Language and Linguistics Compass

WILEY

REVIEW ARTICLE OPEN ACCESS

Introducing the Archive of Pittsburgh Language and Speech, a Publicly Accessible, Richly Annotated Corpus of Sociolinguistic Interviews

Dan Villarreal[1] | Jack Rechsteiner[1] | Barbara Johnstone[2] | Scott Kiesling[1]

[1]University of Pittsburgh, Pittsburgh, Pennsylvania, USA | [2]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Correspondence: Dan Villarreal (d.vill@pitt.edu)

Received: 15 September 2025 | **Revised:** 10 February 2026 | **Accepted:** 20 February 2026

Keywords: computational sociolinguistics | corpus linguistics | open science | Pittsburgh English | research methods | sociolinguistics | speech data

ABSTRACT

The troves of speech data that have driven an increasing orientation towards large-scale methods in linguistics have been, for the most part, available only to closed teams of researchers and their collaborators. The Archive of Pittsburgh Language and Speech (APLS, https://apls.pitt.edu) is a new open data resource, consisting of nearly 46 h of audio from sociolinguistic interviews with 40 speakers of Pittsburgh English. Powered by the corpus management software LaBB-CAT, APLS interviews are richly annotated with multiple layers of linguistic information at the phrase, word, and segment level. Thanks to APLS's graphical user interface, users can access powerful tools for searching the corpus and extracting acoustic measurements with relatively few technical barriers to entry. We describe how APLS fits into the current landscape of (socio)linguistic open data, exemplify APLS's capabilities via a case study of 7137 /aʊ/ tokens, and contextualise the data both in terms of how fieldwork was carried out and Pittsburgh in the early 2000s.

1 | Introduction

As (socio)linguists, we collect far more data than we need, and we use a fraction of the data we collect (Gawne and Styles 2022). The canonical methodology in language variation and change research, the neighbourhood study, involves roughly hourlong audio-recordings of sociolinguistic interviews from dozens of speakers. These dozens of hours of speech may yield a single doctoral dissertation, a series of journal articles, a comprehensive monograph, or (rarely) outputs or activity that benefit the original speech community. This is partially structural: The nature of our research questions means we extract individual tokens from larger stretches of running speech; fieldwork allows researchers to understand their communities with greater ethnographic depth (Schilling 2013); and legal and ethical constraints limits our ability to share human-subjects data freely (Holton et al. 2022). We argue, however, that given the state of linguistics research in 2026—looming threats of deep cuts to university budgets, the increasing expectation of statistical and technical sophistication for quantitative methods (e.g., Sonderegger and Sóskuthy 2025), and persistent inequalities in

access to the professoriate (Charity Hudley et al. 2020)—a status quo that expects community fieldwork as a prerequisite to a sociolinguistics doctoral dissertation and allows the collected knowledge of dozens of community members to gather dust on a shelf is untenable.

Against this backdrop, we introduce the Archive of Pittsburgh Language and Speech (APLS, pronounced *apples*), an open data resource for (socio)linguistics research: https://apls.pitt.edu. APLS contains recordings of sociolinguistic interviews conducted with speakers native to four Pittsburgh-area neighbourhoods; transcripts annotated with information at the level of the phrase, word, and speech segment, which allows these interviews to be used as structured linguistic data; and metadata that facilitate large-scale (socio)linguistic analysis. As of the time of writing, APLS consists of nearly 46 h of audio from 40 interviewees, consisting of over 535,000 word tokens and 1,200,000 force-aligned segments. The data come from standard sociolinguistic interviews with native Pittsburgh English speakers conducted in 2003–2005, some of which have led to well-known findings in sociolinguistics (Bloomquist and Gooden 2015; Eberhardt 2008, 2009a, 2009b, 2009c, 2012; Gooden and Eberhardt 2007; Johnstone 2007, 2013; Johnstone et al. 2015; Johnstone 2018; Johnstone et al. 2006; Johnstone and Kiesling 2008). APLS is a transformation of this original data with the help of the corpus management software LaBB-CAT (Fromont and Hay 2012).

In this article, we first give an overview of the current landscape of (socio)linguistic open data. We then describe how APLS works and what it can do. Finally, we contextualise the data both in terms of how fieldwork was carried out and Pittsburgh in the early 2000s—including why this particular time period in Pittsburgh's history makes it so ripe for sociolinguistic study.

2 | Open Data in (Socio)linguistics

Most data in (socio)linguistics, historically and currently, is closed data (Garellek et al. 2020; Gawne and Styles 2022).[1] Some closed datasets require a fee for access, such as the Switchboard corpus (Godfrey et al. 1992), available from the Linguistic Data Consortium for a one-time $3000 fee (https://catalog.ldc.upenn.edu/LDC97S62). Other datasets are available on a 'freemium' model, where partial access is available for free, but full access is not. For example, the Corpus of Contemporary American English (COCA; Davies 2008) limits free users to 20 searches per day, and premium users ($39.95/year) to 200 searches per day (https://www.english-corpora.org/premium.asp). Most often, however, closed data is closed simply because it is not shared widely beyond the research team that originally collected or created it, or their collaborators, such as the Philadelphia Neighbourhood Corpus (Labov et al. 2013). Gawne and Styles (2022, 15) call this the 'shoebox of tapes on the shelf' problem: 'Linguists ... often collect far more data than they include in their published articles. ... Each data source contains vastly more information than the narrow range of features it was originally evaluated for, and these details may be of interest to generations of future linguists'. We thus define *open data* as that which any user can access without requiring a monetary fee or a private invitation (though potentially requiring, at most, a free user account). In between open and closed data are *semi-open* resources such as Sydney Speaks, for which 'access is managed via an application process' open to members of the research public (Travis 2024, 176). Advocates argue that open data encourages transparent and reproducible methods, situating open data within the broader open science movement (Berez-Kroeker et al. 2018; Casillas et al. 2025; Garellek et al. 2020; Gawne and Styles 2022; Sönning and Werner 2021).

Linguistics encompasses a wide variety of methodological traditions and practices, and the data that different subfields marshal as evidence can also vary widely (Good 2022). This is true within sociolinguistics itself; the data required to carry out conversation analysis (Hoey and Raymond 2022) can be quite different from data for sociophonetics (Fridland and Kendall 2022; Grama 2022; Sonderegger et al. 2022). Nevertheless, a few criteria stand out as making data suitable for the broadest sweep of research questions in sociolinguistics and language variation: spoken or signed language, produced spontaneously/naturalistically, and accompanied by metadata about both speaker demographics and recording context (to situate language practices within social practices and diachronic change). Thus, sociolinguists are unlikely to turn to open linguistics datasets like PHOIBLE (Moran and McCloy 2019), most TalkBank corpora (MacWhinney 2025), and most NLTK corpora (Bird and Loper 2004). In recent years, however, some open data resources have been created with sociolinguistics specifically in mind, such as the Corpus of Regional African American Language (CORAAL; Kendall and Farrington 2023) and the Scotland component of the International Corpus of English (ICE-Scotland; Schützler et al. 2017). CORAAL, for example, contains 159 h of sociolinguistic interviews that are orthographically transcribed and time-aligned at the word and segment level, all freely available to be browsed online and downloaded. Representing a middle ground are openly shared derived datasets, such as over 500,000 vowel formant

measurements from the SPADE project (https://osf.io/4jfrm/overview), which allow researchers to answer a more limited set of research questions based on the data that is shared.

Unfortunately, the *availability* of a data resource does not guarantee that it will be readily or easily *usable* by a wide swath of researchers (Gawne and Styles 2022; Janda 2022; Villarreal and Collister 2024). A researcher who discovers a digitized 'shoebox of tapes' of sociolinguistic interviews on the open internet will still need to expend a tremendous amount of research labour before any data analysis can happen. The challenge is partially due to the nature of speech data, which typically needs to be annotated (e.g., orthographically transcribed) before it can be transformed into structured data. However, even well-organised annotated datasets often require further technical skill (often programming skill) and resources on the part of the user (Fromont and Hay 2008; Sonderegger et al. 2022). CORAAL offers Praat TextGrids aligned at the word and segment level, so extracting vowel measurements requires skill in writing Praat scripts, not to mention computing power to run those scripts over 159 h of audio; ICE-Scotland offers word, segment, and part-of-speech annotations, split across plain-text and XML files. These ease-of-use barriers can have negative downstream effects on theory, since they open the door to 'researcher degrees of freedom' that harm reproducibility (Coretta et al. 2023; Sonderegger et al. 2022). Corpus management software like LaBB-CAT (Fromont and Hay 2012) and ISCAN (Sonderegger et al. 2022) can substantially mitigate these barriers; to date, however, corpora that use these systems, such as ONZE (Gordon et al. 2007) and SPADE (Stuart-Smith et al. 2017), respectively, are accessible only to closed teams of researchers or via an application form.

Finally, while open data has many benefits, one risk is losing sight of the fundamental fact that language is embedded in, and emanates from, communities of speakers. As Holton et al. (2022, 56) remind us, 'linguistic data cannot be divorced from their sources'. To that end, after we describe APLS design and capabilities in the next section, the final two sections of this article discuss the original data collection for the interviews that comprise APLS and the social context of Pittsburgh in the early 2000s.

