MD_NLP: Reconstructing an Australian English Heritage Dialect Corpus from the Mitchell-Delbridge Recordings through LLM-Assisted Speaker Attribution

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

Abstract

We present MD_NLP, a discourse-annotated and georeferenced corpus derived from the Mitchell–Delbridge (MD) recordings, a foundational archive of mid-20th-century Australian English. The corpus comprises word-aligned narrative recordings from 7,735 secondary school pupils across 327 locations, enriched with structured sociodemographic metadata (e.g., sex, birthplace, parental background) and geocoded institutional coordinates. The narratives were reconstructed from archival audio using an integrated pipeline combining WhisperX-based automatic speech recognition, neural speaker diarization, LLM-assisted discourse-role correction, and Montreal Forced Aligner (MFA) boundary refinement. Evaluation on manually annotated data shows that incorporating an LLM-based reasoning step improves turn-level speaker-role attribution from 62.70% (acoustic diarization alone) to 95.68%. Unlike prior uses of the MD archive, which focused on controlled sentence materials, MD_NLP makes the spontaneous narrative component accessible for large-scale analysis. The resulting resource supports research on regional and socially conditioned variation, discourse structure, and corpus phonetics in Australian English. The proposed architecture is directly transferable to other legacy dialect archives, providing a practical pathway for transforming interview-based recordings into temporally aligned, speaker-consistent corpora.

Keywords: dialect corpora, Australian English, speaker diarization, forced alignment, discourse annotation, large language models

1. Introduction

Automatic Speech Recognition (ASR) systems are increasingly used to generate transcripts for dialectal speech in corpus phonetics and variationist linguistics (Ahn et al., 2023; Coto-Solano et al., 2021; Coats et al., 2025a; Ljubešić et al., 2024). However, the transformation of raw ASR output into research-ready dialect resources remains underdeveloped. For many legacy collections, usable corpora require not only transcription, but also reliable speaker attribution, accurate word-level timing, and discourse-structural annotation.

This paper introduces MD_NLP, a new discourse-annotated and word-aligned version of the Mitchell and Delbridge (MD) recordings (Mitchell and Delbridge, 1998). The MD collection comprises recordings of 7,735 Australian secondary school pupils from 327 schools across Australia, recorded in 1959/60 and digitized at the University of Sydney in 1998. The collection constitutes one of the largest geographically distributed archives of mid-20th-century Australian English (AusE). Despite its importance for the study of AusE, including regional and social variation, the narrative component of the corpus has remained largely inaccessible for large-scale analysis due to the absence of transcripts and the variable acoustic quality of the archival recordings.

MD_NLP addresses this gap by providing speaker-labeled, word-aligned transcripts of the MD narratives with precise temporal metadata. In addition, the corpus integrates structured sociodemographic and geographic information derived from the original Mitchell–Delbridge archival records, including sex, birthplace, parental background, school location, and geocoded coordinates. The resulting resource supports not only phonetic investigation, but also lexical, grammatical, discourse-level, and spatial sociolinguistic analyses.

Over the past two decades, research on Australian English (AusE) has increasingly examined regional phonetic differentiation (Schmidt et al., 2021; Loakes et al., 2017; Coats et al., 2025b; Cox and Palethorpe, 2019). Yet large-scale, geographically distributed spontaneous speech data remain scarce. By making nationally distributed narrative recordings from 1959/60 accessible in a temporally aligned and georeferenced format, MD_NLP enables systematic investigation of regional and socially conditioned variation in historical AusE.

To make this historical, geographically distributed material computationally accessible, we developed an integrated processing pipeline. The corpus was constructed using a hybrid pipeline that combines WhisperX-based automatic speech recognition, neural speaker diarization, LLM-assisted

discourse role inference and diarization correction (Gemini), and high-precision forced alignment using the Montreal Forced Aligner (MFA), resulting in automated segmentation into short, role-consistent narrative units.

Building on recent work that leverages Large Language Models (LLMs) to refine speaker attribution in ASR transcripts (Wang et al., 2024; Cheng et al., 2025), our approach addresses two core challenges in dialect resource construction. First, neural diarization models based solely on acoustic information, such as Pyannote (Bredin, 2023; Plaquet and Bredin, 2023), frequently misattribute interviewer speech or collapse speakers in archival classroom-style recordings. We incorporate a reasoning-oriented LLM step to correct these errors by exploiting interactional structure (e.g., question–answer sequences), substantially improving speaker-role consistency. Second, we integrate high-precision word-level alignment via the Montreal Forced Aligner (McAuliffe et al., 2017), producing temporal annotations suitable for corpus-phonetic and other variationist analyses that require fine-grained timing information.

