A Fine-tuned ASR Model for Historical American Dialect Recordings

Steven Coats
English, Faculty of Humanities, University of Oulu
90014 University of Oulu, Finland
steven.coats@oulu.fi

Abstract

This paper introduces DASS2019_NLP, a newly cleaned and curated version of the *Digital Archive of Southern Speech*, a major historical resource for the study of Southern American English, together with six Whisper ASR models fine-tuned on the data. The 344 hours of conversational speech were recorded by fieldworkers between 1969 and 1983 across the Southern United States. Each Whisper model was fine-tuned on DASS2019_NLP, then evaluated on held-out DASS2019_NLP data, a subset of the *Corpus of Regional African American Language* (CORAAL), and a subset of Common Voice. The fine-tuned models show consistent learning trajectories and achieve an average 37% reduction in WER on in-domain data relative to baseline models. Notably, they also improve transcription accuracy on CORAAL, suggesting enhanced robustness to African American English. As expected under read vs. conversational style mismatch, accuracy on CV generally favors the OpenAI baselines. Both the DASS2019_NLP dataset and the best-performing fine-tuned model (whisper-large-v3-DASS-ct2) have been publicly released. These resources provide new tools for quantitative research in historical sociolinguistics, facilitating large-scale analyses of phonological, lexical, and grammatical change in Southern and African American English.

Keywords: Southern American English, dialectology, automatic speech recognition, Whisper, fine-tuning

1. Introduction

Advances in model architectures, the growing availability of training data, and the adoption of specialized computational infrastructures have greatly enhanced Automatic Speech Recognition (ASR) accuracy in recent years. Models trained on large datasets using end-to-end approaches such as Whisper (Radford et al., 2022) can achieve remarkably low Word Error Rates (WER), and newer architectures such as Canary-Qwen (Puvvada et al., 2024) and Granite-speech (Saon et al., 2025), as well as novel approaches to preparing training data (Peng et al., 2025), continue to push performance further.

Nevertheless, a mismatch between training and target domains can substantially degrade model accuracy, particularly for conversational, low-resource, dialectal, or historical speech, a challenge with significant implications for projects with a sociolinguistic or diachronic focus. In many ways, this problem parallels earlier concerns about corpus representativeness (Biber, 1993): just as the composition of a corpus determines the generalizability of linguistic analyses, the composition of a model's training data constrains its performance across linguistic domains, and recent work underscores the importance of tailoring pretraining corpora to accurately represent the linguistic varieties under study (Grieve et al., 2025).

In North America, decades of dialectological fieldwork have produced extensive audio archives documenting historical varieties of English, collected using sampling procedures that carefully balanced demographic and identity factors such as education, social class, ethnicity, and region. Much of this data, crucial for understanding the evolution of American English, remains untranscribed. Fine-tuned ASR models trained on carefully curated sociolinguistic datasets, where speaker and community metadata are known, offer a way to mitigate domain mismatch and extend the reach of computational sociolinguistics (Nguyen et al., 2016) to historical materials.

In this study, we fine-tune Whisper models on conversational speech drawn from historical dialectological recordings from the American South collected during the mid- to late twentieth century. We introduce a new dataset derived from these materials and evaluate a series of fine-tuned ASR models on both in-domain and out-of-domain data.

Our results show that the rate of improvement from fine-tuning does not strongly depend on model size, and that models fine-tuned on historical American English speech yield lower WERs on related dialectal data, including African American English. The fine-tuned model will have immediate applicability for the automated transcription of thousands of hours of hitherto untranscribed dialectological fieldwork recordings, and the findings of the study and associated materials will support corpus-based research into the historical development and evolution of English in the United States. Together, these findings demonstrate the value of

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1372–1381
11–16 May 2026.

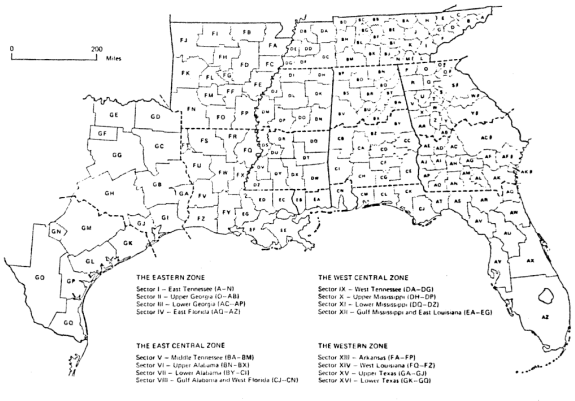

Figure 1: Survey area for LAGS. Image from (Pederson, 1985), p. 5D.

domain-specific fine-tuning for extending ASR to historically and socially meaningful varieties of English.