3 | APLS Design and Capabilities

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. To that end, the development of APLS has focused on two key aspects. First, APLS displays the contents of the corpus in a way that is understandable and searchable. Second, APLS makes it easy to conduct linguistic analyses, whether someone is a veteran of the field or a student in an Introduction to Linguistics course. Much of this is thanks to the LaBB-CAT corpus management software that powers APLS (Fromont and Hay 2012).

APLS is, and always will be, free to use. Users must sign up for an account and agree to the terms of use (https://djvill.github.io/APLS/doc/terms) in order to access APLS: https://djvill.github.io/APLS/sign-up. After the request has been reviewed by an APLS administrator, the user is sent their login information. Unlike with semi-open resources discussed above (e.g., Travis 2024), this is a brief registration form (rather than an application form) that exists only to prevent automated bots from being able to access APLS data; as of the time of writing, all requests have been approved. The APLS terms of use include broad permissions for research and educational uses but prohibit commercial uses or uses 'to create technology designed for policing or law enforcement' (see also Endnote[2]).

One of the first options that you see upon logging into APLS is a link to APLS 101. This takes the user to the online APLS Documentation website—which does not require an APLS login to access—that has been developed in tandem with APLS: https://djvill.github.io/APLS/ (Villarreal and Rechsteiner 2026). While the current article is meant to introduce APLS briefly, the documentation site provides more in-depth tutorials, explanations, and reference guides for all the various features of APLS. In addition to the links to the documentation, there are three primary paths that users can take for exploring APLS: browsing transcripts, browsing participants, and conducting searches. We'll quickly go through each of these exploration paths.

APLS's Transcripts page presents the transcripts in a list with separate columns for selected transcript *attributes* (metadata): the transcript's name, type (sociolinguistic interview section), neighbourhood, and duration.[3] Users can filter the transcripts that are listed based on their attributes. For example, Figure 1 displays the results of a filter for only reading passage transcripts from the Forest Hills and Hill District neighbourhoods that are over 100 s long.[4] The

Transcripts page also presents options for exporting listed transcripts as transcript files, audio files, and attribute information files.

The Transcripts page is also a jumping-off point for viewing individual Transcript pages, where you can listen to and view the content of the transcript. Each transcript in APLS is composed of various *layers*, which are the organizational systems for *annotations* (transcript data) in APLS.[5] Figure 2 shows a few lines from transcript CB01interview3.eaf to demonstrate how APLS displays annotations within a transcript. Layers allow annotations to vary in size and alignment, and the integration of layers is what allows APLS to be a powerful tool for searching and analysing specific tokens of

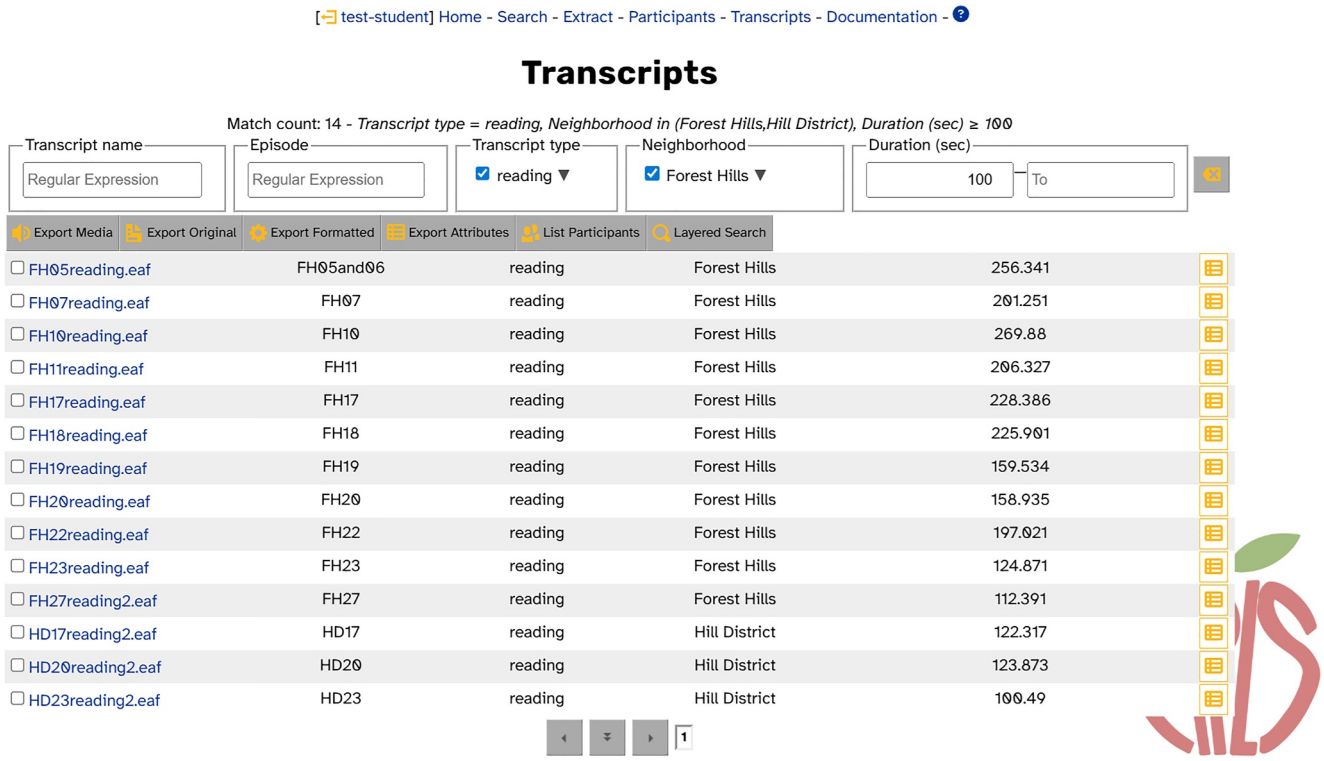

FIGURE 1 | The APLS transcripts page, filtered for reading passage transcripts from the Forest Hills and Hill District neighbourhoods that are over 100 s long.

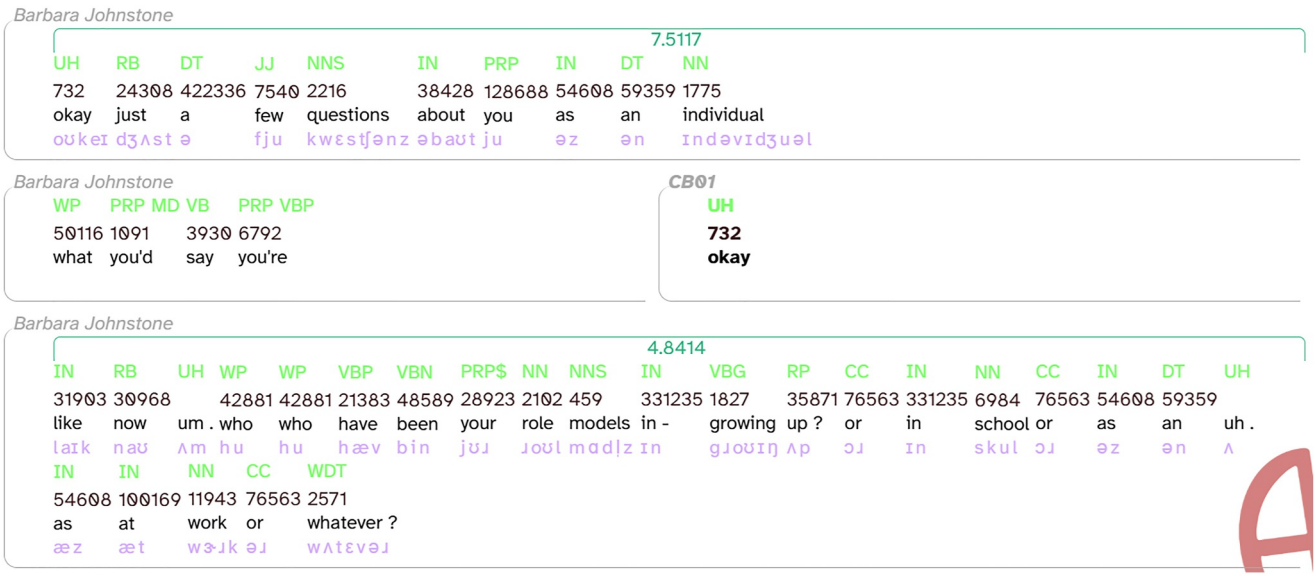

FIGURE 2 | The first few lines from the transcript CB01interview3.eaf, with annotations on several layers: speech_rate (coloured teal), part_of_speech (light green), frequency_from_celex (chocolate), word (black), and segment (lavender).

interest. Annotations in APLS can be thought of as similar to annotations you might make on a printed transcript by hand, where you would mark up the transcript at the sentence, word, or individual sound level, and all these different marks would exist concurrently. These different levels are referred to as a layer's *scope* in APLS, which group layers together based on shared properties to make them more navigable for users. Having annotations separated into different layers also helps the Transcript pages to be easily parsed. A user can toggle layers to be displayed or hidden depending on the content they are interested in, such as phonemes, part of speech tags, occurrences of overlap, and much more.

