RiCL Research in Corpus Linguistics

Compiling a corpus of African American Language from oral histories

Sarah Moeller[a] – Alexis Davis[a] – Wilermine Previlon[a] – Michael Bottini[a] – Kevin Tang[b/a]

University of Florida[a] / United States
Heinrich-Heine University Düsseldorf[b] / Germany

Abstract – African American Language (AAL) is a marginalized variety of American English that has been understudied due to a lack of accessible data. This lack of data has made it difficult to research language in African American communities and has been shown to cause emerging technologies such as Automatic Speech Recognition (ASR) to perform worse for African American speakers. To address this gap, the *Joel Buchanan Archive of African American Oral History* (JBA) at the University of Florida is being compiled into a time-aligned and linguistically annotated corpus. Through Natural Language Processing (NLP) techniques, this project will automatically time-align spoken data with transcripts and automatically tag AAL features. Transcription and time-alignment challenges have arisen as we ensure accuracy in depicting AAL morphosyntactic and phonetic structure. Two linguistic studies illustrate how the *African American Corpus from Oral Histories* betters our understanding of this lesser-studied variety.

Keywords – African American English; oral history; Automatic Speech Recognition; natural language processing; corpus linguistics; morphosyntax

1. Introduction[1]

African American Language[2] (henceforth AAL) is one of the most studied marginalized varieties of American English, yet AAL linguistic data that is annotated, compiled and searchable in the form of a corpus is mostly inaccessible for linguists. Only in 2018 was the first corpus of African American speech, namely the *Corpus of Regional African American Language* (CORAAL; Kendall and Farrington 2021), made available. For the first time, the general population and linguists alike could access interviews of varying African American regional accents. However, with CORAAL being the only such

[1] This research is part of the project *Reanimating African American Histories of the Gulf South* which was supported by the National Endowment for the Humanities (PW–277433–21). We thank the editors (Robbie Love and Carlos Prado-Alonso) and the anonymous reviewers for their valuable feedback.
[2] The term 'African American Language' is also often referred to as 'African American English' (AAE) and 'African American Vernacular English' (AAVE).

resource of its kind, lack of data creates an obstacle for much needed research on language and race in diverse African American communities. Further, the effect of this gap extends to emerging technologies such as automatic speech recognition (henceforth ASR) that perform more poorly for African American speakers (Blackley *et al.* 2019; Koenecke *et al.* 2020; Martin and Tang 2020). As technology expands into everyday life (Lee *et al.* 2022; Yoon *et al.* 2023; Davis *et al*. 2024), healthcare, and hiring (Martin and Wright 2022, and references therein), racial disparity in technology could have dire consequences for African Americans.

A time-aligned corpus of linguistically annotated AAL audio and transcripts will provide a much-needed increase of speech data for African American corpus linguistics. The *Joel Buchanan Archive of African American Oral History* (JBA) at the University of Florida contains a growing collection of over 600 oral history interviews.[3] This is larger than many publicly available spoken corpora of any language or dialect and could have a significant impact on AAL studies, as well as provide a tool to hold major technology developers like *Amazon* and *Apple* accountable to address their racially biased systems, (see Blodgett *et al.* 2020 for a survey of biases in natural language processing, henceforth NLP).

This paper describes the compilation and initial analysis of AAL linguistic features of a time-aligned and linguistically annotated AAL corpus from JBA. Our work aims to produce two major deliverables: 1) a large African American speech corpus enriched for the first time with linguistic annotations and 2) a time-aligned representation of the audio files and transcribed texts. We want to identify and analyze the distinctive features of AAL in hopes of advancing language science, improving NLP, increasing awareness of the richness of language in the USA, and supporting educational endeavors related to African American culture and history. We are also building NLP systems specific to AAL and adapting a forced alignment model from General American English (henceforth GAE) to the recordings and transcripts of AAL.

We provide a background by describing the nature of oral history collections, existing African American linguistic corpora, including the *Joel Buchanan Archive* that we work with, and the AAL (cf. Section 2). We provide a general outline of the compilation process from an oral history collection to corpus for linguistic study (Section

[3] https://ufdc.ufl.edu/collections/ohfb

3). We then describe the specific challenges that have arisen while compiling our corpus, detailing issues related to transcription and time-alignment (Section 4). To illustrate how such a corpus can better our understanding of a lesser-studied language variety, we include two linguistic case studies of the distribution of AAL features (Section 5.2) and the syntactic structures that signal the presence of the AAL feature habitual *be* (Section 5.3) based on an initial sample of the oral history collection described in Section 5.1. Finally, we summarize a computational method that we have developed for the tagging of linguistic features that are distinctive to AAL (Section 5.4).

2. Background

With oral history collections containing a wealth of untapped linguistic data, it is important to understand their construction and the stories they house. Additionally, our methods are also inspired by previous corpus work (cf. Kendall and Farrington 2021 or Fitzgerald 2022, among others). In this section, we describe endeavors to catalog naturalistic AAL data via corpora, as well as oral history collections. Finally, we introduce the oral history collection that we are harnessing for corpus compilation, and briefly describe the specific challenges we face with AAL.

2.1. What are oral history collections?

Oral history interviews are a method of documenting history through audio/video recorded stories of individuals. These projects are centered around the experiences of narrowly defined groups, such as first-generation college students (e.g., the *Machen Florida Opportunity Scholars Oral History Program*)[4] or people who knew or worked with important persons, (e.g., the *John F. Kennedy and Robert F. Kennedy Oral History Collection*)[5] or specific racial groups, such as African Americans. A common method of collecting is using community networks to find volunteers who are willing to share their personal stories. They may also focus on certain regions, or topics such as historical events. However, most oral history collections remain largely inaccessible to corpus linguistics, their power to enlighten through linguistic analysis lying untapped. There are however a few exceptions. Schiffrin (2002) uses an oral history transcription of one

[4] https://oral.history.ufl.edu/projects/machen-florida-opportunity-scholars-program-mfos/
[5] https://www.jfklibrary.org/archives/about-archival-collections/oral-histories

Holocaust survivor to investigate her relationship with her mother and friends. The analysis is based on linguistic construction, such as the variation in the use of referring terms and reported speech. Similarly, Fitzgerald (2022) reports on the compilation and usage of the *Corpus of Irish Historical Narratives*[6] using an archive of Irish oral history documents, the *Irish Bureau of Military History*,[7] which consists of 238 oral testimonies. By applying corpus linguistic methods, Fitzgerald investigates the commitment to truth and what is meant by truth by examining the use of a set of mental process verbs, such as *think*, *remember*, *suppose* and *believe*, and expectation markers such as *actually*, *in fact* and *of course*. Finally, the most note-worthy use of oral history collection is the *Freiburg English Dialect Corpus* (Kortmann and Wagner 2005), which contains transcriptions and audio recordings, totaling 372 interviews which comprise 2.5 million words of text and 300 hours of speech.

Compiling a corpus of linguistically annotated audio and transcripts from existing oral history collections can provide a much-needed increase of speech data for AAL. This increase of data can support more equitable language technology, specifically for ASR. African Americans are a population largely absent from focused corpus linguistics studies as well as NLP (Dacon 2022; Martin 2022; Martin and Wright 2022). African Americans have been the focus of several oral history projects.[8] Generally speaking, these projects serve as chronicles of African Americans who lived through the transatlantic slave trade, the Jim Crow era, the Civil Rights Movement, the wars of the twentieth century, and the first Black presidency. Linguists will find a wealth of conversational speech data, sociolinguistic dynamics, and phonological and morphosyntactic structures.

2.2. African American speech corpora

With AAL's marginalized status being a consistent obstacle for linguistic data collection, linguists have begun to create their own repositories of AAL data. CORAAL (Kendall and Farrington 2021) has been publishing collections of sociolinguistic interviews since 2018 (Kendall and Farrington 2022), and these projects focus on different regional varieties of AAL. CORAAL works in tandem with researchers (often African American)

[6] http://corpas.ria.ie/
[7] https://www.militaryarchives.ie/collections/online-collections/bureau-of-military-history-1913-1921/
[8] For a compiled list of oral history projects with a focus on African Americans, see https://guides.library.duke.edu/africanamericanoralhistories/collections

to make this data accessible to both linguists and the general public. Kendall and Farrington (2022: 191) have also reiterated the importance of their work and argued that their point is "that new advances and better science can be done if there are *more* public and larger data sets." As much AAL data is not readily available for linguists, let alone published in an open space for hobbyists or educators to interact with, CORAAL is pioneering a more accessible approach to disseminate linguistic research. This work has inspired one of our main objectives with this project as well.

As already stated in Section 1, the current paper deals with the compilation of a new speech corpus from an oral history collection called JBA, rather than compiling a corpus using sociolinguistic interviews like CORAAL. JBA is a large and growing collection, containing more than 600 oral history interviews with African Americans that were recorded in the state of Florida and across the Southeastern United States. The archive houses a corpus of approximately 6.5 million words and 1,100 hours of audio. It combines several different regional projects under the *Samuel Proctor Oral History Program* (SPOHP),[9] with the earliest interviews from the 1970s, and is continually updated. Interviewees consist of student participants, community elders, prominent citizens (e.g., pastors, politicians, union workers), ranging from teenage to elderly in age. Through snowball sampling, a non-probability sampling technique taken from sociological methods, where participants recommend other individuals in the community, this collection records individual life experiences to shed light on the complex histories of communities. This collection is maintained by the *University of Florida Digital Collection*[10] through George A. Smathers Libraries.[11]

2.3. AAL

AAL has a rich set of distinctive phonological and morphosyntactic features. Unique syntactic features of AAL can pose challenges for language technology, as many products are built using GAE resources, which often do not account for the differing grammatical structures that are foundational for AAL. Because AAL and GAE do have these differences, it is necessary to build language technology from AAL data. Given the limited number of repositories to make this endeavor easier, we are tasked with adding to

[9] https://oral.history.ufl.edu/
[10] https://ufdc.ufl.edu/
[11] https://ufdc.ufl.edu/collections/flaac

the data that already exists. Most of our effort to date has focused on six morphosyntactic features that are significant and frequent characteristics of AAL and differ from GAE (Green 2002) and are illustrated below. These are: a) Person/number disagreement (absence of third person singular -*s*), as in example (1); b) habitual *be* (cf. 2); c) multiple negation (cf. 3); d) remote past *bin* (cf. 4); e) existential *it/dey* (cf. 5); and f) null copula, as in (6).