MD_NLP provides both a research-ready resource for Australian English dialect research and a transferable pipeline for converting legacy dialect recordings into temporally precise, discourse-annotated corpora.

2. Previous Work

2.1. Use of the MD Recordings in Australian English Research

The MD archive comprises read word lists, two controlled sentence passages, and spontaneous narrative recordings. To date, research has focused exclusively on the controlled sentence materials, while the narrative component has not been systematically analyzed.

Cox et al. (2014) examined the development of /i:/ in the speech of young female Sydneysiders by comparing 168 realizations in the word *speed* from the MD read sentence with an equivalent number of recordings made in the 2000s. The authors found that while the overall quality of the vowel had not changed substantially over time, differences between "broad" and "cultivated" realizations appeared to have lessened. In Cox et al. (2024), static and dynamic formant measurements were taken for ten monophthongs extracted from the MD sentence materials and compared with equivalent recordings from 1989/90 and the 2000s/2010s, shedding light on the diachronic development of Australian English vowels.

The limited use of the narrative material is largely attributable to the absence of transcripts. Manual transcription of thousands of archival recordings is time-consuming and costly, and the variable acoustic quality of the MD recordings poses additional challenges for automated processing.

2.2. LLM-Assisted Diarization and Speaker Attribution

Recent work has explored the use of Large Language Models (LLMs) to improve speaker attribution in ASR transcripts. Wang et al. (2024) introduced DiarizationLM, which leverages semantic and contextual reasoning to correct diarization errors through textual post-processing. Similarly, Cheng et al. (2025) demonstrated that LLM-based approaches can improve speaker consistency by exploiting discourse structure beyond purely acoustic cues.

These studies highlight limitations of diarization systems based solely on acoustic information. In interview-style or classroom recordings, short interviewer turns are frequently misattributed or merged with longer responses. Such errors are particularly problematic for sociolinguistic research, where reliable speaker separation and discourse-role identification are essential.

In addition, although modern ASR architectures such as Wav2Vec 2.0 (Baevski et al., 2020) and NeMo-based systems (Harper et al., 2023) provide word-level timestamps, these timings are primarily optimized for transcription accuracy rather than phonetic precision. For corpus-phonetic research, even small boundary inaccuracies can affect segmental measurements.[1]

Our work extends these lines of research by integrating LLM-based diarization correction with high-precision forced alignment using the Montreal Forced Aligner (MFA), producing temporally refined and speaker-consistent transcripts suitable for dialect and variationist analysis.

3. Source Material

The Mitchell and Delbridge recordings were made by schoolteacher volunteers in 1959 and 1960 in 327 locations across all Australian states and territories.

The recordings contain, for each informant, two to four components:

- The informant reading the words *beat, boot, say, so, high, and how*.
- The informant reading the sentence *Let's pick a good spot near the water and pass the morning surfing and relaxing in the sun*.

[1] See, e.g., MFA alignment benchmarks.

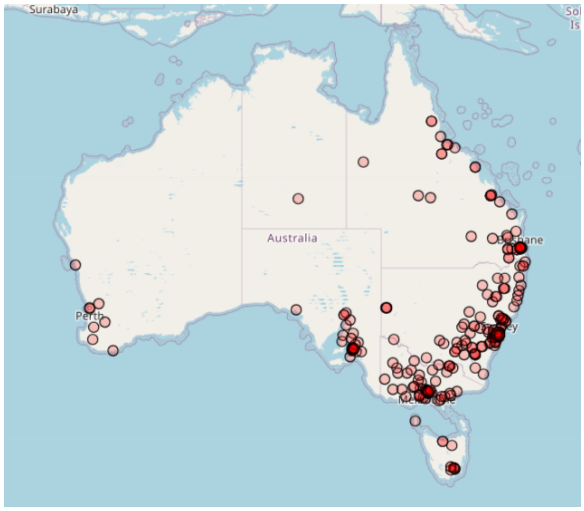

Figure 1: Sample locations for the MD narrative recordings

- The informant reading the sentence *The plane flew down low over the runway then increased speed and circled the aerodrome a second time*.
- The informant providing a short narrative on a topic of their choice. In many recordings, this takes the form of a brief interaction between the pupil and the recording teacher; in others, it consists primarily of a monologic narrative.