Like the Transcripts page, the Participants page allows users to view participant attributes and filter participants based on attributes like speaker gender, birth year, level of education, and neighbourhood. Participants can be filtered by any combination of these attributes, which lets users find groups of speakers that meet certain criteria. Each participant also has an attributes page that contains more fine-grained information, such as occupation type, ethnicities, and childhood neighbourhood.

The final way to navigate the APLS corpus is by searching. Searches can be performed on all the data in APLS, or searches can target specific transcripts or participants. Searching in APLS combines the robust annotations of different layers with the flexibility of regular expressions. As we demonstrate in the following section (Figure 3), the Search page makes it possible to perform searches for complex environments. After finding data of interest, APLS provides a wealth of options to easily export data (Figure 4). The Participants page can export participant attributes as a comma-separated values (.csv) file; Transcript pages can export the transcripts in a wide variety of file formats such as Praat .TextGrid and Elan .eaf (Boersma and Weenink 2025; Sloetjes and Wittenburg 2008), as well as the accompanying audio files and transcript attributes; and Search results pages can export annotation information as a .csv in a 'tidy data' (Wickham 2014) format with one row per matching token, as well as audio clips and/or formatted transcript files of just the utterances that contain matching tokens (Figure 5).

After exporting a results .csv from a search, APLS can further facilitate analyses on the Process with Praat page. As we demonstrate in the following section, this page makes it easy to use Praat (Boersma and Weenink 2025) to measure APLS interviews because it accesses data directly from the corpus without the user needing to download any audio files (Figure 6). The Process with Praat page is streamlined for more common phonetic measurements (formants, pitch, intensity, and centre of gravity); the page also natively supports the Fast Track Praat plugin for more accurate formant measurement (Barreda 2021). Users who need different measurements or additional functionality can use Process with Praat to run custom Praat scripts; this means that any script you have that runs in Praat will run in APLS's in-browser Praat (with a few tweaks to make the script compatible with APLS conventions).

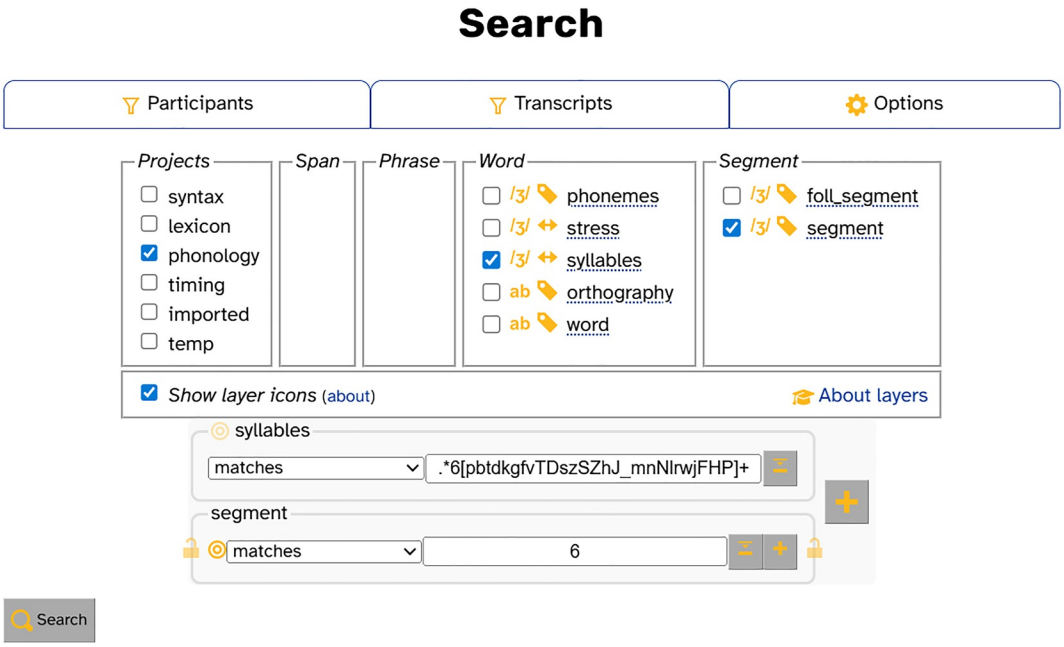

FIGURE 3 | A search in APLS that finds all instances of /aʊ/ when it is followed by at least one consonant in the same syllable.