(1) Saying he just want to remember.
Saying he just wants to remember.

(2) I be in my office by 7:30.
I am usually in my office by 7:30.

(3) I ain't step on no dog.
I didn't step on a dog.

(4) We been adding cinnamon to the cookies.
We've added cinnamon to the cookies for a long time now.

(5) Dey some coffee in the kitchen.
There is some coffee in the kitchen.

(6) We the county champs.
We are the county champs.

3. Compiling Linguistic Corpus from Oral History Collection

This section describes general steps to compile a linguistic corpus from a collection of oral histories. It presents issues that corpus linguists may expect to encounter. The first step is to edit the oral history transcriptions. Second, the transcriptions need to be annotated with linguistic features. Third, compiling a spoken corpus involves aligning the transcribed utterances to the timestamps of corresponding speech in audio files (Harrington 2010).

3.1. Oral histories transcriptions

Oral historians who focus on the experiences of minority communities are likely to encounter language varieties that may have distinctive features which are not accurately represented in the standard orthography. Anyone may struggle to understand another person's accent, recognize regional linguistic features, or correctly interpret

colloquialisms. How oral historians handle these features during transcription is informed by their needs. Therefore, the first step for compiling a linguistically useful corpus is to decide whether to edit the collection's transcriptions to suit the linguists' needs.

A general lack of orthographic norms for regional language varieties complicates transcription (Ghyselen *et al.* 2020). Encountering a less common language variety reduces transcription accuracy by humans. The *National Court Reporters Association*[12] (NCRA) sets a standard at 95 percent accuracy to be certified as a court reporter, yet professional and certified transcribers with thorough training and examination measured as low as 59.5 percent when transcribing African American speakers. In oral history projects, transcribers are often student volunteers and are unlikely to have the same level of training or linguistic awareness to help them hear and transcribe regional linguistic variations (see Appendix B: Data Availability).

Linguists wishing to compile a corpus from an oral history collection should carefully review the oral history program's transcription guidelines. The field of oral history has no universal transcription standards and individual programs post their own guidelines which range in expectations and details. The examination of a handful of these guidelines (Samuel Proctor Oral History Program 2007, 2016; Oregon Department of Transportation Research Section 2010; Strong *et al.* 2018; Samuel Proctor Oral History Project 2020, 2023) reveals a tendency to place priority on informational content, which is a historian's primary interest. Guidelines seem to reflect a philosophy that transcribed dialogue does not need to be completely reflective of a person's speech. For example, the *Columbia Center for Oral History Research's* (CCOHR) transcription style guide (Columbia University Center for Oral History Research 2022: 2, 15) states the following:

> The characteristics of *how* individual speakers communicate —in terms of syntax, grammar, and word usage— are welcome in the transcript so long as they do not interfere with the written clarity of *what* speakers meant to communicate. Fidelity to key characteristics of each individual's speech holds a lower priority.

While oral history guidelines do tend to emphasize authenticity to the spoken word, they often encourage, and sometimes require, adherence to standard orthography. Some programs allow only 'dictionary spelling' and specifically prohibit non-standard representations, including requirements that contracted forms be spelled out (e.g., *won't*

[12] https://www.ncra.org/

rather than *will not*). Some guidelines recommend inserting dropped elements, such as pronouns or the third person singular verb agreement suffix *-s* which may be attributed to 'hurried speech' rather than potential regional linguistic markers. The few oral history guidelines that expect some faithfulness to the spoken sounds find compromises to handle regional linguistic variation. For example, the Samuel Proctor Oral History Project's (2020) transcription guide states that it is important to be faithful to the uniqueness of each person's voice and emphasizes preservation of an individual's personal manner of speaking, allowing reduced forms such as *kinda*, *gonna*, and *wanna* as well as non-standard grammar. The Samuel Proctor Oral History Project's (2016: 8) guidelines also permit transcribers to leave in double negations, explicitly stating "do not change improper grammar said by the speaker." At the same time, the guidelines prohibit transcription of conversational fillers (*um*, *er*, etc.) that might be expected for a linguistic analysis of conversation.

The guidelines vary in how to represent a variety's well-known features. Some recommend that the distinctive features of non-standard varieties be 'corrected' with the explicit purpose of not embarrassing the speaker, as if desiring to present every interviewee as polished, formal, and educated. For example, the oral history transcription guide by Oregon Department of Transportation Research Section (2010: 6) states that "Slang such as 'y'all' is acceptable —very occasionally— if that's what was spoken, although it should not be used extensively for regional approximations *á la* Mark Twain." Notably, no guidelines we examined contain sections dedicated to the transcription of non-standard or regional vocabulary, pronunciation, or grammar. Our corpus adheres to the SPOHP transcription guidelines except for the additional guideline we created to account for African American morphosyntactic features (see section 4).

3.2. Time alignment

Time alignment increases the accessibility and usability of oral histories. Time alignment is the matching of a transcription excerpt to an excerpt of audio. Oral history audio files are often archived separately from transcription files, only sharing a file name (with a different file extension). This lack of alignment between the text and audio makes it difficult to investigate the spoken aspects such as phonetic features tied to the interviewee's social background or emotional state, and non-literal meaning such as irony and humor. Without an aligned connection between the transcription and recordings, it is

more difficult to observe linguistic features (e.g., pronunciation variations) in the acoustics and annotate them in the text. Likewise, addressing transcription errors benefits from immediate access to the spoken utterance.

Fortunately, time alignment can be automated, via a procedure referred to as 'forced alignment'. Forced alignment can be performed at the utterance-level and at the word/phone-level using existing toolkits, such as *Aeneas* (Pettarin 2017). *Aeneas* generates utterance-level alignment by first synthesizing speech from the orthographic representations and then comparing the synthesized speech with the actual speech. This comparison obtains an approximation of each utterance' timestamps. Word/phone-level time alignment can be obtained with another tool, namely the *Montreal Forced Aligner* (MFA; McAuliffe *et al.* 2017), which relies on acoustic models of phonetic units. Both *Aeneas* and MFA were used to align our corpus. Our approach to time alignment is discussed in Section 4.

Unsurprisingly, the quality of alignments depends on the quality of the recordings. Oral history interviews take place in a variety of settings, such as a home, restaurant, school, and even outdoors. This mirrors the environmental settings of linguistic or anthropological fieldwork recordings (Whalen and McDonough 2015). When time-aligning the *Spoken British National Corpus*,[13] Coleman *et al.* (2011) reported several challenges. Everyday conversations were particularly difficult to align due to factors such as overlapping speakers, background noise, variable signal loudness, reverberation, distortion, and poor speaker vocal health. The alignment was suboptimal for phonetic acoustic research (only 24% of the phoneme boundaries were within 20 milliseconds of expert human labels) although it achieved sufficient accuracy for users to navigate to the desired audio portion of the transcription (83% of the phoneme boundaries were within two seconds of their correct positions). For this reason, time-alignment remains an important step when compiling a linguistic corpus from an oral history collection.

Forced aligners suffer from limited or poor-quality data and improve with increased training data. Forced aligners for majority languages have usually been trained on hundreds or even thousands of hours (DiCanio *et al.* 2013). This amount of data is not available for AAL. However, when data is not available, data from related languages or from languages with a similar phone inventory has been shown to improve forced

[13] http://www.phon.ox.ac.uk/AudioBNC

alignment (see Tang and Bennett 2019; Pandey *et al.* 2020, and references therein). So, GAE models can be fine-tuned for AAL.[14] Alternatively, a model trained on data with significant background noise removed can improve the alignment quality (Johnson *et al.* 2018). While such preprocessing of the audio data can improve the alignment, it is labor intensive. Johnson *et al.* (2018) report that to clean one hour and 42 minutes of recordings required an undergraduate research assistant to complete 120 to 150 hours of manual editing.

3.3. Annotating linguistic features of interest

Linguistic annotation is an important tool for linguistic study. Annotation consists of adding information to texts about features such as anaphora, parts-of-speech (POS), phonetics, semantic roles, and syntactic structure (see Schiel *et al.* 2012: Chapter 8). Oral history collections are not usually enriched with linguistic information. Those wishing to undertake in-depth linguistic investigation via oral history must undertake linguistic annotation.

In our experience, annotation of oral history transcriptions does not present special considerations or require a particular tool or method. Linguistic annotation can be undertaken with various tools and guidelines that can be designed to a project's goals to fit the team's workflow. However, if one wants to apply automatic annotation, it should be noted that most state-of-the-art NLP tools are not trained or optimized for spontaneous speech (Moore *et al.* 2015; Dinkar *et al.* 2023) and may need to be fine-tuned (Rohanian and Hough 2021) or require a final step of manual corrections.