1.1. Background

Dialectological fieldwork in the United States has a rich and extensive tradition, particularly through the large-scale efforts of the Linguistic Atlas Project (LAP) in the 20th century. As part of this initiative, begun in 1929, thousands of in-person interviews were conducted with informants throughout the country to systematically document regional variation in American English. The project sought to provide a comprehensive linguistic portrait of contemporary English in the United States by investigating lexical, phonetic, phonological, and syntactic differences, as well as document historical folkways and cultural practices.

The extensive data collection efforts of the LAP led to the creation of several regional linguistic atlases, including the *Linguistic Atlas of New England* (Kurath et al., 1939–43, reprinted 1972), *the Linguistic Atlas of the Middle and South Atlantic States* (McDavid and O'Cain, 1980), and the *Linguistic Atlas of the Upper Midwest* (Allen, 1973–6), among others. Today, the LAP archives are housed at the University of Kentucky and the University of Georgia, where they continue to serve as a valuable resource for linguistic research and historical analysis.[1]

In the context of the broader LAP, from 1968 to 1983, fieldworkers for the *Linguistic Atlas of the Gulf States* (LAGS, Pederson et al., 1986–92) conducted and recorded 1,121 interviews with informants from Tennessee, Georgia, Florida, Alabama, Mississippi, Louisiana, Texas, and Arkansas (see Fig.1). Similar to other LAP projects, much of the content of these interviews comprised responses to a structured survey dealing with topics such as family, the weather, agricultural activities, household articles, or social connections. Although many of the survey items were compiled and published, the recordings were not transcribed. Beginning in the 2000s, heritage LAP recordings were digitized at the University of Georgia under the auspices of William A. Kretzschmar Jr. and his team. In 2012, DASS, the *Digital Archive of Southern Speech* (Kretzschmar et al., 2012), a 64-informant sample of interviews from LAGS, was made available. A manually transcribed and time-aligned version of DASS, produced in the years 2016–2019 in the context of an NSF grant, was made available in 2019 under the moniker DASS2019 (Olsen et al., 2017; Kretzschmar et al., 2019).

1.2. Related work

Data from DASS and DASS2019 have been utilized in a number of studies. Olsen et al. (2018) analyzed the /aɪ/ diphthong in DASS data, finding regional variation in vowel quality within the South. In Renwick and Stanley (2017), formant values for front vowels were extracted from DASS data, then analyzed in terms of speaker characteristics, shedding light on the dynamics of the Southern Vowel Shift and the African American Vowel Shift, two ongoing chain shifts in the vowel inventory of American English. The *whine-wine* merger in the South was analyzed using DASS data in Bridwell and Renwick (2024). Among other findings, the authors noted that the feature has merged for African-American speakers.

Fine-tuned ASR models have been created for dialectal speech in a number of languages, including Swedish, Swiss German, and Norwegian (Vesterbacka et al., 2025; Sicard et al., 2023; Kummervold et al., 2024). For English, Torgbi et al. report a significant improvement in WER on Scottish call-center recordings for a Whisper model fine-tuned on Scottish speech compared to the baseline model (Torgbi et al., 2025). For American English speech, the performance of ASR models on African American English varieties is relatively poor (Chang et al., 2024; Koenecke et al., 2020; Radford et al., 2022), with non-standard phonological features resulting in errors in transcription and alignment (Mojarad and Tang, 2025). As far as is known, models fine-tuned specifically with American dialect recordings have not yet been reported in the research literature.

The heritage recordings of LAGS and of other LAP atlases contain, in addition to cultural, geographical, and local historical information and responses to specific survey items, much spontaneous, unscripted speech, representing a valuable resource for the investigation of the phonetic and

[1] https://linguisticatlasproject.org, https://lap.uga.edu

grammatical properties of English in the United States and its diachronic development.

However, without orthographic transcripts, many potential research questions pertaining to this material cannot be easily investigated. LAGS alone comprises approximately 5,000 hours of heritage recordings that have not yet been transcribed; other LAP regional projects contain many more thousands of hours of historical American English speech that have been carefully organized and curated with detailed speaker metadata, but remain largely untranscribed. Creating accurate transcripts for this dialect material would represent a major advancement. The motivation for the creation of DASS2019_NLP was therefore threefold:

1. To confirm that fine-tuning Whisper with historical dialect recordings, specifically the DASS2019_NLP data, can substantially reduce WER compared to the baseline model.

2. To create and make available whisper-large-v3-DASS2019-ct2, a fine-tuned version of Whisper's large-v3 model for use with LAGS data as well as other historical American recordings from the LAP and other sources.[2]

3. To make DASS2019_NLP available as a dataset on Hugging Face in a format that can be easily utilized for fine-tuning and evaluating ASR systems and for other NLP tasks.[3]

The rest of the paper describes the procedures implemented for retrieving and processing the DASS2019_NLP data, fine-tuning the ASR models, and evaluating the output. In the discussion, a brief error analysis is provided, as well as remarks on limitations and caveats of this approach. The paper concludes with a future outlook for work along these lines.