Recordings made in 1960 generally include all four components, whereas those made in 1959 may contain only the word list and the narrative.

MD_NLP comprises the narrative recordings from the archive. These data are particularly suitable for sociolinguistic and discourse-oriented research, as they contain spontaneous speech, question–answer sequences, and turn-taking phenomena, alongside detailed informant metadata including school, locality, and parental background. At the same time, the archival recording conditions and lack of transcripts present challenges for automated processing.

4. Pipeline Overview

Each input file is a single MD narrative WAV. We distinguish (i) WhisperX *turns* (initial segments), (ii) MFA *alignment windows* (short units used only for forced alignment), and (iii) released *dataset chunks* (final $\leq$12-token, role-consistent segments). Lexical forms always remain the WhisperX tokens; MFA is used only to refine timestamps when alignment succeeds.

Figure 2 outlines the processing pipeline. The system integrates ASR, diarization, LLM-based discourse reasoning, and forced alignment in a reconstruction-oriented workflow that preserves original tokenization while improving temporal precision.

4.1. Automatic Speech Recognition and Alignment

Preliminary transcripts were created using WhisperX (Bain et al., 2023), an implementation of Whisper (Radford et al., 2023) that utilizes the faster-whisper bindings (Klein and Ashraf, 2023) and Wav2Vec2 (Baevski et al., 2020) for word and phone alignment. We used the large-v3 Whisper model. WhisperX timestamps are treated as coarse alignment; we use MFA forced alignment to improve boundary precision for downstream phonetic and timing-sensitive analyses.

4.2. Speaker Diarization and Text Normalization

Speaker turns were identified with Pyannote.audio's speaker-diarization-community-1 model (Bredin, 2023; Plaquet and Bredin, 2023). The conditions of the MD recordings were highly variable in terms of recording equipment, microphone placement and recorded material. Essentially, each school decided how to record the short narrative, resulting in various outcomes for the resulting audio. In some recordings, a microphone was equally spaced between the teacher and the student informant, resulting in both speakers being audible. In others, the microphone was oriented towards the student. In these recordings, the speech of the interviewing teacher is often faint or inaudible. In some recordings, only the student speech has been recorded; presumably due to erasure of the segments containing the interviewer's speech.

We therefore adopted a flexible approach to recording diarization: pyannote's settings were defined to identify a minimum of 1 speaker and a maximum of two speakers. The output for most of the recordings was two speaker labels (SPEAKER_00 and SPEAKER_01). For the recordings with only faint/inaudible interviewer speech, and for those from which the interviewer's speech was removed, pyannote's output was the single speaker label SPEAKER_00.

The initial diarized transcripts were then normalized to ensure compatibility with MFA dictionaries, including expansion of symbols (e.g. "50%" $\rightarrow$ "50 percent") and removal of non-lexical punctuation. Each transcript was then converted to a TextGrid file with speaker tiers corresponding to the number of speakers detected in the previous step, resulting in an initial word-level transcript with timestamps.