4. CHALLENGES DURING COMPILATION

Compiling a linguistic corpus from an oral history collection gives rise to challenges, particularly with a non-standard language variety. This section details challenges we encountered during transcription and time-alignment and summarizes how we handled the challenges.

[14] See Magnotta (2022) for a comparison of forced aligners that were trained on either AAL speech or GAE speech data.

4.1. Transcription challenges and errors

Transcription of spontaneous speech is a challenging task. Transcribers misperceive what they hear due to background noise, poor audio quality, lackluster listening equipment, limited transcriber training, or the transcriber's fatigue (Meyer *et al.* 2013; Tang 2015). Transcription errors are unavoidable but mistranscriptions may have downstream negative effects. Critically, some mistranscriptions may not be considered errors by oral historians (see Section 3.1) but may nevertheless hinder identification of linguistic information. Additionally, if a corpus is used to train ASR systems, the systems may reflect bias against certain speakers due to the transcription choice (Blackley *et al.* 2019; Koenecke *et al.* 2020; Martin and Tang 2020).

Most oral history projects detail a pipeline for transcribing audio files. SPOHP employs its own group of in-house human transcribers who have received training based on the SPOHP's own transcription guidelines (Section 3.1). At SPOHP, each transcription goes through three passes. In the first pass, a first draft transcription is created. This is followed by an audit pass to correct errors and clarify areas marked as unsure. A third pass finalizes the transcription. To understand the transcription of JBA oral histories better, we took a detailed look at 14 transcriptions. We examine first draft transcripts (rather than final drafts) because oral history collections are constantly growing and at any given time, the bulk of the interviews are in first-draft status. First drafts provide an opportunity to examine transcribers' first impressions of AAL. Also, we found that not all mistranscriptions that could be attributed to misperceiving dialectal variation are corrected in the final drafts.

We identified mismatches between audio and transcripts and grouped them into four categories: 'omission', 'insertion', 'substitution', and 'unsure' (Hennink and Weber 2013; Stolcke and Droppo 2017; Zayats *et al.* 2019). Omission takes place when a word or phrase is present in the audio but not in the transcript, as illustrated in (7). Insertion occurs when a word or phrase is not present in the audio but is in the transcript, as in (8). Unsure involves cases in which the transcribers indicated they were unsure of their work, e.g., [*inaudible at 6:04*].

(7) Audio: A lot of my friends *had* got drafted and had already got killed.
Transcript: A lot of my friends got drafted and had already got killed.

(8) Audio: I went to a lady was teaching school in a wooden building.
Transcript: I went to a lady *who* was teaching school in a wooden building.

Substitution is found when the transcript misrepresents a word or phrase in the audio. Word-level substitutions differ from the audio by an entire word, as shown in (9). Character-level substitutions differ in no more than two letters or else a single inflectional morpheme differs, as illustrated in (10).

(9) Audio: In middle school I didn't go to a *white* school…
Transcript: In middle school I didn't go to a *black* school…

(10) Audio: I work for him until I went in the service.
Transcript: I work**ed** for him until I went in the service.

Across the 14 first-draft transcripts we found 1,041 mistranscriptions. The distribution is shown in Table 1. 82 percent were unconscious errors (omission, substitution, insertion). The most common type is that of substitutions (47%). Omissions (27%) and insertions (7%) are less common.

Total Errors	**Substitution**	**Omission**	**Insertion**
1,041	493	283	76

Table 1: Distribution of transcription errors in sample texts

We found notable trends that relate to misrepresentation of AAL. For example, character-level substitutions frequently occurred with verb tense markers which misrepresents AAL characteristic person/number disagreement, as shown in example (11). Also, word-level substitution could result in sparse representation of possible AAL signifiers, as in (12).

(11) Audio: I can't believe that name escape me.
Transcript: I can't believe that name escape**d** me.

(12) Audio: And my *father* came back from the war, and they fell in love…
Transcript: And my *pop* came back from the war, and they fell in love…

There were 38 occurrences of five of the six AAL features (remote past *bin* did not occur) and 31 were transcribed correctly. Occurrences of perfect *done* and multiple negations were always transcribed correctly. Interestingly, AAL features were artificially inserted two times, as in example (13).

(13) Audio: No, **I'd be** passing by…
Transcript: No, **I be** passing by…

Existential *it/dey* proved the most difficult for transcribers (and annotators) to identify. Occurrences were usually substituted for the GAE existential construction, as in example (14).

(14) Audio: …**it was** so many of us today that they had three.
Transcript: …**there were** so many of us today that they had three.

The null copula was mistranscribed three times. In example (15), we see an instance of null copula being incorrectly omitted by the insertion of *was*.

(15) Audio: …the same street the Duncan Brothers funeral home on.
Transcript: …the same street the Duncan Brothers funeral home **was** on.

When compiling an oral history collection into a corpus for linguistic study, transcription issues should be addressed in a way that best serves the compiler's goals. Our project intends to serve oral historians as well as linguists so, when making decisions about handling transcription issues, we sought a balance between accurate linguistic representation and easy-to-read transcripts. We decided to change the transcription guidelines only where they misrepresent the AAL grammatical features we are currently interested in (see Section 2.3). We correct other general issues such as larger missing or incorrect transcriptions to accomplish forced time alignment of the transcriptions and audio files. Such issues that related to several seconds of audio hinder time alignment. Our more accurate representation of the AAE features also seems to improve forced alignment and increase utility for training NLP models for AAE. We work with final draft transcriptions where possible. In the future, we hope to address how best to represent phonological features while still maintaining readability.

In our project, revising transcriptions has been integrated into the pipeline of annotation and time-alignment. We developed simple short recorded video training with exercises about the AAL features of interest. The training makes transcribers aware of these features and AAL in general, guides them to preserve standardized spelling wherever possible, but to refrain from 'correcting' potential dialectal markers. Readers who wish to inform their own corpus compilation work may find our training materials in the publicly-available repository mentioned in Appendix B (Data Availability).

We currently focus annotation efforts on the six distinctive AAL morphosyntactic features listed in Section 2.3. Annotation is performed by students who are native or near native English speakers of any English variety (including AAL), having varying

familiarity with AAL and varying backgrounds in linguistics. Annotators are trained and tested on their ability to identify the six AAL morphosyntactic features before they are allowed to work independently. They listen to the audio recordings and annotate the location and type of feature in the transcriptions using a software tool named 'Rezonator' (DuBois *et al.* 2020). In this tool, annotators highlight a word or phrase with a mouse and select from a list of feature labels. Rezonator text annotations can be exported as CSV files with units in rows and their labels in cells on the same row. In Figure 1, the linguistic feature of person/number disagreement is shown as annotated within an interview. We found it efficient to have multiple rounds of annotation. In each round, annotators focus on one or two features rather than all six at once, while also checking for features they had previously annotated but may have missed.

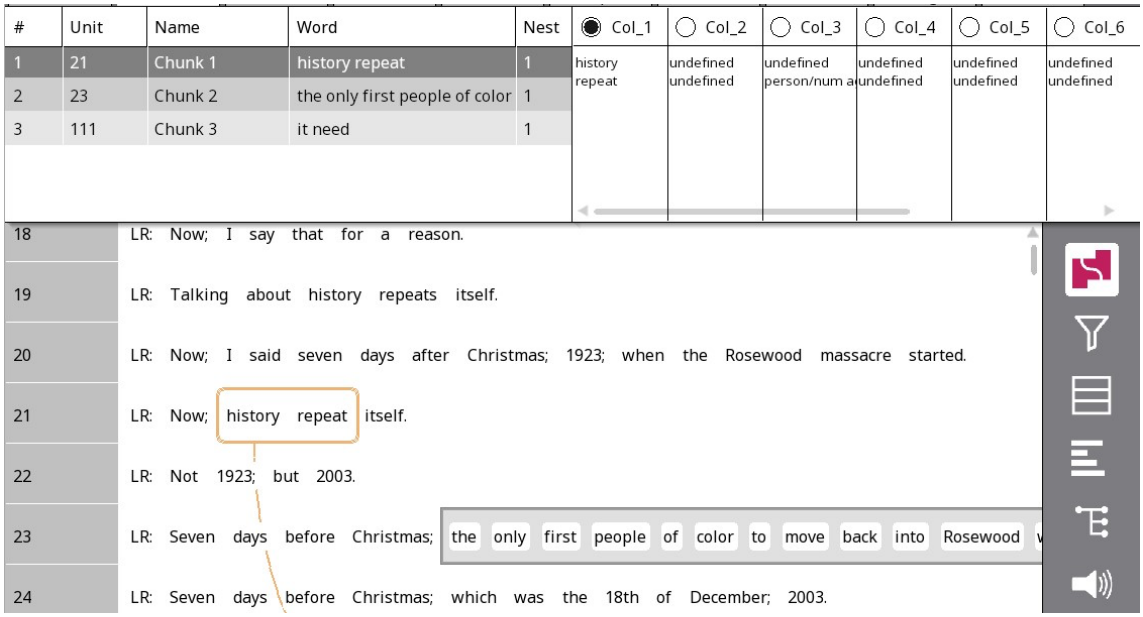

Figure 1: An annotation within an interview

4.2. Time-alignment: Challenges and errors

Time-alignment of audio and text in non-GAE varieties presents a set of challenges that stems from various sources such as mismatches in transcriptions, cross-talk, and lack of customized phonetic resources. In some instances, we have been able to address these challenges. For others, we merely highlight here obstacles that are likely to be encountered by anyone compiling a corpus from oral histories.