2. Methods

2.1. Data Collection and Processing

The DASS2019 data were retrieved from servers hosted at the University of Georgia[4] and its directory structure recreated on a server at Finland's Centre for Scientific Computing (CSC).[5] The timed XML transcript files for each of the 408 recordings were parsed for speaker, speech turn, turn start and end times, turn transcript text, and corresponding audio file. The transcription parsing logic focused on primary speech turns; as a result, while the audio was extracted in continuous chunks, speech occurring within overlapping boundaries (marked by #) was not represented in the final training labels.[6] Consecutive single-speaker turns were then aggregated into chunks of at most 30 seconds by iteratively adding individual speaker turns. The timing information for these chunks was used to cut the audio files. The procedure resulted in 48,214 segments of transcribed audio with speaker labels and a mean segment length of 25.69 seconds. Segments were cut using the parsed timing information and resampled at 16 kHz.

The next step removed several DASS2019 annotations that could affect model accuracy. While overlap markers were handled during the initial parsing of turns, the annotations {X} (unintelligible), {NS} (non-speech such as telephone ringing or dog barking), {NW} (non-word, such as cough), and {C: comment} were removed, including any additional annotation within the corresponding curly brackets. For the annotation {D}, indicating a doubtful transcription, according to the transcriber, the brackets and D: were removed, but not the doubtful transcription; for example, "{D: tobacco shed}" was changed to "tobacco shed". The code {B}, indicating that a beep had been inserted into the audio to mask personal information such as a name or address, was retained. In addition, manual inspection revealed that some unintelligible utterances had been annotated with "?" instead of {X}; these were removed.

The dataset includes columns for segment id, audio segment, text, start time, stop time, and segment duration. Additional metadata for the informants, including geographical location, sex, age, educational attainment, ethnicity, and other variables, can be retrieved from the LAP website.[7] Because each chunk in DASS2019_NLP indexes a particular recording with a particular informant, this metadata can easily be implemented as a lookup table for downstream tasks that focus on informant characteristics (e.g., specific location, educational level, gender, race/ethnicity). The dataset is summarized in Table 1. Here, the number of speakers reflects the 64 unique informants as well as various auxiliary speakers such as spouses, neighbors, children, and interviewers, among others.

2.2. Fine-tuning and evaluation

The Whisper models `tiny`, `base`, `small`, `medium`, `large-v2`, and `large-v3` were re-

[2] https://huggingface.co/stcoats/whisper-large-v3-DASS2019-ct2

[3] https://huggingface.co/stcoats/DASS2019_NLP

[4] https://www.lap.uga.edu/Projects/DASS2019

[5] https://csc.fi

[6] Overlapping segments accounted for approximately 3.99% of the total speech content (643,206 characters out of 16.1 million total).

[7] https://www.lap.uga.edu/Projects/DASS2019/Info

Table 1: DASS2019_NLP summary

Measurement	Value
Speakers	139
Segments	48,214
Word tokens	3,084,208
Audio length (h)	344.04

trieved from Hugging Face and fine-tuned on the DASS2019_NLP data, after additional transcript processing to remove speaker labels. Models were configured for English speech transcription, using an effective batch size of 4, an initial learning rate of 5×10^{-6}, a weight decay value of 0.01, and a linear learning rate scheduler. Gradient checkpointing was enabled to optimize memory usage, and BF16 mixed precision was utilized for faster computations. The models were trained for three epochs, saving checkpoints every 200 steps. Logging was configured to report every 100 steps, and all metrics were tracked using TensorBoard.

The dataset was partitioned into training, validation, and test sets using a randomized 80/10/10 split at the segment level. This approach allows the model to adapt to the specific acoustic conditions, recording equipment, and informant idiolects present across the DASS2019 corpus. While individual speakers may appear in both the training and test sets, the specific 30-second audio segments are strictly isolated, ensuring the model is evaluated on previously unseen utterances. Training was undertaken on CSC's Lumi supercomputer with 16 AMD Instinct MI250X GPUs, each with 128 GB of memory.

For evaluation, three datasets were used: the held-out test data from DASS2019_NLP, a subset of CORAAL, the Corpus of Regional African American Language (Kendall and Farrington, 2023), and a subset of Common Voice v22 (Ardila et al., 2020).

