Datasheet: Archive of Pittsburgh Language and Speech

Jack Rechsteiner and Dan Villarreal

Gebru et al. (2021:1–2) "propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on" for the sake of transparency and accountability. This document is based on their guidelines.

Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
    - The Archive of Pittsburgh Language and Speech, or APLS, was created as an open data resource for (socio)linguistics research in response to the "shoebox of tapes on the shelf" problem (Gawne & Styles, 2022), or the structural problem that researchers often collect large amounts of data that sit unused after the initial project that the data were collected for. Even when data is made available, it is not always available in a way that is easily usable without additional labor and resources. In addition to being an open data resource, APLS has been designed to be of actual use to researchers. This has involved identifying the more tedious aspects of linguistic research, such as correcting transcription intervals, and either manually processing the data before uploading it to APLS or automating the process in APLS itself.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
    - The original data collection was done by Barbara Johnstone (Carnegie Mellon University) and Scott F. Kiesling (University of Pittsburgh) for the Pittsburgh Speech and Society Project, or PSSP. The creation of the APLS dataset was led by Dr. Dan Villarreal on behalf of the University of Pittsburgh. Additional contributors can be found on the credits page of the APLS documentation website at https://djvill.github.io/APLS/doc/credits.
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
    - Funding for the original data collection was provided by:
        - Berkman Fund at Carnegie Mellon University
        - Department of English, Carnegie Mellon University
        - Department of Linguistics, University of Pittsburgh
        - National Science Foundation (Collaborative Research awards BCS-0417657 and BCS-0417684)
    - Funding and resources for the Archive of Pittsburgh Language and Speech was/is provided by:

- Office of Research (via Pitt Momentum Funds), University of Pittsburgh
- Center for Research Computing, University of Pittsburgh
- New Zealand Institute of Language, Brain, and Behaviour, University of Canterbury

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
    - There are two types of instances in the APLS dataset: transcripts and participants.
- **How many instances are there in total (of each type, if appropriate)?**
    - As of APLS version 0.4.3, the dataset contains 274 transcripts and 61 participants.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
    - The APLS dataset is a subset of the audio files from the original Pittsburgh Speech and Society Project (PSSP) sociolinguistic interviews.
    - PSSP contains 589 transcripts and 98 interviewees; 49 transcripts were removed because of poor audio quality; 258 transcripts were excluded because interviewees did not consent to have their data shared publicly and/or the interviewees were not from southwest Pennsylvania; 18 transcripts were excluded because the transcripts contained multiple speakers.
    - Representativeness of APLS vis-a-vis PSSP.

Interviewee Demographic	**PSSP *n* (Percent of Total)**	**APLS *n* (Percent of Total)**
Gender: Female	55 (56%)	24 (60%)
Gender: Male	43 (44%)	16 (40%)
Race: Black	23 (23%)	10 (25%)
Race: White	75 (77%)	30 (75%)
Education: In middle/high school	17 (17%)	0 (0%)
Education: Less than high school	5 (5%)	0 (0%)
Education: High school	37 (37%)	16 (40%)
Education: Undergraduate	16 (16%)	13 (32.5%)
Education: Graduate	23 (23%)	11 (27.5%)

Occupation: Unskilled Manual	16 (16%)	0 (0%)
Occupation: Skilled Manual	20 (20%)	9 (22.5%)
Occupation: Professional	24 (24%)	11 (27.5%)
Occupation: Clerical	38 (39%)	20 (50%)
Neighborhood: Cranberry Township	22 (22%)	6 (15%)
Neighborhood: Forest Hills	27 (28%)	12 (30%)
Neighborhood: Hill District	23 (23%)	10 (25%)
Neighborhood: Lawrenceville	26 (27%)	12 (30%)
Year of Birth: Minimum	1917	1920
Year of Birth: 1st quartile	1941	1941
Year of Birth: Median	1956	1955
Year of Birth: 3rd quartile	1982	1967
Year of Birth: Maximum	1992	1986