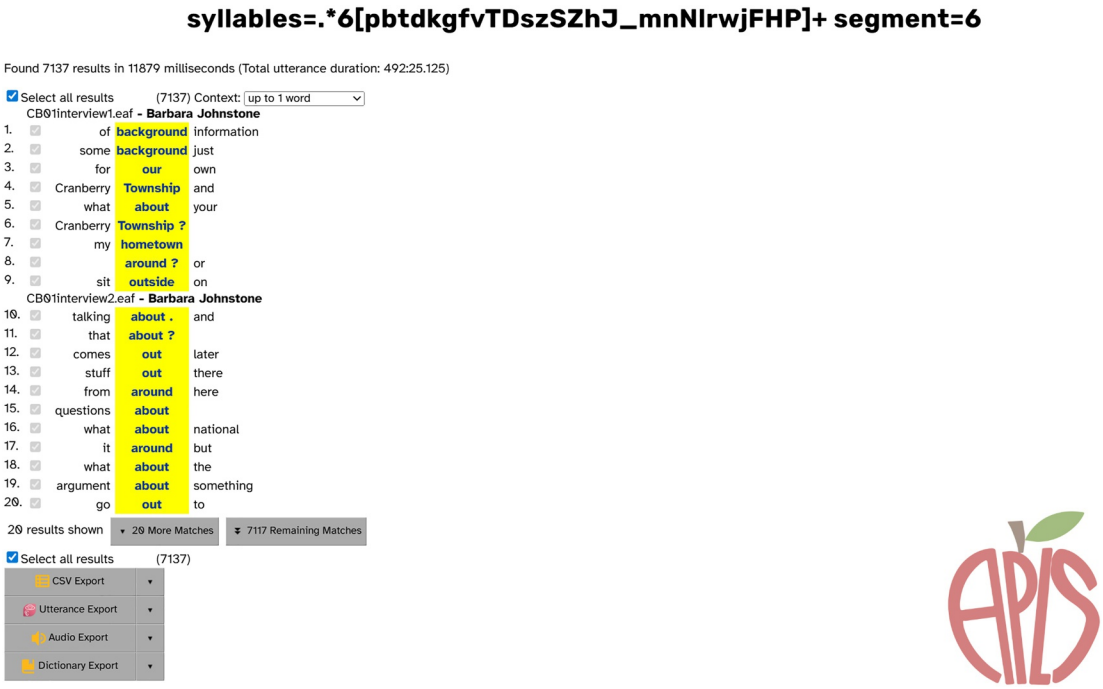

FIGURE 4 | The APLS Search results page displaying the first 20 matches for instances of /aʊ/ in closed syllables.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Title	DataVersi	SearchNa	Number	Transcript	Speaker	Line	LineEnd	MatchId	URL	Before Ma	Text	After Matc	word	word start	word end	syllables	syllables s	syllables (	segment	segment s	segment end
2	APLS: the	0.4.3	syllables=	1	CB01inten	Barbara J	4.359	12.529	g_352;em_	https://ap	of	backgrou	informatic	backgrou	6.019	6.269	'gr6nd	6.109	6.269	6	6.169	6.269
3	APLS: the	0.4.3	syllables=	2	CB01inten	Barbara J	20.86	29.431	g_352;em_	https://ap	some	backgrou	just	backgrou	26.87	27.29	'gr6nd	27.07	27.29	6	27.16	27.22
4	APLS: the	0.4.3	syllables=	3	CB01inten	Barbara J	20.86	29.431	g_352;em_	https://ap	for	our	own	our	27.73	27.83	'6r	27.73	27.83	6	27.73	27.76
5	APLS: the	0.4.3	syllables=	4	CB01inten	Barbara J	33.051	36.396	g_352;em_	https://ap	Cranberry	Township	and	Township	34.331	34.841	't6n	34.331	34.561	6	34.411	34.481
6	APLS: the	0.4.3	syllables=	5	CB01inten	Barbara J	238.681	239.999	g_352;em_	https://ap	what	about	your	about	239.211	239.331	'b6t	239.241	239.331	6	239.271	239.301
7	APLS: the	0.4.3	syllables=	6	CB01inten	Barbara J	262.537	265.795	g_352;em_	https://ap	Cranberry	Township	?	Township	265.347	265.767	't6n	265.347	265.557	6	265.417	265.517
8	APLS: the	0.4.3	syllables=	7	CB01inten	Barbara J	292.018	294.228	g_352;em_	https://ap	my	hometown		hometowr	293.598	294.078	't6n	293.808	294.078	6	293.898	294.018
9	APLS: the	0.4.3	syllables=	8	CB01inten	Barbara J	497.927	498.304	g_352;em_	https://apls.pitt.edu	around ?	or	around ?		497.947	498.267	'r6nd	497.977	498.267	6	498.047	498.187
10	APLS: the	0.4.3	syllables=	9	CB01inten	Barbara J	891.901	897.676	g_352;em_	https://ap	sit	outside	on	outside	896.511	896.661	't6t	896.511	896.571	6	896.511	896.541
11	APLS: the	0.4.3	syllables=	10	CB01inten	Barbara J	119.593	122.137	g_353;em_	https://ap	talking	about .	and	about .	120.503	120.723	'b6t	120.533	120.723	6	120.563	120.693
12	APLS: the	0.4.3	syllables=	11	CB01inten	Barbara J	205.334	205.848	g_353;em_	https://ap	that	about ?		about ?	205.514	205.744	'b6t	205.544	205.744	6	205.574	205.714
13	APLS: the	0.4.3	syllables=	12	CB01inten	Barbara J	240.621	241.834	g_353;em_	https://ap	comes	out	later	out	240.821	241.021	't6t	240.821	241.021	6	240.821	240.981
14	APLS: the	0.4.3	syllables=	13	CB01inten	Barbara J	259.837	260.947	g_353;em_	https://ap	stuff	out	there	out	260.547	260.687	't6t	260.547	260.687	6	260.547	260.647
15	APLS: the	0.4.3	syllables=	14	CB01inten	Barbara J	324.142	325.739	g_353;em_	https://ap	from	around	here	around	324.282	324.582	'r6nd	324.312	324.582	6	324.392	324.482
16	APLS: the	0.4.3	syllables=	15	CB01inten	Barbara J	354.538	360.357	g_353;em_	https://ap	questions	about		about	360.068	360.328	'b6t	360.108	360.328	6	360.158	360.258
17	APLS: the	0.4.3	syllables=	16	CB01inten	Barbara J	448.85	451.964	g_353;em_	https://ap	what	about	national	about	449.18	449.39	'b6t	449.24	449.39	6	449.28	449.36
18	APLS: the	0.4.3	syllables=	17	CB01inten	Barbara J	485.661	487.257	g_353;em_	https://ap	it	around	but	around	486.951	487.221	'r6nd	487.021	487.221	6	487.081	487.161
19	APLS: the	0.4.3	syllables=	18	CB01inten	Barbara J	489.942	494.624	g_353;em_	https://ap	what	about	the	about	493.162	493.412	'b6t	493.212	493.412	6	493.262	493.342
20	APLS: the	0.4.3	syllables=	19	CB01inten	Barbara J	621.492	624.156	g_353;em_	https://ap	argument	about	something	about	622.262	622.472	'b6t	622.292	622.472	6	622.342	622.442
21	APLS: the	0.4.3	syllables=	20	CB01inten	Barbara J	714.653	717.42	g_353;em_	https://ap	go	out	to	out	716.863	717.003	't6t	716.863	717.003	6	716.863	716.973
22	APLS: the	0.4.3	syllables=	21	CB01inten	Barbara J	769.367	773.032	g_353;em_	https://ap	places	around	here	around	770.417	770.697	'r6nd	770.487	770.697	6	770.537	770.607
23	APLS: the	0.4.3	syllables=	22	CB01inten	Barbara J	816.063	817.522	g_353;em_	https://ap	just	around	here	around	817.273	817.483	'r6nd	817.323	817.483	6	817.373	817.423
24	APLS: the	0.4.3	syllables=	23	CB01inten	Barbara J	818.94	823.492	g_353;em_	https://ap	go	out	for	out	823.14	823.28	't6t	823.14	823.28	6	823.14	823.25
25	APLS: the	0.4.3	syllables=	24	CB01inten	Barbara J	903.262	905.801	g_353;em_	https://ap	been	downtowr	on	downtowr	904.822	905.282	'd6n	904.822	905.042	6	904.872	905.012
26	APLS: the	0.4.3	syllables=	25	CB01inten	Barbara J	903.262	905.801	g_353;em_	https://ap	been	downtowr	on	downtowr	904.822	905.282	't6n	905.042	905.282	6	905.142	905.232
27	APLS: the	0.4.3	syllables=	26	CB01inten	Barbara J	957.896	958.145	g_353;em_	https://apls.pitt.edu	around			around	957.916	958.106	'r6nd	957.946	958.106	6	957.996	958.026
28	APLS: the	0.4.3	syllables=	27	CB01inten	Barbara J	992.763	995.977	g_353;em_	https://ap	next	town	over	town	995.443	995.663	't6n	995.443	995.663	6	995.503	995.623
29	APLS: the	0.4.3	syllables=	28	CB01inten	Barbara J	1014.442	1015.996	g_353;em_	https://ap	no	downtowr	Cranberry	downtowr	1014.942	1015.322	'd6n	1014.942	1015.132	6	1014.982	1015.102
30	APLS: the	0.4.3	syllables=	29	CB01inten	Barbara J	1014.442	1015.996	g_353;em_	https://ap	no	downtowr	Cranberry	downtowr	1014.942	1015.322	't6n	1015.132	1015.322	6	1015.222	1015.282
31	APLS: the	0.4.3	syllables=	30	CB01inten	Barbara J	1021.783	1024.603	g_353;em_	https://ap	a	town	before	town	1022.883	1023.253	't6n	1022.883	1023.253	6	1023.033	1023.203
32	APLS: the	0.4.3	syllables=	31	CB01inten	Barbara J	1125.45	1128.108	g_353;em_	https://ap	what	about	sports ?	about	1125.56	1125.77	'b6t	1125.59	1125.77	6	1125.62	1125.73
33	APLS: the	0.4.3	syllables=	32	CB01inten	Barbara J	0.339	2.524	g_354;em_	https://ap	questions	about	you	about	1.339	1.629	'b6t	1.369	1.629	6	1.429	1.599
34	APLS: the	0.4.3	syllables=	33	CB01inten	Barbara J	409.175	413.831	g_354;em_	https://ap	find	out	whether	out	410.395	410.525	't6t	410.395	410.525	6	410.395	410.495
35	APLS: the	0.4.3	syllables=	34	CB01inten	Barbara J	478.696	486.422	g_354;em_	https://ap	changed	about	Pittsburgh	about	481.866	482.106	'b6t	481.896	482.106	6	481.946	482.076

FIGURE 5 | The exported .csv of the search results for /aʊ/ in closed syllables in Microsoft Excel.

In addition to the browser-based graphical user interface (GUI), users have the option of accessing APLS via LaBB-CAT packages for popular programming languages: `nzilbb.labbcat` for R, `nzilbb-labbcat` for Python (Fromont 2025a, 2025b). Like the browser-based GUI, these packages can be used to search the corpus, export layer data, run Praat scripts, etc. They have the added benefit of greater reproducibility, since (e.g.) a particular set of search criteria can be encoded in R/Python code rather than described for copy and paste into the browser. Moreover, these interfaces are interoperable; for example, a user can download search results to CSV using the browser GUI, then import the data into R and use the R package to download audio excerpts containing each of the search results' utterances.

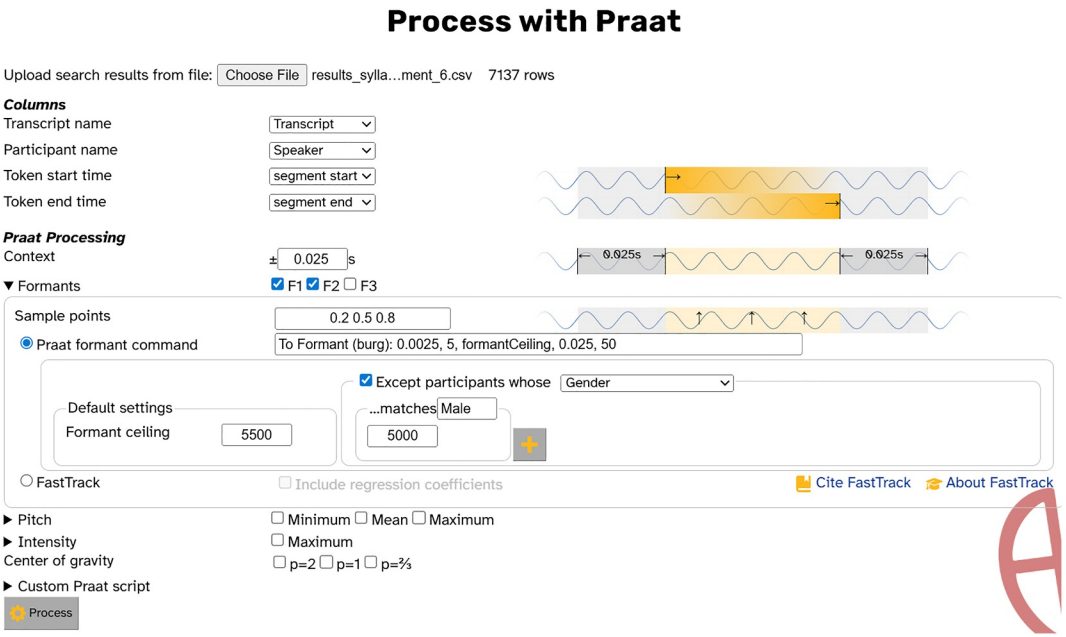

FIGURE 6 | The process with Praat page after a results .csv has been uploaded and formant sample points have been set to 20% (0.2), 50% (0.5), and 80% (0.8).