2.2.1. CORAAL

CORAAL is a publicly available collection of sociolinguistic interview recordings and transcripts representing African American English, intended to be a resource for studying linguistic variation, style, and change within African American speech communities. We utilized a CORAAL subset of speakers from Atlanta in which transcripts were cleaned to remove interviewer speech, markup annotation, pause markers, anonymized names, and very short or very long segments.[8] The dataset comprises 3,474 audio segments, totaling 2.31 hours.

[8] https://huggingface.co/datasets/PardisTaghavi/coraal_chunked.

2.2.2. Common Voice

An additional evaluation was undertaken using a subset of Southern American English from Common Voice. To create a small set of Southern American English-accented speech, the English subcorpus of Common Voice v22 was filtered to include clips that contained any Southern U.S. state in the "accents" field of the metadata,[9] resulting in a test set of 295 segments, totaling 30m 3s in duration. This smaller evaluation test set was used to gauge generalization to read speech. We treat DASS2019_NLP as in-domain. CORAAL is evaluated as a related (near-domain) conversational set, reflecting variation in community and channel. Common Voice (Southern US) is out-of-domain due to style (read vs. conversational) and recording conditions.

3. Results and evaluation

In this section, results for the six Whisper baseline models are compared with those of the six fine-tuned models in terms of WER and CER (Character Error Rate) on the three test datasets.

For all datasets, inference used greedy decoding with the model's English *transcribe* prompt. Audio was resampled to 16 kHz mono. Decoding was capped at `max_new_tokens`=225; all remaining parameters used the `transformers` defaults (e.g., `repetition_penalty`=1.0 with no n-gram blocking) to minimize repetition while ensuring consistent decoding across models. Before scoring, we applied the Whisper English text normalizer. WER and CER were computed with the `evaluate` package.

3.1. DASS2019_NLP dataset

Transcriptions were produced from a subsample of 483 segments from the 10% held-out DASS2019_NLP test data. Figure 2 shows the post-hoc evaluation of the training checkpoints over the course of the 3 training epochs. Training step 0 represents the baseline models.

As can be seen in Fig. 2 and Table 2, for all six models, fine-tuning results in a rapid improvement in WER in the initial 1,000 training steps, followed by a gradual improvement until approximately the 5,000th step. Training after step 6,000 does not improve model performance. The improvement in WER and CER attributable to fine-tuning is consistent across model sizes for the DASS2019_NLP data (Table 3).

[9] Kentucky, Virginia, West Virginia, Missouri, Tennessee, North Carolina, South Carolina, Georgia, Florida, Alabama, Mississippi, Arkansas, Louisiana, and Texas.

Model size	Variant	DASS2019_NLP		CORAAL		CV Southern US	
		WER (%)	CER (%)	WER (%)	CER (%)	WER (%)	CER (%)
tiny	OpenAI	41.93	26.75	48.01	33.41	*25.27*	*13.73*
	Fine-tuned	**25.72**	**15.71**	**43.17**	**27.91**	**21.08**	**8.66**
base	OpenAI	32.09	20.96	36.67	26.99	***12.11***	***4.65***
	Fine-tuned	**20.85**	**13.55**	**35.67**	**23.87**	16.63	6.54
small	OpenAI	22.60	14.67	27.09	19.70	***7.58***	***2.80***
	Fine-tuned	**14.28**	**9.03**	**25.46**	**17.35**	9.21	3.12
medium	OpenAI	20.17	13.46	27.96	20.01	***5.45***	***1.89***
	Fine-tuned	**12.47**	**8.00**	**20.81**	**14.98**	7.15	2.55
large-v2	OpenAI	19.64	13.14	24.54	18.96	***4.66***	***1.62***
	Fine-tuned	**12.09**	**7.62**	**20.14**	**14.72**	5.72	1.93
large-v3	OpenAI	18.50	12.18	22.99	17.17	**5.05**	**1.72**
	Fine-tuned	***11.83***	***7.44***	***19.11***	***14.24***	6.52	1.99

Table 2: WER and CER (%) for Whisper models and their fine-tuned counterparts across three evaluation sets. Within each model size, the better value between OpenAI and Fine-tuned is **bolded** (lower is better). The overall best value within each dataset/metric block is *italicized*.

Model	DASS2019_NLP (Δ%)		CORAAL (Δ%)		CV Southern US (Δ%)	
	WER	CER	WER	CER	WER	CER
tiny	38.66	41.27	10.08	16.46	16.58	36.92
base	35.03	35.35	2.73	11.56	−37.33	−40.65
small	36.81	38.45	6.02	11.93	−21.50	−11.43
medium	38.18	40.56	25.59	25.15	−31.19	−34.92
large-v2	38.44	42.01	17.93	22.36	−22.75	−19.14
large-v3	36.05	38.92	16.88	17.06	−29.11	−15.70

Table 3: Percent improvement (Fine-tuned vs. OpenAI) in WER and CER by model size; positive values indicate lower error after fine-tuning.