4.2.1. Challenges due to mistranscription

One challenge stems from the mistranscriptions of AAL (see Sections 3.1 and 4.1), which can result in misalignment, or even complete omission of distinctive features during forced alignment. As an example, the clause *...when Live Oak got **they** team...* in AAL was mistranscribed as *...when Live Oak got **their** team...* (*they* has the equivalent of *their* in GAE). If time-aligned files output from *Aeneas* and from MFA are constructed from such a mistranscription, the mistranscription would lead MFA to attempt aligning the vowel in *they* [eɪ] using an acoustic model of the vowel in their [ɛ] and it would also try to align some audio signals to an acoustic model of a rhotic phone [ɹ] in *their*, which does not exist in the audio. The misalignment of the phones results in an overall poorer word-level alignment of this and surrounding words, as illustrated in Figure 2.

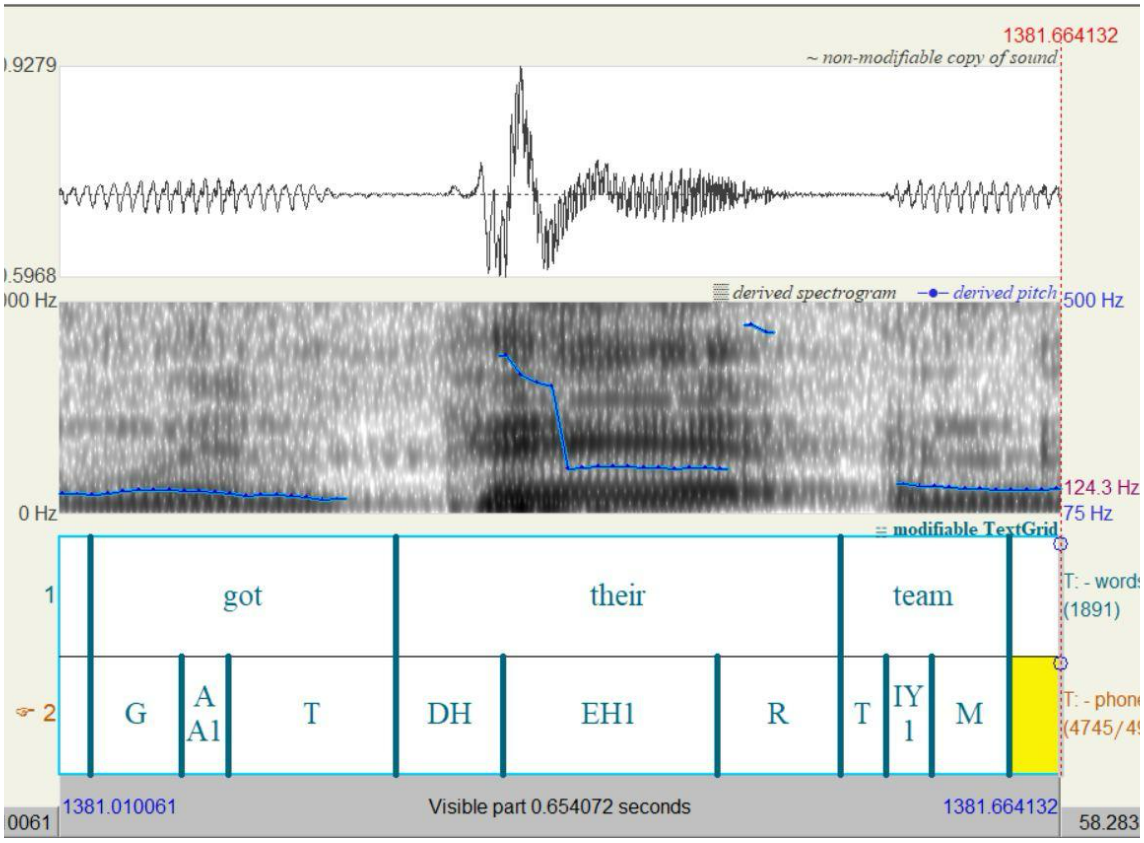

Figure 2: Time-aligned files that, due to mistranscription, do not contain the AAL feature *they*

To mitigate this issue, we are correcting the transcriptions prior to time-alignment (see Section 4.1). While automatic transcription error detection exists, such as Kisler and Schiel (2018), they require a pre-existing set of corrected transcriptions before a language-specific model can be trained. We decided to leave automatic transcription error detection for future work as we must first create a large enough set of corrected

transcriptions. We decided on steps that involve focusing first on interview transcriptions in the final draft version and then manually correcting any transcripts that contain AAL features in the audio but not in the transcripts (see Section 4.1). Once the transcript correctly matches the features in the audio, MFA is able to time-align the audio signals with the transcriptions more accurately. Remaining alignment errors can be quickly corrected by hand.

4.2.2. Cross-talk

Cross-talk occurs when two or more speakers produce an utterance simultaneously. It often results in one speaker cutting off their utterances upon hearing the other. The transcription in Figure 3 illustrates two instances of the interviewer being interrupted by the interviewee (fourth and third-to-last lines). Not all cases of cross-talks are clearly indicated in the transcripts as an interruption (e.g., a dash line at the end of a line) because the speaker being interrupted may finish their utterances as if there is no interruption. Crucially, the onset and offset time of cross-talks (i.e., the exact portion where both speakers overlap) may not be indicated in oral history transcripts.

```
T:      My momma's Naomi Bryan. Father, unknown.

W:      And when was your mother born?

T:      Let me see now; she was born in December the 12th, 1910.

W:      How about her mother and father? What do you know about her—

T:      I don't know nothing, I don't know about them.

W:      Do you have brothers and sisters?

T:      One sister.

W:      Can you state her name?

T:      Sarah Cross.

W:      And is she still—

T:      Deceased.

W:      Deceased.
```

Figure 3: Original transcription of two cross-talk examples

The identification of words and phones can be hampered due to the conflicting voice activity. One solution is to split speakers' audio into separate channels or even entire files if the audio was recorded with a dual channel setup. Another solution is to label an utterance as belonging to a certain speaker (speaker diarization). This can be done automatically using NLP models trained specifically for this task and it is supported by MFA. We are currently testing this method but, as with most language technologies, speaker diarization is known to perform worse for non-standard varieties (Tevissen *et al.* 2023). Currently, we are performing the speaker diarization task manually by listening for cross-talk in the audio and using indicators of interruption in the oral history transcription. We then separate the time-aligned words into speaker tiers in textgrid format, thus allowing cross-talk to be represented in a way that reduces misalignment.

4.2.3. Pronunciation dictionaries missing AAL variants

Pronunciation dictionaries for forced alignment exist for different languages, but their diversity as it relates to English varieties is limited. AAL does not have a pronunciation dictionary in the MFA database. Thus, phonetic representations may not be comprehensive for AAL. For instance, the MFA ARPABET dictionary for English does not include the AAL *th*-stopping pronunciation, as can be seen in Figure 4, which contains only the *DH* mapping to the IPA [ð], but not the *D* mapping to the IPA [d] AAL variants for the words *that* and *they*. This is a challenge we have not yet been able to address adequately.

```
that    DH AE1 T
that    DH AH0 T
that'   DH AE1 T AH0
that'd  DH AE1 T IH0 D
that'll DH AE1 T AH0 L
that'n  DH AE1 T AH0 AH0 N
that're DH AE1 T AH0 R
that's  DH AE1 T S
that'sh DH AE1 T AH0 SH
that'th DH AE1 T IH0 TH
that'ud DH AE1 T IH0 AH0 D
that've DH AE1 T AH0 V
```

```
they    DH EY1
they'd  DH EY1 D
they'l  DH EY1 AH0 L
they'ld DH EY1 D
they'll DH EY1 L
they'm  DH EY1 AH0 M
they'n  DH EY1 AH0 N
they're DH EH1 R
they's  DH EY1 Z
they'se DH EY1 S
they've DH EY1 V
```

Figure 4: Dictionary entries for variations of *that* and *they*, which omit the *th*-stopping feature found in AAL

AAL phonetic/phonological features can be modeled using a phone-level forced-alignment model which can capture pronunciation variation due to accent and regional differences (Yuan and Liberman 2011; McLarty *et al.* 2019; Kendall *et al.* 2021). This is achieved by allowing the model to evaluate multiple pronunciations of the same word. For example, given the acoustic signal of the word *running*, which can be produced with a velar nasal sound or with an alveolar nasal sound (velar nasal fronting), the model will assign the most likely pronunciation. Some well reported phonological features include: *th*-stopping (e.g., *that* [dæt]; Thomas and Bailey 2015), velar nasal fronting (e.g., *running* [ɹʌnɪn]; Tagliamonte 2004), final consonant cluster reduction (e.g., *test* [tes], *hand* [han]; Green 2002: 107), monophthongization of /ai/ (e.g., *buy* [ba:]; Rahman 2008: 147) and /ɹ/ vocalization (e.g., *court* [koət], *bear* [beə]; Green 2002: 120). Some grammatical features that resemble a pronunciation variation can also be automatically annotated as such, for instance, the absence of *-s* verb tense inflection (e.g., *He goes* [goʊ]; Rahman 2008: 147).

The latest version of MFA has an advanced feature that enables the selection of the correct pronunciations given the acoustic likelihood with a set of prior pronunciation probabilities. However, we cannot easily estimate the pronunciation probabilities of each variant without a large and accurate phonetically transcribed spoken AAL corpus.[15] Therefore, we cannot make use of an advanced feature of MFA. For this reason, we decided to generate the possible pronunciation variations of each word type with equal probabilities. We have extracted sentences from JBA interview transcripts to identify any words that had pronunciations missing from the MFA dictionary, then generated more comprehensive dictionaries that account for AAL pronunciations of these words. We are currently testing this method, as we continue to address time-alignment challenges.