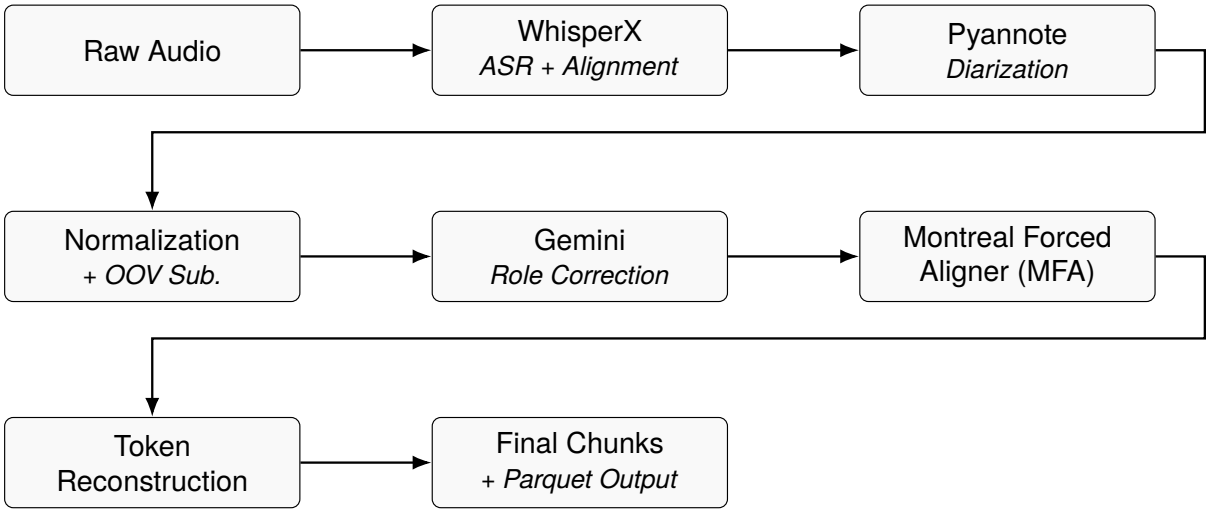

Figure 2: The integrated processing pipeline. Raw audio is transcribed with WhisperX, diarized with Pyannote, and corrected for discourse roles using an LLM. Word-level alignment is refined with MFA on short ($\leq$12-token) windows, after which original WhisperX tokens are reconstructed with selectively updated timestamps.

4.3. LLM-Assisted Discourse Role Inference

While the initial, acoustic-information-based diarization provided speaker separation, many of these labels were erroneous. In addition, the preliminary diarization could not determine discourse roles. In the MD setting, distinguishing between interviewer (Teacher) and respondent (Student) is essential for downstream analyses. We therefore applied an LLM-based reasoning step using Gemini (2.5-flash) to infer discourse roles from transcript context. The model was prompted with the following:

> Stage 1: If turns appear to be answers but no questions are present, set `inaudible_teacher=true` and label all turns STUDENT.
> Stage 2: Otherwise correct speaker attribution: Teachers ask questions; Students provide narrative/answers. Roles should be consistent per diarization label (e.g., `SPEAKER_00`).

A preliminary gate handles recordings where interviewer speech is faint/absent: if turns appear to be answers but no questions are present, all turns are labeled STUDENT to avoid forcing a spurious Teacher role. The LLM returned a structured JSON specifying corrected role assignments per turn. A qualitative example of this correction process is shown in Figure 3.

This step operationalized interpretive reasoning rather than purely acoustic modeling.

4.4. Forced Alignment and Token Reconstruction

Although WhisperX provides word-level timestamps, we further refined alignment using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), with version 3.0.0 of its UK English acoustic model, grapheme-to-phoneme model, and pronunciation dictionary. To increase alignment stability and avoid long-span decoding failures, alignment windows were capped at 12 tokens and further split on pauses > 0.8s or window duration > 8.0s to improve MFA stability.

Before MFA, out-of-vocabulary tokens were replaced with an in-vocabulary placeholder *for MFA input only*. In reconstruction, the original WhisperX token is always restored; if a token was placeholder-aligned (or missing), its original WhisperX timing is retained.

After MFA alignment, word intervals were reconstructed by mapping aligned tokens back onto the original Whisper word sequence. Tokens that were dropped or substituted for alignment were restored with their original Whisper timestamps, ensuring that no Whisper tokens were discarded, original lexical forms were preserved, and alignment precision was improved where MFA word intervals were available. Otherwise, Whisper timings were retained. This reconstruction step produced a temporally refined but lexically faithful word-level representation.

Baseline: WhisperX + Pyannote (Acoustic Diarization Only)

SPEAKER_01: Could you describe water skiing, what you do and what gear you need and where you do it?
SPEAKER_00: Well, the gear, first you need a boat, which seems essential.
SPEAKER_00: And the boat ranges from about, say, 13, 14 feet onwards.
SPEAKER_00: If you're going to learn to ski, you need two skis.
SPEAKER_00: How fast do you do it on that?
SPEAKER_00: Oh, about 30 is a pretty good speed because you can do certain different things on it when you're up.
SPEAKER_00: Do you hurt yourself if you come off?
SPEAKER_00: Oh, I've never hurt myself except once when I land on my head.

Full Pipeline: LLM-Assisted Role Attribution

Teacher: Could you describe water skiing, what you do and what gear you need and where you do it?
Student: Well, the gear, first you need a boat, which seems essential.
Student: And the boat ranges from about, say, 13, 14 feet onwards.
Student: If you're going to learn to ski, you need two skis.
Teacher: How fast do you do it on that?
Student: Oh, about 30 is a pretty good speed because you can do certain different things on it when you're up.
Teacher: Do you hurt yourself if you come off?
Student: Oh, I've never hurt myself except once when I land on my head.