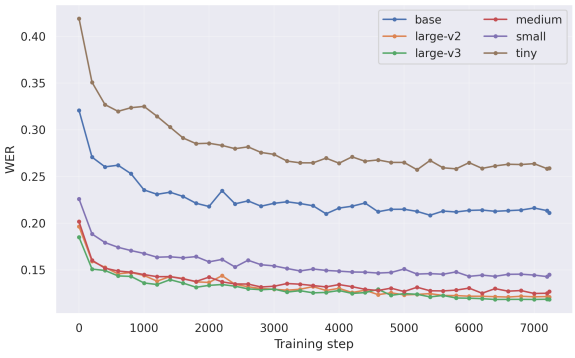

Figure 2: WER by training step, DASS2019_NLP test data

3.2. CORAAL dataset

Fine-tuning improves performance on CORAAL across all model sizes, though the gains are smaller than for in-domain data (Table 2). Relative to the OpenAI baselines, WER reductions range from 2.7–25.6% and CER from 11.6–25.2% (Table 3). The best absolute CORAAL scores are obtained by large-v3 (fine-tuned) with WER 19.11% and CER 14.24%. These results suggest that adaptation on historical conversational speech transfers within conversational style to related communities and channels, albeit with an attenuated effect compared to strictly in-domain evaluation.

3.3. Common Voice dataset

On Common Voice (Southern US), which consists of read speech, fine-tuning generally hurts performance: five of six model sizes show higher error than the OpenAI baselines (WER change −21.5% to −37.3%; CER −11.4% to −40.7%), with the exception of tiny, which improves (WER +16.6%, CER +36.9%) (Table 3). The best absolute CV scores come from the OpenAI large-v2 baseline (WER 4.66%, CER 1.62%). This pattern is consistent with style mismatch: models adapted to

conversational, noisy, spontaneous speech may over-specialize and lose accuracy on clean, read speech.

4. Discussion

4.1. Overall effects of fine-tuning

Fine-tuning on DASS2019_NLP yields consistent in-domain gains across all model sizes because the training and test distributions match (spontaneous Southern U.S. English, similar acoustic properties and annotation conventions). Transfer to CORAAL is positive but more variable: although it is also conversational, differences in speaker population, microphone/noise profiles, and discourse features reduce the benefit for smaller models and reward larger models with more capacity.

In contrast, performance on Common Voice (read speech) often degrades because fine-tuning steers the model toward colloquial pronunciations and disfluencies that are scarce in read speech. These findings are consistent with broader evidence that domain-specific fine-tuning improves ASR robustness and transferability when linguistic and acoustic conditions differ between training and evaluation data. Prior work on accent and domain adaptation similarly shows that targeted fine-tuning on dialectal or conversational speech can substantially reduce error rates compared with generic pretraining (Qian et al., 2022; Maison and Esteve, 2023).

4.2. Dialectal relatedness (SAmE–AAE)

The transfer we see into CORAAL is plausibly helped by well-documented overlap between Southern American English (SAmE) and African American English (AAE). Because AAE developed largely in the U.S. South, long and intense contact produced shared features such as monophthongization of diphthongs or mergers (Labov et al., 2008; Wolfram and Schilling, 2015; Green, 2002; Rickford, 1999; Thomas, 2007). These contact-historical ties (plantation South, subsequent migration patterns) mean SAmE-trained models can partially generalize to AAE data.

4.3. Error analysis

Examination of model outputs for selected passages in the held-out test dataset provides some insights into the types of errors that occur with the models. Here, we compare transcript outputs for our fine-tuned whisper-large-v3-DASS2019-ct2 model with outputs from the baseline whisper-large-v3. Table 4 shows examples from the 4,821 held-out segments. The model outputs do not retain the speaker labels, which were removed from the training data for fine-tuning, nor are they normalized. Links to the corresponding audio are provided in footnotes. Preliminary error categorization suggests that the fine-tuned model's performance gain is twofold: it improves the retention of verbatim disfluencies (e.g., filler particles and discourse or hesitation markers often elided by the base model), but also provides higher accuracy for domain-specific content words, such as regional vocabulary and non-standard morphological variants (see Examples 2 and 3).