3.1 | Case Study

In this section, we hope to make the practicality of APLS more concrete by presenting a case study of an analysis that can be done using APLS. This example will focus on the diphthong /aʊ/ in closed syllables. This diphthong is interesting to look at in Pittsburgh English because pronouncing /aʊ/ like [aː] is emblematic of Pittsburghese, with the stereotype that *downtown* is pronounced 'dahntahn' (Gooden and Eberhardt 2007; Johnstone 2013; Johnstone et al. 2015; Johnstone and Kiesling 2008). If a researcher was planning to analyse this feature on their own, they would have to conduct sociolinguistic interviews, transcribe the recordings, identify instances of the diphthong, filter the instances to match the phonetic context of a closed syllable, and write Praat scripts to extract the acoustic measurements. This would take many hours of the researcher's effort to see the results. APLS, however, makes it possible to do this analysis in three steps.

As APLS already has the interviews transcribed, aligned, and annotated, the first step to analysing the diphthong /aʊ/ in closed syllables is to find the relevant tokens by searching the corpus. Figure 3 shows how to specify this search in APLS. The *segment* layer is specified as 6, which corresponds to /aʊ/ in the CELEX DISC alphabet that APLS uses to represent phonological data (Baayen et al. 1995). The *syllables* layer is given the regular expression `.*6[pbtdkgfvTDszSZhJ_mnNIrwjFHP]+`, which searches the corpus for words containing syllables of any length (indicated by .*) where the diphthong /aʊ/ (indicated by 6) is followed by one or more of any consonant (indicated by `[pbtdkgfvTDszSZhJ_mnNIrwjFHP]+`). As of APLS version 0.4.3, this search returns 7137 time-aligned /aʊ/ tokens. (Any user of APLS can replicate this search by following these steps, but we would like to note that the number of results may change as more data is added to the corpus.)

The next step is to export the search result matches as a .csv file. Figure 4 shows the Search results page, which displays results in context and provides further options for exporting them to various file formats. This can be done after a successful search by clicking the CSV Export button on the results page and downloading the generated results file. Figure 5 shows what this .csv file looks like in Microsoft Excel. This .csv file primarily contains information about the word that the token occurs in and the timing of the token relative to the rest of the transcript; however, users can configure the CSV Export to include additional layers and transcript/participant attributes.

The final step is to upload the .csv file to the Process with Praat page. All that is required of the user after uploading a valid .csv file is to specify the measurements they would like to receive from Praat. Figure 6 shows a configuration to extract measurements for F1 and F2 at the 20%, 50%, and 80% timepoints of the vowel tokens. APLS performs the

measurements internally once the Process button is clicked, and the resulting .csv file contains F1 and F2 measurements for all 7137 tokens.

These measurements can then be imported into data analysis software like R (R Core Team 2025) for further processing and analysis. Figure 7 represents one possible visualisation of this data: /aʊ/ vowel trajectories in F1/F2 space. For this plot, we excluded speakers with fewer than 75 tokens, then we used the data pipeline recommended by Stanley (2022): we removed outliers (per speaker) and incomplete words, normalised measurements using the Atlas of North American English method (Labov et al. 2006), and calculated vowel trajectory length based on three timepoints (Farrington et al. 2018), in that order. An APLS user with a high-speed internet connection can reproduce these results in 5 minutes or less.

As a final note, the ease with which users can perform this sort of analysis brings a risk beyond the previously discussed risk of divorcing linguistic data from their sources (Holton et al. 2022, 56). Ease of use is no substitute for knowing how to use tools like formant-tracking, code for outlier removal or normalisation, and other methods properly. (Indeed, a visual inspection of Figure 7 suggests that several of the trajectories are highly likely to be mismeasurements, despite these methods.). While there is no way for APLS to ensure that users can use it responsibly, we suggest several correctives. First, users should make use of the extensive APLS Documentation website (Villarreal and Rechsteiner 2026) and similar resources for understanding other methods, to avoid treating APLS like a 'black box' whose results are accepted uncritically. Second, users should make their data and methods visible so reviewers and readers can reproduce and audit their workflows.[2] Finally, users, manuscript reviewers, and readers should not assume that APLS is infallible; as with any dataset created by humans, APLS is prone to human error. To that end, the next section discusses how data get into APLS and our protocol for users to suggest corrections and/or extensions to APLS.

3.2 | The APLS Data Workflow

The transcripts in APLS have undergone two primary processing steps: initial transcription prior to upload and the generation of layer data upon upload. All transcriptions were created by transcribers (mostly undergraduate research assistants) who received extensive training, including several rounds of practice transcriptions with feedback. Once transcribers were fully trained, their transcriptions were not checked for accuracy; however, all Black speakers' transcriptions were checked by at least one trained transcriber who was familiar with African American Language and corrected where necessary. Transcriptions were newly created for the APLS project between 2021 and 2025. About half

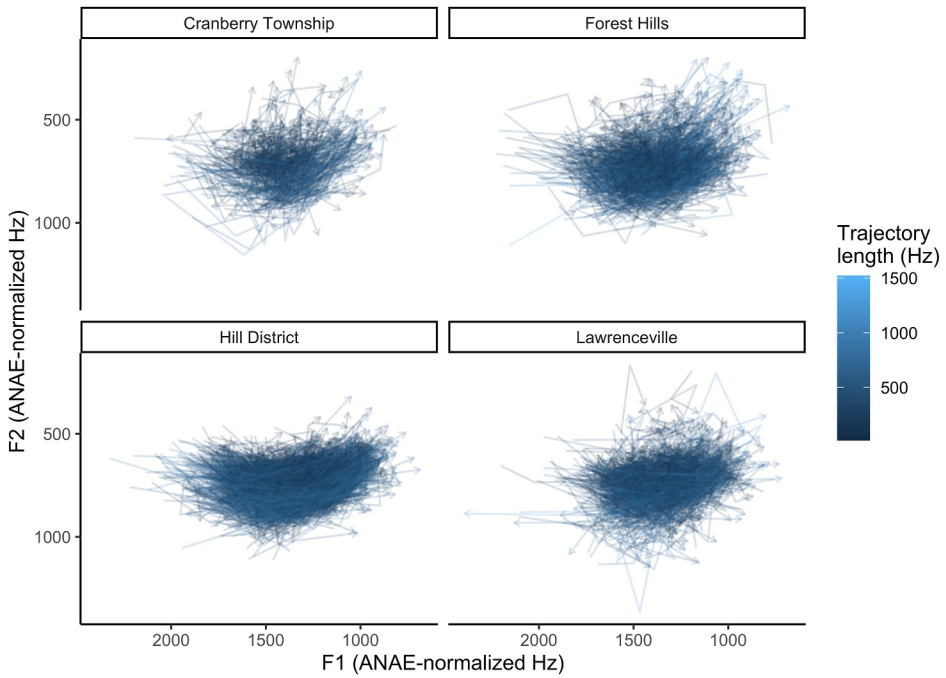

FIGURE 7 | ANAE-normalised F1 and F2 midpoint measurements for /aʊ/ tokens after removing outliers and incomplete words. See https://github.com/jackrechsteiner/introducing-APLS-case-study for plotting code.

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. Because the primary goal of transcription was to create data amenable to processing by LaBB-CAT (Fromont and Hay 2012), the APLS transcription convention is mostly orthographic (https://djvill.github.io/APLS/doc/transcription-convention).

Once a transcription and its corresponding audio file were uploaded to APLS (along with transcript/participant attributes), layer data was generated via LaBB-CAT-internal processes. Phone alignments (i.e., the *segment* layer) were generated via LaBB-CAT's HTK Layer Manager (Fromont 2019), which uses a per-speaker train-and-align procedure with the Hidden Markov Toolkit (Young et al. 2006); Gonzalez et al. (2020, 10) found that both this procedure and the Montreal Forced Aligner (McAuliffe et al. 2017) produce 'the highest quality alignments for English sociophonetic work'. Other layers were generated via: looking up annotations in reference corpora like CELEX2 (Baayen et al. 1995), such as finding a word's lemma to generate the *lemma* layer; calculating annotations based on other annotations, such as comparing the current word's end timepoint to the following word's start timepoint in order to generate the *foll_pause* layer; and/or utilising other algorithms such as the Stanford Part-of-Speech Tagger (Toutanova et al. 2003) for the *part_of_speech* layer. For more information on how layer data was generated, readers can consult the 'layer field guide' in the APLS Documentation: https://djvill.github.io/APLS/doc/layer-field-guide.

In all, this process resulted in over 7.5 million annotations across 13 generated layers (as of APLS version 0.4.3). Given the enormity of this dataset, hand-checking these annotations at scale is not feasible. However, in order to allow APLS to evolve to better fit future users' needs, we have created a protocol for users to suggest corrections (e.g., to segmental alignments or morphemic parses) and/or extensions (e.g., manual codes for a sociolinguistic variable, manual categorisation of participants into local orientation based on discourses in their interviews) to existing APLS data (see https://djvill.github.io/APLS/doc/citing-contributing#contributing-back).

Finally, we do not (yet) have concrete plans to add more interviews to APLS—especially given the 'transcription bottleneck' (e.g., Bird 2020; Do et al. 2014; Shi et al. 2021; Zahrer et al. 2020)—but we are certainly interested in performing more fieldwork in the future.

4 | Data Collection

The recordings that comprise this corpus were made as part of a multi-method, multi-year project aimed at describing the speech of people in the Pittsburgh area and exploring the relationship between Pittsburgh speech and 'Pittsburghese', the local term for local speech as it is imagined by Pittsburghers. Led by Barbara Johnstone and Scott Kiesling, the Pittsburgh Speech and Society Project (PSSP) as a whole began in 2001 and continued through 2015, when Johnstone (2013) and her colleagues (Johnstone et al. 2015) pulled the project's results together in book form. The PSSP involved the sociolinguistic interviews included here, along with historical research, collection and analysis of artefacts ranging from newspaper articles and cartoons to coffee mugs and T-shirts, and participant observation as well as other ethnographic methods.