Figure 3: Qualitative comparison of acoustic diarization and the LLM-assisted pipeline. The baseline system collapses interviewer and respondent speech into a single speaker label after the initial turn. Interviewer questions (highlighted in red) are incorrectly attributed to the same speaker as narrative responses. The LLM-assisted step recovers consistent Teacher–Student roles by exploiting interactional structure.

5. Dataset Structure

5.1. Segment-Level Structure

The final dataset is stored as Parquet files with corresponding audio segments, and is available as a Hugging Face dataset.[2] Each row corresponds to a short ($\leq$12-token), role-consistent segment and contains: (i) interviewee identifier, (ii) start and end time (in seconds), (iii) discourse role (Student/Teacher), (iv) transcript text, (v) word-level tokens with timestamps, and (vi) a path to the corresponding audio segment.

Word-level tokens are stored as structured objects containing lexical form, start time, and end time, preserving the original WhisperX tokenization while incorporating refined timestamps where available.

5.2. Sociodemographic and Geographic Metadata

In addition to transcript-level annotations, MD_NLP incorporates structured informant metadata derived from the original Mitchell–Delbridge archival records provided by the University of Sydney. The original metadata were distributed in heterogeneous spreadsheet formats and required normalization and cleaning prior to integration.

For each informant, the metadata include: identifier (ID#), state, town/city, institution (school), recording year, tape identifier, sex, birthplace, father's birthplace, father's occupation, and mother's birthplace. Institutional location fields were normalized and geocoded, resulting in latitude–longitude coordinates associated with each recording site.

These metadata are linked to each narrative segment via the informant identifier, enabling analyses that integrate discourse structure, phonetic timing, demographic attributes, and geographic location within a unified sociolinguistic resource.

5.3. Corpus Size

Table 1 summarizes corpus size and discourse distribution. The dataset comprises 214.14 hours of recordings, of which 137.95 hours are attributed to active speech following diarization and role assignment.

Table 1 shows the size of the dataset.

The corpus reflects broad geographic and gender representation across Australian states (Table 2).

[2] https://huggingface.co/datasets/stcoats/MD_NLP

Metric	Student	Teacher	Total
Recording Length (h)	–	–	214.14
Speech (h)	92.71	45.24	137.95
Turns	46,026	25,903	71,929
Word Count	1,155,994	635,862	1,791,856

Table 1: Comparative analysis of classroom discourse: Recording vs. active speech duration, turn frequency, and lexical volume by role.

State	Female	Male	Unk.
Australian Capital Territory	24	60	0
New South Wales	2,071	1,729	1
Queensland	617	617	0
Northern Territory	7	13	0
South Australia	402	424	1
Tasmania	173	116	0
Victoria	676	615	0
Western Australia	117	72	0

Table 2: Distribution of informants by state and recorded sex (archival metadata field).

6. Evaluation

6.1. Evaluation Setup

We randomly selected 10 narratives (approximately 30 minutes of speech; 185 WhisperX-derived turns) for manual evaluation.[3]

Gold discourse-role labels (TEACHER/STUDENT) were assigned at the turn level based on auditory inspection and interactional function by a single annotator. Because the recordings are dyadic, roles induce a consistent two-speaker mapping per file, standardized as SPEAKER_00 (Student) and SPEAKER_01 (Teacher).

We compare (i) a baseline WhisperX (large-v3) + Pyannote diarization system without role correction and (ii) the full pipeline including LLM-based role correction. Performance is measured as turn-level speaker accuracy. For the full pipeline, predicted and gold turns share identical timestamps and are matched by exact key after rounding to three decimals. For the baseline, gold turns are assigned the speaker label with maximum temporal overlap; non-overlapping cases are excluded.

6.2. Results

The LLM-assisted pipeline substantially improves role attribution over acoustic diarization alone, yielding a 33-point absolute gain in turn-level accuracy on this sampled evaluation set.

[3]We use *turn* to refer to WhisperX segments defined by start/end timestamps and an associated text span.

System	Accuracy (%)
Baseline (WhisperX + Pyannote)	62.70
Full Pipeline (LLM-assisted)	**95.68**

Table 3: Turn-level speaker accuracy on 10 sampled narratives. Baseline labels are mapped by maximum temporal overlap.