Example (1) exhibits a typical characteristic of the Whisper base model: it tends to normalize utterances by eliding discourse and filler particles. In this excerpt,[10] the baseline model has changed *mm-kay* to *okay*, omitted the filler *um*, and elided a superfluous *to*, while the fine-tuned model has correctly transcribed these forms. Default Whisper models, it has been suggested, have been designed to emphasize clarity of meaning rather than verbatim accuracy of utterances (Lea et al., 2023). In Example (2),[11] an excerpt in which the interviewer is discussing words used to describe groups of people with a 57-year-old English/Spanish bilingual female informant from Laredo, Texas, the baseline model has misspelled the Spanish borrowing *pachucas* and rendered *chucs* as *chooks*. The baseline model has also incorrectly transcribed *Do you ever hear of them* as *Do you remember they were*.[12]

The fine-tuned model produces superior transcripts for some segments. Example (3) is an excerpt from an interview with a 73-year-old man from Huntsville, Texas, recorded in 1976.[13] In the excerpt, the informant, a farmer, uses the nonstandard weak preterite form *growed*, before uttering *grew* a few seconds later; the Whisper base model has corrected the first utterance to standard *grew*, as well as eliding *uh*, while the fine-tuned model transcribes the correct forms. The correct transcription of nonstandard forms such as these is especially important for tracing the development of American English dialects in the Southern US.

Example (4) is a segment from an interview with a 74-year-old speaker from Encinal, Texas.[14] Here, the default model has made a simple error, transcribing *iodine* as *I don't*, while the fine-tuned model correctly transcribes the segment.

[10] 863_2_hpf.wav, beginning at 00:35:52.5

[11] 893_5_hpf.wav, beginning at 00:6:31.

[12] In addition, the manual transcriber has failed to transcribe *and with a lot of grease or something on it*.

[13] 856C_4_hpf.wav, beginning at 00:09:50.

[14] 894_3_hpf.wav, beginning at 00:28:05.

Table 4: Examples of transcription differences between base and fine-tuned models

Ref.	Mm-kay. Any, um, things that people used to make to, similar to a bedspread?	
Base	Okay. Any things that people used to make similar to a bedspread?	(1)
Fine-Tuned	Mm-kay. Any, um, things that people used to make to similar to a bedspread?	
Ref.	<893>: and with their hair real slick <Interviewer>: Mm-hmm. <893>: And we used to call them pachucas. <Interviewer>: Mm-hmm. <893>: But <Interviewer>: Do you ever hear of them just called chucs? <893>: Chucs? <Interviewer>: Uh-huh.	
Base	And with their hair real slick and with a lot of grease or something on it. And we used to call them pachukas. But... Do you remember they were just called chooks? Chooks.	(2)
Fine-Tuned	And uh with their hair real slick and with a lot of grease or something on it. Mm-hmm. And we used to call them pachucas. Mm-hmm. But uh. Do you ever hear them just called chucs?. Chucs. Uh-huh.	
Ref.	<Interviewer>: and uh what other crops did you grow around here? <856C>: well that's about all I growed was cotton and corn, that's all I grew	
Base	And what other crops did you grow around here? Well, that's about all I grew was cotton and corn. That's all I grew.	(3)
Fine-Tuned	and uh what other crops did you grow around here. well that's about all I growed was cotton and corn that's all I grew.	
Ref.	<Interviewer>: Say if you had a little cut on your finger a brown liquid medicine you could put on that stains a lot <894>: Iodine	
Base	Say if you had a little cut on your finger, a brown liquid medicine you could put on that stings a lot. I don't.	(4)
Fine-Tuned	Say if you had a little cut on your finger a brown liquid medicine you could put on that stings a lot. Iodine	

5. Summary and outlook

The DASS2019_NLP dataset brings Linguistic Atlas Project data into the era of large language models, opening up new possibilities for training and fine-tuning models with heritage dialectology data from the American South. The whisper-large-v3-DASS2019-ct2 model, fine-tuned on the dataset, shows a significant improvement in WER on in-domain data, compared to the baseline, as well as improvement on an unrelated AAE dataset. These resources will facilitate research into the linguistic properties of Southern American English.

While the current evaluation utilizes pre-segmented audio, applying this model to raw, untranscribed recordings would require an integrated pipeline featuring voice activity detection (VAD) and speaker diarization. Future work could implement a multi-stage approach where acoustic diarization (e.g. with pyannote, Bredin 2023) is combined with LLM-assisted speaker role correction (see, e.g., Wang et al. 2024; Cheng et al. 2025) using the linguistic context of the transcript to distinguish between 'Interviewer' and 'Informant'. Such a procedure could automate the transcription of the vast archives of untranscribed Linguistic Atlas recordings.

Whisper-large-v3-DASS2019-ct2 more accurately records word repetitions, fillers, and self-repairs than the baseline model and is also better at transcribing content words. For Southern American English speech, a variety that can differ significantly from standard American English in its phonetics/phonology and grammar, whisper-large-v3-DASS2019-ct2 achieves fewer errors.