The interviews in APLS were conducted for the 'Neighbourhood Studies' phase of the PSSP, between 2003 and 2005. Johnstone and Kiesling chose four Pittsburgh-area neighbourhoods, representing different stages of Pittsburgh's history, where they thought they would find people of different social-class and ethnic backgrounds, and which might be characterised by different social networks (Milroy 1987) and degrees of affiliation with the area and the city.

- The *Hill District*, on top of the central spine between Pittsburgh's two rivers, is where immigrants first tended to settle for much of Pittsburgh's early history. There were waves of Eastern and Southern industrial workers, Jews and African Americans. More African Americans arrived later during the Great Migration. For a time in the 1930s and 1940s, the neighbourhood was known nationwide for its jazz clubs, and the *Pittsburgh Courier* was the largest-circulation newspaper for African Americans in the US. Early housing consisted of tenement-style apartment buildings with shared courtyard outhouses. By the early 2000s, these had been replaced by detached houses on small lots. After the city razed part of the neighbourhood in 1961 for a sports arena, displacing 8000 residents, and Pittsburgh's steel industry collapsed in the 1970s and 80s, the Hill District went into serious decline (Bodnar et al. 1983; Carnegie Library of Pittsburgh 1994; Glasco 1991). By the early 2000s, when the choices of

neighbourhoods were made, most commercial businesses had shut down or moved elsewhere, and there were many overgrown lots where buildings had collapsed or been demolished. The population was mostly African American and poor, though there was a small amount of newer housing attracting middle-class African Americans. Johnstone and Kiesling knew that African Americans often thought of Pittsburgh as a particularly racist city, so they expected little local affiliation.

- *Lawrenceville* is an older neighbourhood adjacent to a riverside where heavy industry (particularly steelmaking) was once concentrated (Toker 1986). In the early 2000s, many of Lawrenceville's residents were the children or grandchildren of immigrant industrial workers from Eastern Europe and Ireland. Most were white and working-class. Many lived close to family members and relatives, sometimes in the houses where they grew up. Housing in Lawrenceville consisted largely of narrow row houses along narrow streets leading up steep hills from the riverfront. Johnstone and Kiesling expected the neighbourhood to be characterised by dense, multiplex social networks and local pride.
- *Forest Hills*, 10 miles east of Pittsburgh, was founded by German settlers. For a time, coal was mined there. From 1910, Forest Hills was connected by trolley to downtown Pittsburgh and to the nearby headquarters of the Westinghouse company. Beginning in 1916 and ending in the 1970s, the borough was the site of a Westinghouse research centre, attracting engineers and scientists. Westinghouse employees comprised an estimated 90% of the borough's population by the late 1920s; Westinghouse subsidised much of the town's early residential development. After World War II, when a restricted-access highway connected downtown Pittsburgh with its eastern suburbs, Forest Hills became attractive to commuters who worked in Pittsburgh proper (Borough of Forest Hills 1997; Walter 2015). Forest Hills residents in the early 2000s were mainly white and middle-class. Older housing in Forest Hills consisted of row houses; newer housing was detached houses built largely in the second half of the 20th century. Johnstone and Kiesling expected to find looser social networks and less affiliation with Pittsburgh than in Lawrenceville.
- *Cranberry Township* is 25 miles north of Pittsburgh, at the intersection of two Interstate highways. The Cranberry area was agricultural, but by the early 2000s, shopping centres, hotels, and housing developments had replaced some of the farmland. There was still an active farming community in the area, however. Newer residents of Cranberry tended to be recent arrivals from elsewhere in the US, not natives of southwestern Pennsylvania. They commuted to Pittsburgh by car (Uram 2023). Some chose Cranberry because it is in a different county from Pittsburgh, with lower real estate taxes; others liked the new housing stock and convenient shopping. Johnstone and Kiesling expected to find loose social networks and relatively little affiliation with Pittsburgh.

In each neighbourhood, 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 neighbourhood events.

There were three field workers. All the interviews and field experiments in Lawrenceville were carried out by Johnstone, as were all the Forest Hills interviews. A white male graduate student assisted with some of the field experiments in Forest Hills. Johnstone was assisted in Cranberry Township by a white female graduate student. The Hill District interviews were carried out by a Black female graduate student. Neither Johnstone nor any of these assistants were from the Pittsburgh area. Interviews were recorded digitally as 44.1 kHz .wav files using Marantz PMD-660 recorders and Sony ECM-44B electret condenser microphones.

The interview protocol was meant to elicit a sizeable sample of the interviewee's speech and, in the process, find out about their social networks, the degree to which they were oriented to local events and objects, their personal identities, and their attitudes about Pittsburgh and Pittsburgh speech. Hill District participants were also asked some questions about African American speech. The questions were open-ended and meant to elicit conversational answers.

The researchers also conducted four field experiments meant to supplement data from the interviews with more focused production and attitudinal data[6]:

- Two short reading passages designed to elicit either local or non-local variants of a number of phonetic variables. These included /aʊ/ and /aɪ/ monophthongisation in a variety of environments, the *cot–caught* merger, the *peel–pill* merger (/i/ and /ɪ/ before /l/), /u/ and /oʊ/ fronting, /ʌ/ lowering, pre-nasal /æ/ raising and fronting, /l/ vocalisation, and epenthetic /ɹ/ before /ʃ/ ('warsh').
- A minimal pairs task which asked interviewees to read sets of two or three words and identify which ones sounded the same to them.
- A matched-guise task in which interviewees listened to a sentence read aloud twice, once with a more local variant and once with a less local one, and answered evaluative questions (Johnstone and Kiesling 2008). These tasks are not in APLS due to poor recording quality.

In all, PSSP comprised 95.1 h of interviews from 98 interviewees. APLS includes only the interviews for which participants gave their permission, in writing, to have the recordings and any transcripts archived for future research and for the historical record. The interviews with the youngest group of participants (minors at the time of recording) are also not included, nor are interviewees not originally from the Pittsburgh area.[7] APLS (as of version 0.4.3) comprises 45.7 h from 40 interviewees, and is not as demographically balanced as the original PSSP sample. All 10 interviewees from the Hill District identified as Black (17.4 h, 38.0% of the total); the remaining 30 interviewees identified as white (28.3 h, 62.0%).

5 | Pittsburgh at the Beginning of the Millennium

The PSSP captured the voices of Pittsburghers at a particular time in the history of Pittsburgh. The city at the time was still known as the Steel City or the Iron City because of its history of industrial capacity, but its population had been declining since the 1960s, and the city was only beginning to recover from the rapid, traumatic collapse of the coal and domestic steel industry in the late 1970s and 1980s. In that collapse, unemployment reached a peak of 17% and people emigrated from the city looking for work, leading to a population decline of over 8% in one decade. The decade following the year 2000 saw the end of this precipitous population loss; although the population continued to decline, the steepest declines were over (Deitrick 2015).

All the neighbourhoods in our study were impacted by this region-wide collapse, although the Hill District, Lawrenceville, and Forest Hills were hardest hit. Cranberry Township was a largely rural area which was growing rapidly in the 2000s as it shifted from its original rural character to a relatively affluent exurb. Many interviewees thus had memories of a very different Pittsburgh from before the 1970s.

At the same time, Pittsburgh's population loss was beginning to be offset by new residents arriving for jobs in the higher education and healthcare sectors, largely driven by the technology-focused Carnegie Mellon University and biomedical research at the University of Pittsburgh (Venkatu 2018). Artists of all sorts were also drawn to the urban area with a relatively low cost of living (Tierney 2014).

The Pittsburgh environment in which the interviews were performed was thus one in which there was considerable nostalgia for Pittsburgh's past as an economic (and professional sports) powerhouse, and an orientation to a changing economic base of 'eds and meds'. Interviewees were sometimes polarised on these topics. The timing of the interviews was fortuitous, although the study was not designed with this specifically in mind: It was time in which the past was still in memory and relevant for many, and the future seemed relatively bright for others.

This impacted the neighbourhoods in different ways. The Hill District was seeing little of this rejuvenation, despite its location adjacent to several universities. Lawrenceville was just beginning a process of gentrification, now in 2025 one that is going strong and nearly complete. Forest Hills had not yet seen the benefits of the new economy, and Cranberry Township was growing and changing rapidly. Pittsburgh is often described as a 'city of neighbourhoods' (cue Mister Rogers' Neighborhood theme), and the differences and segregation of these neighbourhoods, some very small geographically, underscores this character and especially their differences in the mid-2000s.