7. Implications for Dialect Research and Corpus Construction

The evaluation results suggest that the proposed architecture is not only effective for the MD corpus, but also relevant for other interview-based dialect collections.

7.1. LLM-Assisted Role Attribution in Interview Corpora

Many dialect archives consist of interview-style or field recordings with asymmetric speaker roles.[4] In such settings, acoustic diarization alone often collapses interviewer and respondent speech or inconsistently labels short question turns. Our results suggest that incorporating a lightweight reasoning-oriented LLM step can substantially improve role consistency without manual reannotation.

This approach is directly transferable to other large-scale dialect collections, including 20th-century dialectological or sociolinguistic interviews from the US or UK, where interviewer–informant structure is systematic but acoustically variable.

7.2. Temporal Precision for Variationist Analysis

Dialect research frequently depends on accurate segment boundaries for vowel measurement, duration studies, and turn-taking analysis. By combining WhisperX transcription with MFA-based boundary refinement while preserving original tokenization, the pipeline produces word-level timestamps suitable for corpus-phonetic workflows.

The short alignment-window strategy and token reconstruction mechanism are not specific to Australian English and can be applied to other legacy dialect recordings with comparable acoustic quality.

7.3. Scalability and Cross-Linguistic Potential

The pipeline is modular and language-agnostic in design. While this study focuses on Australian English, the same architecture—ASR, diarization, LLM-based role correction, and forced

[4]For example, field recordings from the 20th-century American *Linguistic Atlas Project*.

alignment—can be adapted to other dialect continua and minority language archives, provided appropriate acoustic and pronunciation models are available.

Because many historical dialect archives remain untranscribed, the approach offers a practical pathway for transforming large volumes of interview data into searchable, temporally aligned corpora. For example, MD_NLP supports geographically indexed extraction of spontaneous-speech tokens in targeted phonological environments, enabling spatial modelling of regional patterning in mid-20th-century AusE using the linked coordinates and informant metadata.

8. Using MD_NLP for Dialect Research

MD_NLP is designed to support workflows that require (i) searchable text, (ii) reliable speaker-role separation, (iii) word-level timing suitable for phonetic extraction, and (iv) linkage to demographic and geographic metadata. Because segments are short and role-consistent, researchers can retrieve targeted stretches of spontaneous speech without extensive manual cleaning.

8.1. Example queries and extraction patterns

The dataset supports common dialectological and corpus-phonetic extraction tasks, including: (i) retrieval of lexical or grammatical patterns conditioned by region (e.g., discourse markers, quotatives, negation patterns), (ii) extraction of phonological environments for vowel/consonant measurement (e.g., pre-lateral contexts, pre-nasal contexts, stress-conditioned subsets), and (iii) modelling of interactional structure (e.g., teacher prompt types vs. student narrative responses).

For example, researchers can (a) filter to STUDENT turns only to avoid interviewer speech, (b) restrict to segments from specific states or subsets of locations using geocoded coordinates, and (c) extract aligned word tokens for downstream forced-alignment or formant measurement pipelines. The availability of school location, birthplace, and parental background variables enables analyses that jointly model regional location and social stratification within the same historical dataset.

8.2. Spatially explicit analyses

Because each segment is linked to geocoded recording sites, MD_NLP can be used in spatially explicit dialectology. Researchers can aggregate measurements (lexical rates, phonetic summaries, discourse-role distributions) at the level of locations, regions, or states and apply spatial statistical methods (e.g., clustering, hotspot analysis, spatial autocorrelation) to test whether observed patterns exhibit geographic structure beyond chance. This is particularly useful for evaluating claims of regional differentiation in AusE, where effects may be subtle and geographically diffuse.

8.3. Diachronic linkage to contemporary corpora

Although MD_NLP focuses on historical narratives, its structure is compatible with modern large-scale speech corpora that provide transcripts and audio. This supports direct comparison of regional patterns across time (e.g., 1959/60 vs. present-day corpora), allowing researchers to test whether regional structure is stable, emerging, or reorganizing. In this sense, MD_NLP can serve as an historical anchor point for contemporary regional studies of AusE.

9. Conclusion

We presented MD_NLP, a discourse-annotated, georeferenced corpus of Mitchell–Delbridge narratives constructed using an integrated ASR, diarization, LLM-based role correction, and forced-alignment pipeline. The resulting resource combines temporally precise word-level timestamps with structured sociolinguistic metadata and national geographic coverage.