5. Discovering Linguistic Information in Oral Histories

The process of compiling a linguistically enriched corpus from oral histories provides insights into the language variety spoken by the interviewees in natural settings. This section describes insights we have gleaned so far about AAL. The insights pertain to the

[15] CORAAL can potentially be used for this purpose as it has been fully forced aligned with phonetic transcriptions (Farrington and Kendall 2019). However, we have not investigated to what extent CORAAL handled pronunciation variations and whether they were manually verified.

representativeness of an oral history collection (Section 5.1), the distribution of AAL features in speech (Section 5.2) and modeling syntactic structures that signal the presence of the AAL feature habitual *be* (Section 5.3). Finally, we discuss the potential for modeling AAL linguistic information preserved in oral history collections with NLP systems, which in turn can provide automated assistance to annotators (Section 5.4).

5.1. Preliminary description of the UF AAL Spoken Corpus

The JBA (see Section 2.2) was chosen due to its abundance of AAL speakers and its conversational nature that is conducive to naturally occurring instances of AAL features (Roller 2015). Our goal is to compile a corpus with no less than 500 JBA interviews that are completely time-aligned and featurally annotated. The compiled corpus will aid the linguistic investigation of AAL and the development of NLP tools that reduce bias against AAL.

The notion of representativeness is an important consideration in corpus compilation for linguistic investigation. However, oral historians' sampling technique is not governed by linguistic representativeness (as mentioned in Section 2.2, snow-ball sampling was used in JBA). Egbert *et al.* (2022: 28–51) report a survey of conceptualizations of representativeness in corpus linguistics, but all these conceptualizations that tie to random or stratified sampling are not applicable to a corpus compiled from an oral history collection. However, one conceptualization, which Egbert *et al.* (2022) argue against, is that a very large corpus is a *de facto* representative corpus. In this view, corpus size is the primary consideration of corpus design and, as Sinclair (1991: 8) states, "... a corpus should be as large as possible and should keep on growing." A spoken corpus with 500 JBA interviews is arguably a very large corpus, indeed larger than the other significant AAL corpus (CORAAL). When completed, the full corpus should serve as a corpus of the AAL spoken in the state of Florida and across the Southeastern United States. What specific linguistic features and socio-demographic dimensions it will be representative of remain to be determined upon completion.

For the purpose of NLP tool development, we chose a subset of 58 interviews. It is important to note that not all African Americans speak AAL. Therefore, our initial corpus subset was compiled from a preliminary inspection of the JBA transcribed materials conducted to identify speakers who consistently used AAL, while giving strong weight

to audio quality (see Section 3.2). We also made practical considerations, such as if the transcription was completed and if the recording had no more than three AAL interviewees. Our current compiled corpus comprises speech from 18 interviewees across 16 interviews that have been fully annotated. In the 18 interviewees, gender is evenly distributed (nine male speakers and nine female speakers). 15 speakers are adults ranging from 26–80 years while the remaining three are teenagers. The interviews were recorded over a ten-year span from 2008 to 2018 and were located through Mississippi and northern Florida. The corpus contains 59,388 tokens, averaging 4,568 per document (see Appendix A for more details).

5.2. Distribution of AAL features

Early in the project, we performed a pilot annotation round on 16 interviews that were annotated fully in one round for all six features. These interviews were annotated by our trained annotators, but they were also annotated independently by a class of graduate and undergraduate students who were given an abbreviated version of the training. Although the more thoroughly trained annotators were better at hearing some features, the two teams of annotators confirmed each other's findings regarding the distribution of features.

The pilot study allowed us to investigate the distribution of AAL features and understand the prevalence of each feature better. The results are displayed in Table 2. Null copula (34%) and person/number disagreement (30%) together comprise over half the AAL features in the data. Multiple negation (16%) and existential *it/dey* are less frequent but still prominent and together comprise one third of the AAL features. The other three features are much less frequent. Remote past *bin* and habitual *be* are the rarest, while perfect *done* occurs only slightly more frequently. Our observations on remote past *bin* match other work where it is noted as a rare feature (Green *et al.* 2022).

Text	Null copula	Person/ Number	Multiple Negation	Existential *it/dey*	Perfect *done*	Remote past *bin*	Habitual *be*	Total
1	25	9	1	6	0	0	1	42
2	0	1	0	3	0	0	0	4
3	3	15	4	3	2	0	0	27
4	2	0	0	3	0	1	0	6
5	10	7	5	4	3	1	0	30
6	2	7	4	5	0	0	0	18
7	1	1	0	0	0	0	0	2
8	5	7	8	3	0	0	0	23
9	7	3	4	0	0	0	0	14
10	2	3	3	2	0	0	0	10
11	15	13	7	1	1	0	1	38
12	6	2	1	0	0	0	1	10
13	0	0	0	0	0	0	0	0
14	4	0	1	0	0	0	0	5
15	5	7	5	6	1	0	0	24
16	0	1	0	4	0	0	0	5
Total	87	76	43	40	7	2	3	258

Table 2: AAL features found by annotation team in a pilot study of 16 oral histories[16]

We found that the features are not equally recognizable to annotators. Existential *it/dey* is difficult to identify despite its frequency and requires second passes. This feature may be difficult to identify because it requires attention to the larger context, whereas the other features can mostly be identified by examining the sentence or phrasal context. Another feature that is problematic for annotators is remote past *bin*. We suspect the motivation for this are the feature's rarity, semantic load, and contextual constraints which are more complex than what is described in the literature. On the other hand, multiple negation, which is found in other varieties of English, is one of the most frequent features and seemingly the easiest to recognize.

5.3. Syntactic environments of habitual be

It is known that certain syntactic environments correlate with AAL features like habitual *be* (Fasold 1972; Green 2002). We used the oral history data to investigate whether these described environments hold in naturalistic data. We investigated POS and syntactic dependency relations in the environment of habitual and non-habitual occurrences of *be*. To analyze the structures, we first ran the *NLTK POS* tagger (Bird 2009) and the *spaCy*

[16] The list of texts by interviewee is shown in Appendix A.

Universal Dependency syntactic parser[17] on all sentences containing *be*. Both models are originally trained on GAE. Then, we analyzed the syntactic patterns and, from this analysis, built a rule-based machine learning classifier to identify *be* as either habitual or non-habitual.[18] In the process, we confirmed POS and syntactic dependencies commonly found in the environment of habitual *be* and uncovered new ones. Further information about the NLP results of the efforts described here are available in Previlon *et al.* (2024).

5.3.1. POS environments

We first leveraged POS patterns distinguishing habitual and non-habitual usage of *be* as described by Green (2002). For example, patterns indicative of the habitual meaning include a pronoun immediately preceding *be*, as in *...they be like, what you finna do?* or a verb ending in *-ing* immediately following it, as in *But LeBron be passing though*. We coded these POS environments described in Green (2002) as Boolean (True/False) Python rules to filter out many non-habitual *be* instances. These filtering rules do not capture all instances of non-habitual *be*, but they also do not flag any false positives.

Examining non-habitual instances that were not filtered, we uncovered POS patterns not described in the literature. Because our data is limited, we labeled these new patterns *ad-hoc*. Future study may determine if they are generally applicable. *Ad-hoc* POS rule 1 states that *be* is non-habitual if it is immediately followed by a deverbal noun and immediately preceded by neither a personal pronoun nor a noun, as in (16) below. *Ad-hoc* POS rule 2 states that *be* is non-habitual if it is immediately preceded by an adverb and immediately preceding that adverb is either a verb or modal verb, as in (17).

(16) I will **never** be **going** there again.

(17) You **should regularly** be trimming the dog's nails.

The power of the known and the *ad-hoc* POS patterns to disambiguate habitual/non-habitual meanings was tested by applying a machine learning classifier trained on the output of the filtering rules to 5,133 instances of *be* in the CORAAL corpus (Santiago *et al.* 2022). The results were compared to the manually annotations and are displayed in

[17] https://spacy.io/

[18] It would be more accurate to say that we disambiguate standard uses of *be* from non-standard, because we do not distinguish between habitual *be* and other non-standard invariant forms, such as emphatic *be*, which occur in nearly identical syntactic environments (Harris 2019).

Table 3. The POS-based classifier correctly tags 79 percent of non-habitual instances, and only 13 percent of instances of habitual *be* were false positives.

The *ad-hoc* rules increased non-habitual filtering accuracy over just the known environments, but unlike the known POS patterns, they also incorrectly filtered some habitual instances. Future analysis may show whether these false positives can be reduced by refining the *ad-hoc* analysis or whether these environments are indeed ambiguous.

	Tagged as 'non-habitual'	**Not tagged**	**Total**
Non-habitual	3,662	994	4,656
Habitual	61	416	477

Table 3: Number of habitual and non-habitual usages of *be* and how they are tagged by rules based on POS environments[19]

5.3.2. Dependency syntax environments

As part of the process of building NLP tools for AAL, we explored how GAE-trained NLP tools model the syntactic environments that signal the presence of habitual *be*. We contrasted those environments with the modeling of non-habitual *be* syntax. We parsed the dependency trees of 250 sentences (132 habitual and 118 non-habitual) containing *be* in the JBA data. Nine significant patterns were identified. Of note for developers of NLP tools to disambiguate habitual and non-habitual *be* is that the relevant dependency relations are primarily constrained by immediate parent, child, and sibling relations between *be* and other words. Zoomable figures illustrating these dependency structures are available on the Open Science Framework (see Appendix A: Data availability).