6 | Conclusion

In this article, we have described the Archive of Pittsburgh Language and Speech (APLS, https://apls.pitt.edu), a corpus of Pittsburgh sociolinguistic interviews that is free of charge (and always will be). We hope that APLS will make a valuable contribution to the landscape of (socio)linguistic open data, in three ways—and we welcome user feedback on how APLS can better meet any of these aims. First, while many (socio)linguistic data resources exist, to our knowledge, none combines openness, suitability to sociolinguistic research questions, and relatively low technical barriers to entry, quite like APLS does. Thus, we hope that APLS enables future generations of researchers access to the type of high-quality corpus data that is normally available only to a select few. Second, we hope that APLS can serve as an example of an open data resource for language and linguistics researchers who have an interest in sharing their data. Finally, APLS is more than simply a collection of data and metadata with a convenient interface. Its interviews offer a glimpse into the lives of Pittsburghers—a snapshot of an American city at a time when its economic circumstances and civic identity were at an inflection point. By sharing these interviews, we hope to facilitate a greater understanding and appreciation for these speakers, their identities, their communities, and all the richness that their speech represents.

Acknowledgements

Funding for the original data collection was provided by the Berkman Fund at Carnegie Mellon University, the Carnegie Mellon University Department of English, the University of Pittsburgh Department of Linguistics, and National Science Foundation (Collaborative Research awards BCS-0417657 and BCS-0417684). Funding and resources for the Archive of Pittsburgh Language and Speech have been provided by the Office of Research (via Momentum Funds) and Center for Research Computing at the University of Pittsburgh, and the New Zealand Institute of Language, Brain, and Behaviour at the University of Canterbury. Thanks are due to members of the Computational Sociolinguistics Lab at the University of Pittsburgh and to Robert Fromont. Most importantly, we gratefully acknowledge the speakers who have generously shared their voices with us.

Endnotes

[1] Information about dataset availability, format, and pricing is current as February 2026.

[2] The APLS Terms of Use allow users to share derived data 'so long as they do not permit readers or viewers to reconstruct substantial portions of audio data, transcripts, and/or annotations' (https://djvill.github.io/APLS/doc/terms). For example, an APLS user who carried out the case study and downloaded the .csv file depicted in Figure 5 would be permitted to post the file on a publicly accessible GitHub or OSF repository.

[3] There are additional attributes not displayed on the Transcripts page, such as recording date. For a listing of all attributes, see https://djvill.github.io/APLS/doc/attribute-reference-card.

[4] All screenshots are from APLS version 0.4.3.

[5] For a listing of all layers, see https://djvill.github.io/APLS/doc/layer-reference-card.

[6] The protocols for interviews and field experiments are available at https://github.com/djvill/APLS/tree/main/files/data-collection.

[7] The exclusions break down as follows:

- Participants who did not give permission for data to be public (including minors): 32.7 h, 44 interviewees.
- Participants not originally from the Pittsburgh area: 8.9 h, 8 interviewees.
- Participants meeting both exclusion criteria: 3.3 h, 2 interviewees.
  In addition, APLS currently excludes files with multiple interviewees, since these are more difficult to transcribe (save for a pair of Cranberry Township interviewees, to round out low numbers of Cranberry Township speakers). This excludes 4.5 h of data and 4 interviewees.

References

Baayen, H., R. Piepenbrock, and H. van Rijn. 1995. *The CELEX Lexical Database (Release 2, CD-ROM), LDC Catalogue No.: LDC96L14*. University of Pennsylvania, Linguistic Data Consortium.

Barreda, S. 2021. "Fast Track: Fast (Nearly) Automatic Formant-Tracking Using Praat." *Linguistics Vanguard* 7, no. 1: 20200051. https://doi.org/10.1515/lingvan-2020-0051.

Berez-Kroeker, A. L., L. Gawne, S. S. Kung, et al. 2018. "Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field." *Linguistics* 56, no. 1: 1–18. https://doi.org/10.1515/ling-2017-0032.

Bird, S. 2020. "Sparse Transcription." *Computational Linguistics* 46, no. 4: 713–744. https://doi.org/10.1162/coli_a_00387.

Bird, S., and E. Loper. 2004. "NLTK: The Natural Language Toolkit." In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Association for Computational Linguistics. https://aclanthology.org/P04-3031/.

Bloomquist, J., and S. Gooden. 2015. "African American Language in Pittsburgh and the Lower Susquehanna Valley." In *The Oxford Handbook of African American Language*, edited by J. Bloomquist, L. J. Green, and S. L. Lanehart, 236–255. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199795390.013.35.

Bodnar, J., R. Simon, and M. P. Weber. 1983. *Lives of Their Own: Blacks, Italians, and Poles in Pittsburgh, 1900–1960*. The Working Class in American History. University of Illinois Press. https://www.press.uillinois.edu/books/?id=p010637.

Boersma, P., and D. Weenink. 2025. "Praat." Version 6.4.50. https://praat.org/.

Borough of Forest Hills. 1997. A Brief History of Forest Hills. https://www.foresthillspa.gov/about/history.php.

Bredin, H. 2023. "pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe." In *INTERSPEECH 2023*, 1983–1987. ISCA. https://doi.org/10.21437/Interspeech.2023-105.

Carnegie Library of Pittsburgh. 1994. "The Hill District: History." *Bridging the Urban Landscape*. https://web.archive.org/web/20150807080932/http:/www.carnegielibrary.org/exhibit/neighborhoods/hill/hill_n4.html.

Casillas, J. V., G. Constantin-Dureci, I. A. Rascón, et al. 2025. "Opening Open Science to All: Demystifying Reproducibility and Transparency Practices in Linguistic Research." *Linguistics*. https://doi.org/10.1515/ling-2023-0249.

Charity Hudley, A. H., C. Mallinson, and M. Bucholtz. 2020. "Toward Racial Justice in Linguistics: Interdisciplinary Insights into Theorizing Race in the Discipline and Diversifying the Profession." *Language* 96, no. 4: e200–e235. https://doi.org/10.1353/lan.2020.0074.

Coretta, S., J. V. Casillas, S. Roessig, et al. 2023. "Multidimensional Signals and Analytic Flexibility: Estimating Degrees of Freedom in Human-Speech Analyses." *Advances in Methods and Practices in Psychological Science* 6, no. 3: 1–29. https://doi.org/10.1177/25152459231162567.

Davies, M. 2008. The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/.

Deitrick, S. 2015. "Cultural Change in Pittsburgh: A Demographic Analysis at City and County Scales." *Pennsylvania Geographer* 52, no. 5: 71–92. https://drive.google.com/file/d/1kuUW2xJAYQ5TMuo1ACru6WaxnGOZyAMe/view.

Do, T.-N.-D., A. Michaud, and E. Castelli. 2014. "Towards the Automatic Processing of Yongning Na (Sino-tibetan): Developing a 'Light' Acoustic Model of the Target Language and Testing 'Heavyweight' Models From Five National Languages." In *SLTU-2014*, 153–160. https://www.isca-archive.org/sltu_2014/do14_sltu.html.

Eberhardt, M. 2008. "The Low-Back Merger in the Steel City: African American English in Pittsburgh." *American Speech* 83, no. 3: 284–311. https://doi.org/10.1215/00031283-2008-021.

Eberhardt, M. 2009a. "The Sociolinguistics of Ethnicity in Pittsburgh." *Language and Linguistics Compass* 3, no. 6: 1443–1454. https://doi.org/10.1111/j.1749-818X.2009.00157.x.

Eberhardt, M. 2009b. *Identities and Local Speech in Pittsburgh: A Study of Regional African American English*. PhD dissertation. University of Pittsburgh. https://d-scholarship.pitt.edu/7433/.

Eberhardt, M. 2009c. "African American and White Vowel Systems in Pittsburgh." *Publication of the American Dialect Society* 94, no. 1: 129–157. https://doi.org/10.1215/-94-1-129.

Eberhardt, M. 2012. "Enregisterment of Pittsburghese and the Local African American Community." *Language & Communication* 32, no. 4: 358–371. https://doi.org/10.1016/j.langcom.2012.08.002.

Farrington, C., T. Kendall, and V. Fridland. 2018. "Vowel Dynamics in the Southern Vowel Shift." *American Speech* 93, no. 2: 186–222. https://doi.org/10.1215/00031283-6926157.

Fridland, V., and T. S. Kendall. 2022. "Managing Sociophonetic Data in a Study of Regional Variation." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, 237–247. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0023.

Fromont, R. 2019. "Forced Alignment of Different Language Varieties Using LaBB-CAT." In *Proceedings of ICPHS*, Vol. 19, 1327–1331. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1376.pdf.

Fromont, R. 2025a. nzilbb-labbcat: Client Library for Communicating With LaBB-CAT Servers Using Python. https://nzilbb.github.io/labbcat-py/.

Fromont, R. 2025b. nzilbb.labbcat: Accessing Data Stored in "LaBB-CAT" Instances. https://nzilbb.github.io/labbcat-R/.

Fromont, R., and J. Hay. 2008. "ONZE Miner: The Development of a Browser-Based Research Tool." *Corpora* 3, no. 2: 173–193. https://doi.org/10.3366/E1749503208000142.

Fromont, R., and J. Hay. 2012. "LaBB-CAT: An Annotation Store." In *Proceedings of Australasian Language Technology Association Workshop*, 113–117.

Garellek, M., M. Gordon, J. Kirby, et al. 2020. "Toward Open Data Policies in Phonetics: What We Can Gain and How We Can Avoid Pitfalls." *Journal of Speech Science* 9, no. 1: 3. https://doi.org/10.20396/joss.v9i00.14955.

Gawne, L., and S. Styles. 2022. "Situating Linguistics in the Social Science Data Movement." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, 9–25. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0006.