Evaluation shows that incorporating a lightweight LLM-based reasoning step substantially improves role attribution in interview-style recordings, increasing turn-level accuracy from 62.70% to 95.68%. This improvement is particularly relevant for dialect corpora in which interviewer–informant asymmetry and variable recording conditions challenge acoustic diarization.

Beyond the MD corpus, the proposed architecture is directly applicable to other large-scale dialect archives, including mid-20th-century North American and British sociolinguistic interviews and comparable legacy collections in other languages. Many such archives remain only partially transcribed and lack reliable speaker attribution; the present approach offers a practical pathway toward transforming them into searchable, temporally aligned, and socially annotated research corpora.

For Australian English specifically, MD_NLP provides new opportunities for investigating regional variability, discourse structure, and corpus-phonetic phenomena in spontaneous mid-20th-century speech at a national scale. By combining alignment precision with rich demographic and geographic metadata, the corpus supports emerging research on spatial and socially conditioned

variation in AusE.

Future work will expand quantitative evaluation, explore additional discourse annotations, and further assess cross-dialect and cross-linguistic transferability of the pipeline.

The integration of structured metadata with discourse-aware speaker attribution positions MD_NLP as a foundation for large-scale, spatially explicit dialect research in Australian English.

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.

10. Bibliographical References

Emily P. Ahn, Gina-Anne Levow, Richard A. Wright, and Eleanor Chodroff. 2023. An Outlier Analysis of Vowel Formants from a Corpus Phonetics Pipeline. In *Interspeech 2023*, pages 2573–2577.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *INTERSPEECH 2023*, pages 4489–4493.

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

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.

Steven Coats, Carmelo Alessandro Basile, Cameron Morin, and Robert Fuchs. 2025a. The YouTube corpus of Singapore English podcasts. *English World-Wide*, 46(3):274–298.

Steven Coats, Chloé Diskin-Holdaway, and Debbie Loakes. 2025b. Regional Distribution of the /el/-/æl/ Merger in Australian English. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 147–156, Abu Dhabi, UAE. Association for Computational Linguistics.

Rolando Coto-Solano, James N. Stanford, and Sravana K. Reddy. 2021. Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems With DARLA. *Frontiers in Artificial Intelligence*, 4.

Felicity Cox and Sallyanne Palethorpe. 2019. Vowel variation in a standard context across four major Australian cities. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, pages 577–581. Australasian Speech Science and Technology Association.

Felicity Cox, Sallyanne Palethorpe, and Samantha Bentink. 2014. Phonetic archaeology and 50 years of change to Australian English /i:/. *Australian Journal of Linguistics*, 34(1):50–75.

Felicity Cox, Joshua Penney, and Sallyanne Palethorpe. 2024. Australian English monophthong change across 50 Years: Static versus dynamic measures. *Languages*, 9(3).

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Jason Li, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Nithin Koluguri, Jocelyn Huang, Jia Fei, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Naren Sean, and Boris Ginsburg. 2023. NeMo: a toolkit for Conversational AI and Large Language Models.

Guillaume Klein and Mahmoud Ashraf. 2023. faster-whisper: Faster whisper transcription with ctranslate2.

Nikola Ljubešić, Peter Rupnik, and Tea Perinčić. 2024. Mići Princ – A Little Boy Teaching Speech Technologies the Chakavian Dialect. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 232–250. Institute of Contemporary History.

Deborah Loakes, John Hajek, and Janet Fletcher. 2017. Can you t[æ]ll I'm from M[æ]lbourne?: An overview of the dress and trap vowels before /l/ as a regional accent marker in Australian English. *English World-Wide. A Journal of Varieties of English*, 38(1):29–49.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger.

2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech 2017*, pages 498–502.

Alexander George Mitchell and Arthur Delbridge. 1998. The speech of Australian adolescents: Research data and recordings collected by AG Mitchell and Arthur Delbridge in 1959 and 1960. The University of Sydney.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH 2023*, pages 3222–3226.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28448–28481. PMLR.

Penelope Schmidt, Chloé Diskin-Holdaway, and Debbie Loakes. 2021. New insights into /el/-/æl/ merging in Australian English. *Australian Journal of Linguistics*, 41(1):66–95. _eprint: https://doi.org/10.1080/07268602.2021.1905607.

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.