    - The sampling strategy for the original PSSP interviews was to focus on four Pittsburgh-area neighborhoods, representing different stages of Pittsburgh's settlement history, where the researchers thought they would find people of different social-class and ethnic backgrounds, and which might be characterized by different social networks and degrees of affiliation with the area and the city. In each neighborhood, researchers interviewed three males and three females from each of four age groupings: people born pre-World War II, people born between 1946 and 1964, people born between 1965 and 1984, and people born between 1985 and 1997. Participants were recruited by means of snowball sampling: the researchers started with people they knew and asked them if there were other people they knew who would fit our criteria. They also placed ads in local newsletters, talked to the leaders of community organizations, and attended neighborhood events.
- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
    - Each transcript instance consists of an audio file of a speech recording from sociolinguistic fieldwork, annotations for that audio file, and metadata. Annotations include linguistic information, like part of speech tags or phonetic transcription, and extralinguistic information, such as background noise labels or indications that some speech has been redacted. Metadata fields include recording date, `episode` (interview series), and duration.
    - Each participant instance, which represents an interviewee, interviewer, or bystander whose speech was captured in the recording, consists of a speaker code and associated metadata. Interviewees in APLS are identified by an anonymized speaker code that includes their neighborhood abbreviation (e.g., CB01, HD17).

- **Is there a label or target associated with each instance?** If so, please provide a description.
    - N/A
- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
    - Yes. Information that could personally identify the interviewees has been redacted from the transcripts.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.
    - Yes. Transcripts that contain recordings from the same interviewee all begin with the same anonymized speaker code and share the same value for the `episode` metadata field (e.g., transcripts for audio from CB01's interview all begin with CB01 and have CB01 as the `episode` in their metadata).
- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.
    - N/A
- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
    - Currently known errors in the dataset will eventually be described in the APLS data coverage checks. There is also potential for errors to be inherited from the tools used to create the APLS dataset; see preprocessing for more information.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
    - The dataset is self-contained. All the data and tools for the dataset exist on the APLS server.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.
    - No. Any such data has been redacted from the dataset.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.
    - Some transcripts might be considered to contain moderately inappropriate or offensive language. Some speakers were asked if they had an experience where they thought they might die and what the experience was. Outside of this, speakers were recorded talking about a variety of topics, some of which may contain incidental occurrences of troubling content. Occurrences of this kind, however, are infrequent in the data set.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
    - Yes. See the Composition section above for information about participant subpopulations.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.
    - The only individual identifiable from the dataset is the fieldwork lead and interviewer for the white speakers, Dr. Barbara Johnstone. It is not possible to identify any other individuals from the dataset. Instances of identifying information for individuals other than Dr. Barbara Johnstone have been redacted from the dataset.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.
    - Race and ethnic origins are metadata attached to participants. Other potentially sensitive information may have arisen during the interviews, but it was not systematically categorized. None of the potentially sensitive information can be used to identify participants.

Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

    - The audio files came from fieldwork recordings conducted directly with interviewees. The transcripts of those audio files have undergone two primary processing steps: initial transcription prior to upload to APLS and the generation of annotations upon upload. About half of transcriptions were created with the assistance of predictive speech-technology tools: CLOx (Wassink et al. 2018) or Batchalign (Liu et al. 2023) for speech annotation, pyannote (Bredin 2023) for turn-segmentation. Trained transcribers always hand-checked and corrected any predictive outputs. Details about the generated annotations are given in "Preprocessing/cleaning/labeling" below.
    - Metadata on participants was reported directly by participants. Metadata on transcripts was coded by fieldworkers.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?
    - See "Preprocessing/cleaning/labeling" below.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
    - The APLS dataset is a subset of the audio files from the original sociolinguistic fieldwork. The APLS dataset contains specifically the audio files from interviewees who were natives of the Pittsburgh area that consented to make their data publicly available.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
    - Three fieldworkers (one professor and two graduate students) were involved in the collection of the initial recordings. Fourteen transcribers (one professor, one graduate student, and twelve undergraduate students) were involved in transcribing the recordings.
    - Undergraduate transcribers were compensated at a rate of $15/hour; some also received 2 credits toward their Linguistics major or minor. The graduate transcriber was compensated at a rate of $35/hour.
- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
    - The interviews were conducted between 2003 and 2005. The transcriptions and annotations were created between 2021 and 2026.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

    - The initial collection of interviews was reviewed and approved by the Institutional Review Board of Carnegie Mellon University.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
    - Data was collected directly from the interviewees in the transcripts.
- **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
    - Individuals were notified of data collection and had to sign consent forms before being interviewed. The Informed Consent form is available at https://github.com/djvill/APLS/blob/main/files/data-collection/Neighborhood_studies_consent_form.doc.
- **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
    - Yes, individuals were notified of the collection of their data and provided consent to have their data be shared and used publicly. The Informed Consent form is available at https://github.com/djvill/APLS/blob/main/files/data-collection/Neighborhood_studies_consent_form.doc.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
    - Individuals were able to indicate optional consent to have their audio recordings used for certain uses and/or saved indefinitely in the archives of the University of Pittsburgh. No mechanism was provided for consenting individuals to revoke their consent at a future date. More information can be found in the Informed Consent form at https://github.com/djvill/APLS/blob/main/files/data-collection/Neighborhood_studies_consent_form.doc.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
    - N/A

Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of**

instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- Preprocessing: 14 minutes' worth of audio in the audio files were redacted based on transcribers' judgements of information that could uniquely identify the speaker in the original transcription. Any instances of the N-word were also redacted. This was performed using a low-pass filter at 40 Hz.
- Labeling: Standard procedures in LaBB-CAT were used for labeling. Information on LaBB-CAT labeling processes can be found at https://labbcat.canterbury.ac.nz/#AutomaticAnnotation

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
  - Unredacted versions of the audio files are available only to members of the APLS team and will not be made public.
- **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
  - Yes. Links to software are categorized below by whether they were used for preprocessing or for labeling.
    - Preprocessing:
      - Praat: https://www.fon.hum.uva.nl/praat/
      - ELAN: https://archive.mpi.nl/tla/elan
      - CLOx: https://clox.ling.washington.edu/#/
      - Batchalign: https://github.com/TalkBank/batchalign
      - pyannote: https://github.com/pyannote/pyannote-audio
      - APLS Transcription Checker: https://djvill.shinyapps.io/apls_elan_file_checker/
    - Labeling:
      - LaBB-CAT: https://labbcat.canterbury.ac.nz/
      - LaBB-CAT custom scripts: https://github.com/nzilbb/LaBBCAT-Layer-Scripts

Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.
  - N/A
- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.
  - Published academic work that is based on the original fieldwork from which APLS data is drawn or uses the APLS dataset is listed on the APLS documentation website at https://djvill.github.io/APLS/doc/bibliography

- **What (other) tasks could the dataset be used for?**
    - The APLS dataset has been designed with the intention of being used for linguistic analyses, but the dataset is provided for free, public use for research and educational purposes. For instance, APLS data could be used for tasks related to ASR performance for Black vs. White speakers, part-of-speech tagging in spontaneous non-sentential speech, or sentiment analysis. More details about the terms of using APLS are provided on the documentation website at https://djvill.github.io/APLS/doc/terms
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
    - To avoid stereotyping bias, it is important that dataset consumers are aware that the speakers in this dataset should not be considered as being representative of all speakers who share their demographic attributes, such as age, class, race, education, neighborhood, and other categories.
    - Additionally, dataset consumers should be aware that there is room for error (human and/or mechanical) at every step of the annotation and labeling processes due to the methods used to create the dataset.
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
    - The dataset may not be used for commercial purposes. It may not be used to create technology designed for policing or law enforcement. More details about the terms of using APLS are provided on the documentation website at https://djvill.github.io/APLS/doc/terms

Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
    - Yes, the APLS dataset is publicly available online to any users who sign up for an APLS user account.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
    - The dataset is self-hosted on the APLS website: https://apls.pitt.edu/labbcat
- **When will the dataset be distributed?**
    - The dataset was first made available in October 2023.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
    - The terms of use for APLS are provided on the documentation website at https://djvill.github.io/APLS/doc/terms
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
    - No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
    - No.

Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
    - Dr. Dan Villarreal is supporting/maintaining the dataset. The dataset is hosted by the University of Pittsburgh.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
    - The APLS support team can be contacted by email at apls@pitt.edu
- **Is there an erratum?** If so, please provide a link or other access point.
    - The version history of APLS is provided on the documentation site at https://djvill.github.io/APLS/doc/version-history
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
    - The APLS dataset is currently still in "preview" versions. The dataset is subject to change in minor ways. Any changes to the dataset will be detailed on the version history page of the documentation website at https://djvill.github.io/APLS/doc/version-history
    - After the initial 1.0.0 release of APLS, updates may still be made on the basis of user contributions. See below for information about how users can extend/augment/build on/contribute to the APLS dataset.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

    - No.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
    - Older versions of APLS data will be kept by dataset maintainers, but only the most recent version of the dataset will be hosted and available on the APLS website. Older versions of the dataset can be made available upon request.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
    - Users can notify corpus maintainers about inaccuracies in the transcripts, annotations, and/or metadata by email at apls@pitt.edu. Suggestions of contributions to the dataset that would be useful are detailed on the APLS documentation website at https://djvill.github.io/APLS/doc/citing-contributing#contributing-back