The fine-tuned model can be used to generate transcripts for the almost 5,000 hours of LAGS recordings that have yet to be transcribed, as well as for other heritage dialect recordings in the LAP collections. Accurate transcripts for these interviews could enable new analyses of the historical development of American English phonetics/phonology, lexis, and grammar, especially in terms of regional variation. The methods used to create DASS2019_NLP and to fine-tune the model could be applied to other existing resources of transcribed dialectological field recordings or sociolinguistic interviews such as the Sociolinguistic Archive and Analysis Project (Kendall, 2007).

In a broader perspective, it is hoped that DASS2019_NLP and whisper-large-v3-DASS2019-ct2, by making these heritage recordings more accessible and providing an ASR model suitable for these kinds of materials, will contribute to the continued research of an important aspect of the intangible cultural heritage of the American South, namely the speech of its inhabitants, by enabling systematic linguistic analyses of historical dialect recordings and

facilitating new insights into how these speech patterns continue to evolve over time.

6. Caveats and limitations

The audio quality of the DASS2019 recordings (and the LAGS recordings from which they were sampled) varies greatly. The quality of the recordings, made in informants' homes with a microphone and a reel-to-reel tape recorder, is affected by factors including room acoustics, background noise, ability of the interviewer to correctly calibrate the recording apparatus, and subsequent degradation of the physical tape prior to digitization in the 2000s. Although the whisper-large-v3-DASS2019-ct2 model results in an improved WER for many interviews, for recordings of poor quality, the ASR transcripts will still contain many errors.

Examination of the transcriptions and the corresponding audio revealed many errors in the manual transcripts. For example, the recording 494_5_hpf.wav, beginning at 00:09:27, from a passage in which the interviewer is discussing farming and plowing with the informant and the informant's husband, was transcribed manually as *{X} we break much ground down here.*, where {X} is the code for unintelligible. Manual inspection reveals that the actual speech in the segment is *'course they don't do, they don't even break much ground now here.* Both the base and the fine-tuned models provide mostly accurate transcriptions of the segment: *Well, if they don't do, they don't even break much ground down here* for the base model, and the slightly better *Well if they don't do they don't even break much ground now here* for the fine-tuned model.

Other inaccuracies in the manual transcripts include *whole {X} porch* for *whole side of pork*, *You ever use the word gully or ravine or column?* instead of *You ever use the word gully or ravine or hollow?*, and many others. In many cases, both the base model and the fine-tuned model correctly transcribe items that the manual transcriber has incorrectly transcribed, but ultimately, the extent to which inaccuracies in the gold-standard manual transcripts affect the fine-tuned model is not clear. In general, manual comparisons of the base and fine-tuned transcripts for test data showed the fine-tuned model to be more accurate.

One final consideration concerns the data processing pipeline: the procedure for the preparation of training data focused on primary speech turns, meaning that while audio was extracted in continuous chunks, any speech occurring within overlapping boundaries (marked in the source files by the # annotation) was not represented in the training labels. This effectively treated overlapping vocalizations as non-transcribed acoustic context. While this reduced the total volume of available training text, the model's performance suggests that it successfully learned to prioritize the primary speaker, contributing to the observed improvements in transcription accuracy even in naturalistic, conversational audio.

Despite these limitations, the results demonstrate that fine-tuning ASR models on carefully curated dialect data can substantially improve transcription accuracy for historical speech. The DASS2019_NLP dataset and the whisper-large-v3-DASS2019-ct2 model therefore provide both a practical resource for working with Linguistic Atlas recordings and a methodological template for extending ASR to other heritage dialect archives.

7. Acknowledgements

This work was supported by the European Union – NextGenerationEU instrument and was funded by the Research Council of Finland under grant number 358720. Computational resources were provided by Finland's Centre for Scientific Computing.

8. Bibliographical References

Harold B. Allen, editor. 1973–6. *The Linguistic Atlas of the Upper Midwest. 3 vols.* University of Minnesota Press, Minneapolis.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus.

Doug Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Hervé Bredin. 2023. Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe. In *INTERSPEECH 2023*, pages 1983–1987.

Keiko Bridwell and Margaret E. L. Renwick. 2024. Race, place, and education: Charting the wine-whine merger in the U.S. South. *American Speech*, 99(4):441–467.

Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R. Mortensen. 2024. Self-supervised speech representations still struggle with African American Vernacular English. In *Interspeech 2024*, pages 4643–4647.

Luyao Cheng, Hui Wang, Chong Deng, Siqi Zheng, Yafeng Chen, Rongjie Huang, Qinglin Zhang, Qian Chen, Xihao Li, and Wen Wang. 2025. Integrating audio, visual, and semantic information for enhanced multimodal speaker diarization on multi-party conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19914–19928, Vienna, Austria. Association for Computational Linguistics.

Lisa J. Green. 2002. *African American English: A linguistic introduction*. Cambridge University Press, Cambridge.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411. Publisher: Frontiers.