The nine patterns can be stated as rules. Rules 1–3 identify habitual patterns while rules 4–9 identify non-habitual patterns. In large part, our analysis of computational modeling confirms Green's (2002) description of habitual *be*. Rule 1, 'main verb', is found when *be* is POS-tagged as a verb and has a child that is an adjective, adverb, or preposition (48 instances). Rule 2, 'aux to main verb', is when *be* is both POS-tagged and has a dependency relation of auxiliary while its parent is labeled as a verb (70 instances). Rule 3, 'subordinate clause', occurs when *be* is POS-tagged as an auxiliary with the

[19] It should be borne in mind that the POS rule-based tagger identifies non-habitual usages, so 'not tagged' does not necessarily imply habitual meaning.

dependency relation of either a relative clause modifier or a clausal complement (6 instances).

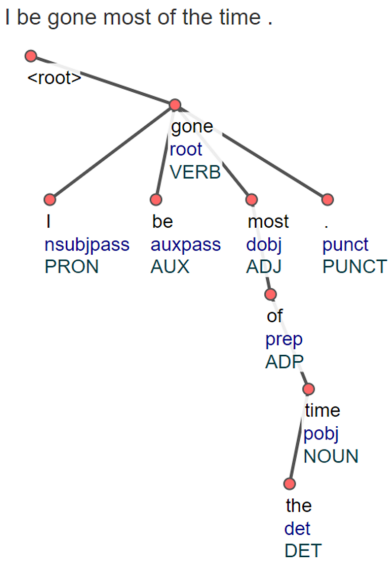

Figure 5: Illustration of habitual dependency rule 2

Under the pattern of rule 2, an interesting phenomenon is found when *be* is followed by *gone* as in *I be gone most of the time*. This is illustrated in Figure 5, which shows that the parser, although useful for capturing the habitual pattern described in rule 2, nevertheless mischaracterizes the syntactic and semantic structure of AAL. It labels *be* as an auxiliary and *gone* as a verb. In GAE, *gone* functions as a verb when used as the past participle of *go* (*She had gone to the store*), but takes on an adjectival role when describing a state of existence, as in *The bad days are gone*. In this example, it can be argued that, with the habitual meaning, *gone* leans more towards an adjectival interpretation than a verbal one, since the speaker is not actively engaging in the act of 'gon-ing', but rather *gone* describes a repeated state of being. That is, the subject is often not here, but it is not the case that it has often left for a place.

Five patterns can be stated as dependency rules of non-habitual meanings. Rule 4, 'auxiliary child', takes place when *be* has a child that is both POS-tagged as auxiliary and has a dependency relation of auxiliary (36 instances). Rule 5, 'auxiliary sibling', is when *be* has a sibling that is both POS-tagged as auxiliary and has a dependency relation of auxiliary (22 instances). Rule 6, 'governed by conjunction', is when *be* has *and* as a sibling and *be* is POS-tagged as an auxiliary with a dependency relation of conjunction. Also, *be* has a child that is an adjective, adverb, or noun (four instances). Rule 7, 'noun child', is when *be* is immediately preceded by a pronoun and has a child that is a noun

(one instance). Rule 8, 'infinitival child', is when *be* has an immediate child that is POS-tagged as a particle and has a dependency relation of auxiliary (51 instances). Typically, the child is infinitival *to* (46 instances), or its phonetic variations *ta* (two instances), or *na* as in *gonna* and *wanna* (three instances). Finally, rule 9, 'particle child', is when *be* has a child that is POS-tagged as a particle with the dependency relation of auxiliary (one instance).

The nine rules discussed above capture all but eight of the 250 sentences. Examining these eight sentences, we learned that the parser also mischaracterizes AAL because of non-standard orthography, such as *g*-dropping, when the velar nasal fronting or variable (ING) is represented orthographically as *-in*,[20] presumably as a reflection of the pronunciation (Tagliamonte 2004; Hazen 2008). Similarly, the non-standard spelling of *Imma*, as in *Imma be talking in a minute*, is incorrectly labeled as a proper noun and *wanna*, as in *...and you never wanna be here tomorrow*, is sometimes labeled as a verb. Similar cases of tagging errors by GAE-trained models on common AAL lexical items have previously been reported (Dacon 2022).

Our analysis of the oral history data reveals that both POS patterns and dependency structures are needed to describe and identify the habitual *be* construction. We tested the expediency of including the syntactic environments in an automatic habitual *be* tagger and this reveals that POS-based descriptions do not sufficiently disambiguate habitual and non-habitual meanings. The addition of dependency-based rules allows the tagger to correctly identify habitual *be* when the POS-based rules do not. For example, the use of syntactic dependencies allows the identification of a *be* that is POS-tagged as a verb (matching the first dependency parsing rule) as habitual, but a POS-only model would flag it as non-habitual. Similarly, in another example, the second dependency parsing rule ('aux to main verb') correctly flags habitual *be* where the POS-only approach does not.

5.4. Automatic annotation of syntactic features

Annotation of distinctive AAL morphosyntactic features is necessary for the study of the language variety. Unfortunately, manual annotation is prohibitive in terms of time and cost, particularly because oral history programs continually add interviews and transcriptions. To assist annotation, innovative methods with NLP can be applied.

[20] The six instances were *tryin*, *walkin*, *willin*, *throwin*, *hurtin*, and *laughin*.

Numerous NLP tools exist that work well for GAE and the same cannot be said for AAL with some exceptions (Jørgensen *et al.* 2016; Blodgett *et al.* 2018). As seen in Section 5.3.2, pre-trained models for GAE cannot however be readily transferred to AAL (see Ziems *et al.* 2022). Therefore, we are designing our own AAL feature taggers.

State-of-the-art NLP models that will identify the features automatically are dependent on massive amounts of annotated data. Manual annotation is necessary to create training examples. When data is limited, then linguistic analysis can compensate, to a certain extent. We decided to leverage syntactic patterns that statistically correlate with AAL features as input data to machine learning models.

Our general steps for developing an AAL feature tagger are 1) identifying linguistic contexts of the feature in the data, 2) analyzing the contextual patterns, 3) coding those patterns as Boolean (True/False) rules that filter non-occurrences and flag likely occurrences of the feature, and 4) using the rules to train a classifier to identify whether a string of text contains the feature of interest. The results can then be integrated into the annotation process by 5) presenting the computer's 'annotations' to human annotators for checking, and 6) using the human corrections to retrain and improve the model. This builds a machine-in-the-loop cycle of annotation that should be faster and more accurate than either purely manual or purely automatic work.

After developing the habitual *be* tagger, we tested it on 5,133 manually annotated sentences of CORAAL (Kendall and Farrington 2021), which were kept separate from the data used to analyze the syntactic environments. Four machine learning models were implemented with *scikit-learn* (version 1.2.2; Pedregosa *et al.* 2011) and a transformer was implemented with *fairseq* (Ott *et al.* 2019). The best model achieved a 0.96 F1 score and beat a baseline that does not leverage syntactic patterns. This shows robust results even in the face of data sparsity. The detail of these model developments and experiments can be found in Previlon *et al.* (2024).

However, the results included many false negatives, indicating additional analysis is needed. In the meantime, an effective machine-in-the-loop annotation cycle only requires annotators to verify true positives and false positives of habitual *be*. We were able to reduce the tagger's false tagging of the habitual *be* as non-habitual by increasing the model's recall. We also applied data augmentation with synthesized habitual *be* sentences *á la* Santiago *et al.* (2022). This created a more balanced training corpus because habitual *be* is rare. A corpus with nearly equal examples of the two classes

positively impacts results by reducing statistical bias towards the more frequent non-habitual *be*.

6. CONCLUSION

This study presented initial efforts in compiling a spoken corpus of AAL using recordings and transcriptions from the oral history discipline (Section 2). This corpus will both address the AAL data gap and allow technology developers to correct racially biased systems. We acknowledge that our project is not the first to compile a set of 'legacy' audio recordings into a linguistic corpus. For instance, Olsen *et al.* (2017) reported on a pipeline to deal with transcription and time alignment issues with a legacy speech corpus consisting of sociolinguistic interviews (Pederson *et al.* 1986) which contain African American speakers. Nonetheless, our project involves the compilation of data that is not only historical but that was originally collected for purposes other than linguistic research, which brings a unique set of challenges (Sections 3 and 4). Furthermore, additional challenges arise due to a lack of reliable computational tools and established transcription standards for AAL, which we are addressing through the process of compiling this corpus (Sections 4 and 5).

In Section 5, we demonstrate the wealth of linguistic information that can be extracted from oral histories, such as a frequency distribution of AAL linguistic features and models of the features' syntactic patterns. As our annotation continues, these analyses will be updated. Discovering distributional information from just a handful of interviews reveals the possibilities for oral history work to enhance corpus linguistics research. No longer limited to self-collected data in small amounts, or publicly accessible data that contains unverifiable sources (e.g., Blodgett *et al.* 2018), linguists may expand their research to address concerns in specific populations. It has enabled us to examine speech features of AAL speakers across the Gulf South on data compiled for non-linguistic purposes. Following the guidelines of Kendall and Farrington (2022) about managing the African American sociolinguistic data, in the future, we may find connections between regionality and the use of various features by annotating what sociolinguistic metadata the oral historians collected or that are found within the oral histories themselves.