Glasco, L. 1991. "Double Burden: The Black Experience in Pittsburgh." In *City at the Point: Essays on the Social History of Pittsburgh*, 69–110. University of Pittsburgh Press. https://muse.jhu.edu/pub/49/edited_volume/chapter/3959384.

Godfrey, J. J., E. C. Holliman, and J. McDaniel. 1992. *SWITCHBOARD: Telephone Speech Corpus for Research and Development*, 517–520. IEEE Computer Society. https://doi.org/10.1109/ICASSP.1992.225858.

Gonzalez, S., J. Grama, and C. E. Travis. 2020. "Comparing the Performance of Forced Aligners Used in Sociophonetic Research." *Linguistics Vanguard* 6, no. 1: 20190058. https://doi.org/10.1515/lingvan-2019-0058.

Good, J. 2022. "The Scope of Linguistic Data." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, 27–47. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0007.

Gooden, S., and M. Eberhardt. 2007. "Local Identity and Ethnicity in Pittsburgh AAE." *Penn Working Papers in Linguistics* 13, no. 2: 81–94. https://repository.upenn.edu/handle/20.500.14332/44655.

Gordon, E., M. Maclagan, and J. Hay. 2007. "The ONZE Corpus." In *Creating and Digitizing Language Corpora: Volume 2: Diachronic Databases*, edited by J. C. Beal, K. P. Corrigan, and H. L. Moisl, 82–104. Palgrave Macmillan UK. https://doi.org/10.1057/9780230223202_4.

Grama, J. 2022. "Managing Legacy Data in a Sociophonetic Study of Vowel Variation and Change." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, 221–236. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0022.

Hoey, E. M., and C. W. Raymond. 2022. "Managing Conversation Analysis Data." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, 257–266. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0025.

Holton, G., W. Y. Leonard, and P. L. Pulsifer. 2022. "Indigenous Peoples, Ethics, and Linguistic Data." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. Collister, 49–60. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0008.

Janda, L. A. 2022. "Managing Data and Statistical Code According to the FAIR Principles." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, 447–452. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0042.

Johnstone, B. 2007. "Linking Identity and Dialect Through Stancetaking." In *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, edited by R. Englebretson, 49–68. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.04joh.

Johnstone, B. 2013. *Speaking Pittsburghese: The Story of a Dialect*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199945689.001.0001.

Johnstone, B. 2018. "Southern Speech With a Northern Accent: Performance Norms in an Imitation." *American Speech* 93, no. 3–4: 497–512. https://doi.org/10.1215/00031283-7271294.

Johnstone, B., J. Andrus, and A. E. Danielson. 2006. "Mobility, Indexicality, and the Enregisterment of 'Pittsburghese'." *Journal of English Linguistics* 34, no. 2: 77–104. https://doi.org/10.1177/0075424206290692.

Johnstone, B., D. Baumgardt, M. Eberhardt, and S. Kiesling. 2015. *Pittsburgh Speech and Pittsburghese*. De Gruyter Mouton. https://doi.org/10.1515/9781614511786.

Johnstone, B., and S. F. Kiesling. 2008. "Indexicality and Experience: Exploring the Meanings of/aw/-monophthongization in Pittsburgh." *Journal of SocioLinguistics* 12, no. 1: 5–33. https://doi.org/10.1111/j.1467-9841.2008.00351.x.

Kendall, T., and C. Farrington. 2023. *The Corpus of Regional African American Language*. Online Resources for African American Language Project. https://doi.org/10.7264/1ad5-6t35.

Labov, W., S. Ash, and C. Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Mouton de Gruyter.

Labov, W., I. Rosenfelder, and J. Freuhwald. 2013. "One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis." *Language* 89, no. 1: 30–65. https://doi.org/10.1353/lan.2013.0015.

Liu, H., B. MacWhinney, D. Fromm, and A. Lanzi. 2023. "Automation of Language Sample Analysis." *Journal of Speech, Language, and Hearing Research* 66, no. 7: 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642.

MacWhinney, B. 2025. "Understanding Language Through TalkBank." *Current Directions in Psychological Science* 34, no. 2: 75–81. https://doi.org/10.1177/09637214241304345.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner & M. Sonderegger. 2017. "Montreal Forced Aligner: Trainable text-speech Alignment Using Kaldi." In *Proceedings of 18th Interspeech*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386.

Milroy, L. 1987. *Language and Social Networks* (Language in Society). 2nd ed. Basil Blackwell. https://www.wiley.com/en-ca/Language+and+Social+Networks%2C+2nd+Edition-p-9780631153146.

Moran, S., and D. McCloy. 2019. PHOIBLE. https://phoible.org/.

R Core Team. 2025. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Schilling, N. 2013. *Sociolinguistic Fieldwork* (Key Topics in Sociolinguistics). Cambridge University Press.

Schützler, O., U. Gut, and R. Fuchs. 2017. *New Perspectives on Scottish Standard English: Introducing the Scottish Component of the International Corpus of English*, edited by J. C. Beal and S. Hancil, 273–302. De Gruyter Mouton. https://doi.org/10.1515/9783110450903-012.

Shi, J., J. D. Amith, R. C. García, E. G. Sierra, K. Duh, and S. Watanabe. 2021. "Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, edited by P. Merlo, J. Tiedemann, and R. Tsarfaty, 1134–1145. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.96.

Sloetjes, H., and P. Wittenburg. 2008. "Annotation by Category – ELAN and ISO DCR." In *6th International Conference on Language Resources and Evaluation*.

Sonderegger, M., and M. Sóskuthy. 2025. "Advancements of Phonetics in the 21st Century: Quantitative Data Analysis." *Journal of Phonetics* 111: 101415. https://doi.org/10.1016/j.wocn.2025.101415.

Sonderegger, M., J. Stuart-Smith, M. McAuliffe, R. Macdonald, and T. Kendall. 2022. "Managing Data for Integrated Speech Corpus Analysis in Speech Across Dialects of English (SPADE)." In *The Open Handbook of Linguistic Data Management*, edited by A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, 195–207. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0020.

Sönning, L., and V. Werner. 2021. "The Replication Crisis, Scientific Revolutions, and Linguistics." *Linguistics* 59, no. 5: 1179–1206. https://doi.org/10.1515/ling-2019-0045.

Stanley, J. A. 2022. "Interpreting the Order of Operations in a Sociophonetic Analysis." *Linguistics Vanguard* 8, no. 1: 279–289. https://doi.org/10.1515/lingvan-2022-0065.

Stuart-Smith, J., M. Sonderegger, and J. Mielke. 2017. Speech Across Dialects of English (SPADE): Large-Scale Digital Analysis of a Spoken Language Across Space and Time. https://spade.glasgow.ac.uk/.

Tierney, J. 2014. "How the Arts Drove Pittsburgh's Revitalization." *Atlantic*, August 13. https://www.theatlantic.com/business/archive/2014/12/how-the-cultural-arts-drove-pittsburghs-revitalization/383627/.

Toker, F. 1986. *Pittsburgh: An Urban Portrait*. Pennsylvania State University Press.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer. 2003. "Feature-Rich Part-of-Speech Tagging With a Cyclic Dependency Network." In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – NAAC '03*, Vol. 1, 173–180. Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073478.

Travis, C. E. 2024. "Sydney Speaks Corpus: An Overview." *Australian Journal of Linguistics* 44, no. 2–3: 163–181. https://doi.org/10.1080/07268602.2024.2386387.

Uram, A. 2023. "Cranberry Township: An Explosive History." *Butler Eagle*, August 13. https://www.butlereagle.com/20230912/cranberry-township-an-explosive-history/.

Venkatu, G. 2018. *Rust and Renewal: A Pittsburgh Retrospective*. White Paper. Federal Reserve Bank of Cleveland. https://www.clevelandfed.org/regional-analysis/pittsburgh-retrospective.

Villarreal, D., and L. Collister. 2024. "Open Methods: Decolonizing (or Not) Research Methods in Linguistics." In *Decolonizing Linguistics (Oxford Collections in Linguistics)*, edited by A. C. Hudley, C. Mallinson, and M. Bucholtz, 263–288. Oxford University Press. https://doi.org/10.1093/oso/9780197755259.003.0014.

Villarreal, D., and J. Rechsteiner. 2026. "Archive of Pittsburgh Language and Speech Documentation." https://djvill.github.io/APLS.

Walter, M. B. 2015. "An Unlikely Atomic Landscape." *Western Pennsylvania History* 98, no. 3: 36–49. https://journals.psu.edu/wph/article/view/60195.

Wassink, A. B., R. Squizzero, C. Fellin, and D. Nichols. 2018. Client Libraries Oxford (CLOx): Automated Transcription for Sociolinguistic Interviews. https://clox.ling.washington.edu/.

Wickham, H. 2014. "Tidy Data." *Journal of Statistical Software* 59, no. 10: 1–23. https://doi.org/10.18637/jss.v059.i10.

Young, S., G. Evermann, M. Gales, et al. 2006. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

Zahrer, A., A. Zgank, and B. Schuppler. 2020. "Towards Building an Automatic Transcription System for Language Documentation: Experiences From Muyu." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, edited by N. Calzolari, F. Béchet, P. Blache, et al., 2893–2900. European Language Resources Association. https://aclanthology.org/2020.lrec-1.353/.