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

Allison Koenecke, 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. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

William A. Kretzschmar, Margaret E. L. Renwick, Lisa M. Lipani, Michael L. Olsen, Rachel M. Olsen, Yuanming Shi, and Joseph A. Stanley. 2019. Transcriptions of the Digital Archive of Southern Speech.

Per E. Kummervold, Javier de la Rosa, Freddy Wetjen, Rolv-Arild Braaten, and Per Erik Solberg. 2024. Whispering in Norwegian: Navigating orthographic and dialectic challenges. In *Interspeech 2024*, pages 3984–3988.

Hans Kurath, Marcus L. Hansen, Bernard Bloch, and Julia Bloch, editors. 1939–43, reprinted 1972. *Linguistic Atlas of New England. 3 vols.* Brown University Press, Providence, RI.

William Labov, Sharon Ash, and Charles Boberg. 2008. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. De Gruyter Mouton.

Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Lucas Maison and Yannick Esteve. 2023. Improving accented speech recognition with multi-domain training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Raven I. McDavid and Raymond K. O'Cain, editors. 1980. *Linguistic Atlas of the Middle and South Atlantic States, Fascicles 1–2*. University of Chicago Press, Chicago.

Hamid Mojarad and Kevin Tang. 2025. Automatic Speech Recognition of African American English: Lexical and Contextual Effects. In *Interspeech 2025*, pages 3883–3887.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska De Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Rachel M. Olsen, Michael L. Olsen, and Margaret E. L. Renwick. 2018. The impact of sub-region on /ai/ weakening in the U.S. South. *Proceedings of Meetings on Acoustics*, 31(1):060005.

Rachel M. Olsen, 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*, 30(1):060001.

Lee Pederson. 1985. A matrix for word geography (LAGS working papers, 3rd series, introduction).

Lee Pederson, Susan L. McDaniel, and Carol M. Adams, editors. 1986–92. *Linguistic Atlas of the Gulf States. 7 vols.* University of Georgia Press, Athens, Georgia.

Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning. In *Interspeech 2025*, pages 2225–2229.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is More: Accurate Speech Recognition

Translation without Web-Scale Data. In *Interspeech 2024*, pages 3964–3968.

Yanmin Qian, Xun Gong, and Houjun Huang. 2022. Layer-wise fast adaptation for end-to-end multi-accent speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2842–2853.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Margaret E. L. Renwick and Joseph A. Stanley. 2017. Static and dynamic approaches to vowel shifting in the Digital Archive of Southern Speech. *Proceedings of Meetings on Acoustics*, 30(1):060003.

John Rickford. 1999. *African American Vernacular English: Features, evolution, educational implications*. Wiley.

George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Morais, Gakuto Kurata, Hagai Aronowitz, Ibrahim Ibrahim, Jeff Kuo, Kate Soule, Luis Lastras, Masayuki Suzuki, Ron Hoory, Samuel Thomas, Sashi Novitasari, Takashi Fukuda, Vishal Sunder, Xiaodong Cui, and Zvi Kons. 2025. Granite-speech: Open-source speech-aware LLMs with strong English ASR capabilities.

Clement Sicard, Kajetan Pyszkowski, and Victor Gillioz. 2023. Spaiche: Extending state-of-the-art ASR models to Swiss German dialects.

Erik R. Thomas. 2007. Phonological and phonetic characteristics of African American Vernacular English. *Language and Linguistics Compass*, 1(5):450–475.

Melissa Torgbi, Andrew Clayman, Jordan J. Speight, and Harish Tayyar Madabushi. 2025. Adapting Whisper for regional dialects: Enhancing public services for vulnerable populations in the United Kingdom.

Leonora Vesterbacka, Faton Rekathati, Robin Kurtz, Justyna Sikora, and Agnes Toftgård. 2025. Swedish Whispers; Leveraging a Massive Speech Corpus for Swedish Speech Recognition. In *Interspeech 2025*, pages 758–762.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. 2024. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models. In *Interspeech 2024*, pages 3754–3758.

Walt Wolfram and Natalie Schilling. 2015. *American English: Dialects and variation, 3rd Edition*. John Wiley & Sons.

9. Language Resource References

Tyler Kendall and Charlie Farrington. 2023. *The Corpus of Regional African American Language*. PID https://doi.org/10.7264/1ad5-6t35.

William A. Kretzschmar and Paulina Bounds and Jacqueline Hettel and Steven Coats and Lee Pederson and Lisa Lena Opas-Hänninen and Ilkka Juuso and Tapio Seppänen. 2012. *Digital Archive of Southern Speech*. Linguistic Data Consortium. PID https://doi.org/10.35111/5bnt-r659.