Finally, our work contributes to the development of NLP for AAL. Modern speech technology is not possible without time-aligned language data. We are currently testing

the ability of forced alignment to perform on multiple pronunciation features of AAL. Since AAL cannot be fully represented within the limits of GAE orthography, additional linguistic annotation is needed. This annotation, in turn, provides training data for NLP models that will be effective on AAL. The last part of Section 5 demonstrates development of NLP tools designed for annotating AAL linguistic features. We developed an automatic classifier for tagging the AAL feature called habitual *be*. Our work improves an NLP model using insights from syntax. We are extending this method to automatically tag other features such as the neutralized person/number agreement characteristic of AAL. Crucially, our spoken corpus of oral histories can contribute both in providing authentic spoken data for AAL as well as the cultural references mentioned in the topic-guided oral history interviews.

References

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. California: O'Reilly Media Inc.

Blackley, Suzanne V., Jessica Huynh, Liqin Wang, Zfania Korach and Li Zhou. 2019. Speech recognition for clinical documentation from 1990 to 2018: A systematic review. *Journal of the American Medical Informatics Association* 26/4: 324–338.

Blodgett, Su Lin, Johnny Wei and Brendan O'Connor. 2018. Twitter universal dependency parsing for African-American and mainstream American English. In Iryna Gurevych and Yusuke Miayo eds. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 1415–1425.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel Tetreault eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online publication: Association for Computational Linguistics, 5454–5476.

Coleman, John, Mark Liberman, Greg Kochanski, Jiahong Yuan, Sergio Grau, Chris Cieri and Lou Burnard. 2011. Mining years and years of speech. *Phonetics Laboratory of the University of Oxford*: 1–23. https://diggingintodata.org/sites/diggingintodata.org/files/miningayearofspeechwhitepaper.pdf.

Columbia University Center for Oral History Research. 2022. *Columbia University Oral History Transcription Style Guide*. https://www.ccohr.incite.columbia.edu/s/CCOHR-Transcript-Style-Guide-2022-htpm.pdf (accessed 31 January 2024.)

Dacon, Jamell. 2022. Towards a deep multi-layered dialectal language analysis: A case study of African-American English. In Su Lin Blodgett, Hal Daumé III, Michael Madaio, Anika Nenkova, Brendan O'Connor, Hanna Wallach and Qian Yang eds. *Proceedings of the 2nd Workshop on Bridging Human–Computer Interaction and*

Natural Language Processing. Seattle: Association for Computational Linguistics, 55–63.

Davis, Alexis, Joshua L. Martin, Eric Cooks, Melissa J. Vilaro, Danyell Wilson-Howard, Kevin Tang and Janice Krieger. 2024. From English to "Englishes": A process perspective on enhancing the linguistic responsiveness of culturally tailored cancer prevention interventions. *Journal of Medical Internet Research* preprint: 57528. https://preprints.jmir.org/preprint/57528

DiCanio, Christian, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America* 134/3: 2235–2246.

Dinkar, Tanvi, Chleé Clavel and Ioana Vasilescu. 2023. Fillers in spoken language understanding: Computational and psycholinguistic perspectives. *arXiv* preprint arXiv:2301.10761: 1–20. https://arxiv.org/pdf/2301.10761.pdf

DuBois, John W., Terry DuBois, Georgio Klironomos and Brady Moore. 2020. From answer to question: Coherence analysis with rezonator. In Sophia Malamud, James Pustejovsky and Jonathan Ginzburg eds. *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Short Papers*. Waltham, New Jersey: SEMDIAL, 1–4. http://semdial.org/anthology/Z20-Bois_semdial_0031.pdf

Egbert, Jesse, Biber Douglass and Betanny Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness.* Cambridge: Cambridge University Press.

Farrington, Charlie and Tyler Kendall. 2019. *The Corpus of Regional African American Language: MFA-Aligned*. Version 2019.06. http://lingtools.uoregon.edu/coraal/aligned/.

Fasold, Ralph. 1972. *Tense Marking in Black English: A Linguistic and Social Analysis*. Washington: Center for Applied Linguistics.

Fitzgerald, Chris. 2022. *Investigating a Corpus of Historical Oral Testimonies: The Linguistic Construction of Certainty*. London: Routledge

Ghyselen, Anne-Sophie, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen and Arjan van Hessen. 2020. Clearing the transcription hurdle in dialect corpus building: The Corpus of Southern Dutch Dialects as case study. *Frontiers in artificial intelligence* 3/10. https://doi.org/10.3389/frai.2020.00010

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

Green, Lisa, Kristine M. Yu, Anissa Neal, Ayana Whitmal, Tamira Powe and Deniz Özyıldız. 2022. Range in the use and realization of BIN in African American English. *Language and Speech* 65/4: 958–1006.

Harris, A. Nicole. 2019. *The Non-Aspectual Meaning of African American English Aspect Markers*. New Haven: Yale University ProQuest Dissertations Publishing.

Harrington, Jonathan. 2010. *Phonetic Analysis of Speech Corpora*. Hoboken: John Wiley & Sons.

Hazen, Kirk. 2008. A vernacular baseline for English in Appalachia. *American Speech* 83/2: 116–140.

Hennink, Monique and Mary B. Weber. 2013. Quality issues of court reporters and transcriptionists for qualitative research. *Qualitative Health Research* 23/5: 700–710.

Johnson, Lisa M., Marianna Di Paolo and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation & Conservation* 12: 80–123.

Jørgensen, Anna, Dirk Hovy and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In Kevin Knight, Ani Nenkova and Owen Rambow eds. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego: Association for Computational Linguistics, 1115–1120

Kendall, Tyler and Charlie Farrington. 2021. *The Corpus of Regional African American Language.* Version 2020.05. http://oraal.uoregon.edu/coraal (accessed 25 June 2023.)

Kendall, Tyler and Charlie Farrington. 2022. Managing sociolinguistic data with the Corpus of Regional African American Language (CORAAL). In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister eds. *The Open Handbook of Linguistic Data Management.* Massachusetts: The MIT Press, 185–94.

Kendall, Tyler, Charlotte Vaughn, Charlie Farrington, Kaylynn Gunter, Jaidan McLean, Chloe Tacata and Shelby Arnson. 2021. Considering performance in the automated and manual coding of sociolinguistic variables: Lessons from variable (ING). *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.648543

Kisler, Thomas and Florian Schiel. 2018. MOCCA: Measure of confidence for corpus analysis: Automatic reliability check of transcript and automatic segmentation. In Nicoletta Calzolari ed. *Proceedings of the 11th International Conference on Language Resources and Evaluation*. Miyazaki: European Language Resources Association, 1781–1786.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky and Sharad Goel. 2020. Racial disparities in automated speech recognition. In Judith T. Irvine and Ann Arbor eds. *Proceedings of the National Academy of Sciences* 117/14: 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kortmann, Bernd and Wagner, Susanne. 2005. The Freiburg English Dialect Project and Corpus (FRED). In Bernd Kortmann, Tanja Herrmann, Lukas Pietsch and Susanne Wagner eds. *Volume 1 Agreement, Gender, Relative Clauses.* Berlin: De Gruyter Mouton, 1–20.

Lee, Donghee N., Myiah J. Hutchens, Thomas J. George, Danyell Wilson-Howard, Eric J. Cooks and Janice L. Krieger. 2022. Do they speak like me? Exploring how perceptions of linguistic difference may influence patient perceptions of healthcare providers. *Medical Education Online:* 27/1: 2107470. https://doi.org/10.1080/10872981.2022.2107470

Magnotta, Sierra. 2022. *Analysis of Two Acoustic Models on Forced Alignment of African American English*. Georgia, U.S.: University of Georgia dissertation.

Martin, Joshua L. 2022. *Automatic Speech Recognition Systems, Spoken Corpora, and African American Language: An Examination of Linguistic Bias and Morphosyntactic Features*. Gainesville, Florida: University of Florida dissertation.

Martin, Joshua L. and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual 'be'. *Interspeech*: 626–630.

Martin, Joshua L. and Kelly E. Wright. 2022. Bias in automatic speech recognition: The case of African American language. *Applied Linguistics* 44/4: 613–630.

McAuliffe, Michael, Michaela Socolof, Michael Wagner and Morgran Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *INTERSPEECH:* 498–502.

McLarty, Jason, Taylor Jones and Christopher Hall. 2019. Corpus-based sociophonetic approaches to postvocalic R-lessness in African American language. *American Speech* 94/1: 91–109.

Meyer, Julien, Laure Dentel and Fanny Meunier. 2013. Speech recognition in natural background noise. *PloS one* 8/11. https://doi.org/10.1371/journal.pone.0079279

Moore, Russell, Andrew Caine, Calbert Graham and Paula Buttery. 2015. Incremental dependency parsing and disfluency detection in spoken learner English. In Pavel Král and Václav Matoušek eds. *Text, Speech, and Dialogue.* New York: Springer International Publishing, 470–479.

Olsen, Rachel M., Michael L. Olsen, Joseph A. Stanley, Margaret E. L. Renwick and William Kretzschmar. 2017. Methods for transcription and forced alignment of a legacy speech corpus. *Proceedings of Meetings on Acoustics,* 1–13. https://doi.org/10.1121/2.0000559

Oregon Department of Transportation Research Section. 2010. *Guide to Transcribing and Summarizing Oral Histories.* https://www.oregon.gov/odot/Programs/ResearchDocuments/guide_to_transcribing_and_summarizing_oral_histories.pdf (accessed 25 June 2023.)

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier and Michael Auli. 2019. Fairseq: A fast, extensible toolkit for sequence modelin. In Ammar Waleed, Annie Louis and Nasrin Mostafazadeh eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis: Association for Computational Linguistics, 48–53.

