed from only transcripts, audio censored from the speech data, or both?

The redaction was performed on both audio and transcription files. The participants were asked whether their names should be redacted or not, and if they consented to their name being disclosed, personal names were not redacted. All other PII was redacted and a corresponding ‘bleep’ was made to take its place in the audio file.

Is there any part of this dataset that is privately held but can be requested for research purposes?

There are multiple other linguistic tasks some of the interviewees performed but were not released (yet).

Is there a sample dataset distributed? If so, how well does the sample represent the actual dataset? Do they include all forms of speech included in the dataset? How big is the sample?

There is no separate dataset.

Aside from this datasheet, is other documentation available about the data collection process (e.g., agreements signed with data subjects and research methodology)?

Yes. There is a user guide with detailed information (http://lingtools.uoregon.edu/coraal/userguide/CORAALUserGuide_current.pdf).