Pandey, Ayushi, Pamir Gogoi and Kevin Tang. Understanding forced alignment errors in Hindi-English code-mixed speech–a feature analysis. 2020. In *Proceedings of First Workshop on Speech Technologies for Codeswitching in Multilingual Communities,* 13–17. http://festvox.org/cedar/WSTCSMC2020.pdf

Pederson, Lee, Susan Leas McDaniel and Carol M. Adams eds. 1986. *Linguistic Atlas of the Gulf States*. Georgia: University of Georgia Press.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Pettarin, Alberto. 2017. *Aeneas: Automagically Synchronize Audio and Text*. https://www.readbeyond.it/aeneas/ (accessed 29 June 2023.)

Previlon, Wilermine, Alice Rozet, Jotsna Gowda, Bill Dyer, Kevin Tang and Sarah Moeller. 2024. Leveraging syntactic dependencies in disambiguation: the case of African American English. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. (Preprint available at https://doi.org/10.31234/osf.io/ph7q8).

Rahman, Jacquelyn. 2008. Middle-class African Americans: Reactions and attitudes toward African American English. *American Speech* 83/ 2: 141–76.

Rohanian, Morteza and Julian Hough. 2021. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In Chengqing Zong, Fei Xia, Wenjie Li and Roberto Navigli eds. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.* Online publication: Association for Computational Linguistics, 3693–3703.

Roller, K. 2015. Towards the 'oral' in oral history: Using historical narratives in linguistics. *Oral History*: 73–84.

Samuel Proctor Oral History Program. 2007. *Style Guide: Guidelines for Transcribing and Editing Oral Histories*. https://ufdc.ufl.edu/IR00002513/00001 (accessed 25 June 2023.)

Samuel Proctor Oral History Program. 2016. *Style Guide: Guidelines for Transcribing and Editing Oral Histories.* https://oral.history.ufl.edu/wp-content/uploads/sites/15/SPOHP-Style-Guide-2016.pdf (accessed 25 June 2023.)

Samuel Proctor Oral History Project. 2020. *Learn to Transcribe Oral History the SPOHP Way*. https://www.youtube.com/watch?v=_aKXmOLQINw (accessed 23 June 2023.)

Samuel Proctor Oral History Project. 2023. *Machen Florida Opportunity Scholars Program (MFOS).* https://oral.history.ufl.edu/projects/machen-florida-opportunity-scholars-program-mfos/ (accessed 27 June 2023.)

Santiago, Harrison, Joshua Martin, Sarah Moeller and Kevin Tang. 2022. Disambiguation of morpho-syntactic features of African American English: The case of habitual be. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, Paul Buitelaar eds. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin: Association for Computational Linguistics, 70–75.

Schiel, Florian, Christoph Draxler, Angela Baumann, Tania Ellbogen and Alexander Steffen. 2012. *The Production of Speech Corpora*. München: Open Access Ludwig-Maximilians-Universität München. https://doi.org/10.5282/ubm/epub.13693.

Schiffrin, Deborah. 2002. Mother and friends in a holocaust life story. *Language in Society* 31/3: 309–353.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stolcke, Andreas and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. *Interspeech*: 137–141.

Strong, Liz, Mary Marshall Clark and Caitlin Bertin-Mahieux. 2018. *Columbia University Oral History Transcription Style Guide*. Columbia: Columbia University Center for Oral History Research. https://incite.columbia.edu/publications-old/2019/3/13/oral-history-transcription-style-guide (accessed 25 June 2023.)

Tagliamonte, Sali A. 2004. Someth[in]'s go[ing] on!: Variable *ing* at ground zero. In Britt-Louise Gunnarsson, Lena Bergström, Gerd Eklund, Staffan Fidell, Lise H. Hansen, Angela Karstadt, Bengt Nordberg, Eva Sundergren and Mats Thelander eds. *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe.* Uppsala: Uppsala Universitet, 390–403.

Tang, Kevin. 2015. *Naturalistic Speech Misperception*. London: University College London dissertation.

Tang, Kevin and Ryan Bennett. 2019. Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (mayan). In Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren eds. *Proceedings of the 19th International Congress of Phonetic Sciences*. Canberra: Australasian Speech Science and Technology Association Inc, 1719−1723-

Tevissen, Yannis, Jérôme Boudy, Gérard Chollet and Frédéric Petitpont. 2023. Towards measuring and scoring speaker diarization fairness. *CoRR* abs/2302.09991. https://doi.org/10.48550/arXiv.2302.09991.

Thomas, Erik R. and Guy Bailey. 2015. Segmental phonology of African American English. In Jennifer Bloomquist, Lisa J. Green and Sonja L. Lanehart eds. *The*

Oxford Handbook of African American Language. Oxford: Oxford University Press, 403–419.

Whalen, Douglas H. and Joyce McDonough. 2015. Taking the laboratory into the field. *Annual Review of Linguistic*s 1/1: 395–415.

Yoon, Sunmoo, Peter Broadwell, Frederick F. Sun, Maria De Planell-Saguer and Nicole Davis. 2023. Application of topic modeling on artificial intelligence studies as a foundation to develop ethical guidelines in African American dementia caregiving. *Studies in Health Technology and Informatics* 305, 541–544.

Yuan, Jiahong and Mark Liberman. Automatic detection of "g-dropping" in American English using forced alignment. 2011. In the *2011 IEEE Workshop on Automatic Speech Recognition and Understanding*. Hawaii: Curran Associates Inc, 490–493. https://doi.org/10.1109/ASRU.2011.6163980

Zayats, Vicky, Trang Tran, Richard Wright, Courtney Mansfield and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. *Interspeech:* 3088–3092.

Ziems, Caleb, William Held, Jingfeng Yang and Diyi Yang. 2022. Multi-VALUE: A framework for cross-dialectal English NLP. *CoRR* abs/2212.08011. https://doi.org/10.48550/arXiv.2212.08011.

Corresponding author
Kevin Tang
Heinrich-Heine-University Düsseldorf
Department of English and American Studies
Sekretariat Ulrike Kayser
Geb. 23.21.02.102
Universitätsstraße 1
40225 Düsseldorf
Germany
Email: kevin.tang@hhu.de

received: July 2023
accepted: February 2024

APPENDIX A: LIST OF TEXTS BY INTERVIEWEE

	Interviewee	Gender	Age at recording	Date of interview	Place of interview	Tokens
(1)	Alexis Cooper	Female	Teen	09/21/2013	Moorhead, MS	5,231
(2)	Breyanna Hooper	Female	Teen	09/21/2013	Sunflower, MS	4,137
(3)	Cornelius Towns	Male	83	11/25/2012	Bland, FL	3,181
(4)	Darron L Edwards	Male	42	09/23/2011	Ruleville, MS	3,181
(5)	David Faison	Male	84	03/11/2016	Silver Springs, FL	5,184
(6)	Deloris Johnson	Female	65	05/25/2010	Gainesville, FL	8,316
(7)	Diana Bell	Female	60s	06/12/2009	Alachua County	4,711
(8)	Ernest Sneed	Male	Elder	09/06/2016	Alachua County	2,548
(9)	Eugene Martin	Male	70s	06/24/2016	High Springs, FL	5,323
(10)	Jeanette G. Jackson	Female	80s	02/26/2008	Alachua County	3,148
(10)	Doris Marie Perry Ryan	Female	70s	02/26/2008	Alachua County, FL	3,148
(10)	Orien A. Hills	Male	Elder	02/26/2008	Alachua County, FL	3,148
(11)	John Booth	Male	62	01/13/2009	Gainesville, FL	4,969
(12)	John Due	Male	83	06/18/2017	Unknown, MS	3,212
(13)	Shirely Felton	Female	63	06/21/2018	Fort Myers, FL	4,977
(14)	Vendarae Lewis	Female	56	07/27/2012	Bartow, FL	3,225
(15)	Yolanda Veal	Female	26	09/24/2010	Indianola, MS	4,862
(16)	Zacchaeus McEwen	Male	17	09/18/2013	McComb, MS	4,266

APPENDIX B: DATA AVAILABILITY AND AUTHORSHIP CONTRIBUTION STATEMENT

All figures are available on Open Science: http://doi.org/10.17605/OSF.IO/4HGBW. Our training materials for transcribers to become aware of AAL features and AAL in general may be found at https://doi.org/10.17605/OSF.IO/X9WHN.

Sarah Moeller and Kevin Tang contributed equally to this paper, and they served as the senior and corresponding authors. We follow the CRediT taxonomy (https://credit.niso.org/).

Conceptualization: Sarah Moeller and Kevin Tang; **Formal Analysis:** Sarah Moeller, Kevin Tang, Alexis Davis, Wilermine Previlon and Michael Bottini; **Funding acquisition:** Sarah Moeller and Kevin Tang; **Investigation:** Sarah Moeller, Kevin Tang, Alexis Davis, Wilermine Previlon and Michael Bottini; **Methodology:** Sarah Moeller, Kevin Tang, Alexis Davis, Wilermine Previlon and Michael Bottini; **Resources:** Sarah Moeller and Kevin Tang; **Software:** Wilermine Previlon; **Supervision:** Sarah Moeller, Kevin Tang and Alexis Davis; **Visualization:** Alexis Davis and Wilermine Previlon; **Writing original draft:** Sarah Moeller, Kevin Tang, Wilermine Previlon and Alexis Davis; **Writing review and editing:** Sarah Moeller, Kevin Tang, Wilermine Previlon and Alexis Davis.