

Towards an Australian Parliamentary Speech Corpus

Steven Coats

University of Oulu, Finland

steven.coats@oulu.fi

CLARIN Annual Conference, Brighton

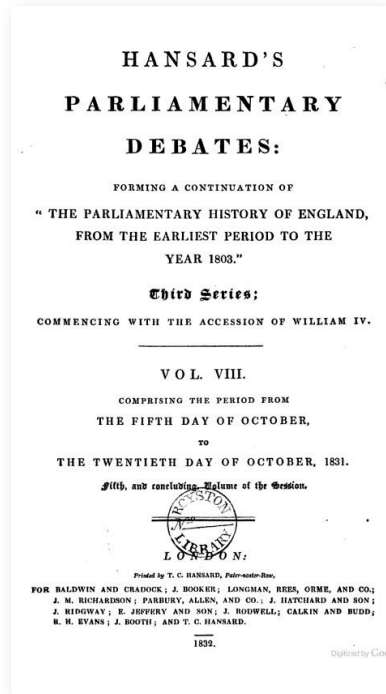
September 30th, 2026



Overview

1. **Motivation:** Spoken reality vs. the Hansard record
2. **Background and Data:** Parliamentary corpora and Australian sources
3. **Proof of Concept:** Speech alignment and metadata enrichment
4. **Preliminary extensions:** Acoustic speaker matching and Hansard-supported name correction
5. **Outlook:** Verbatim speech and AU ParlaSpeech

The official record: Hansard



Hansard (1832)

[Wikimedia Commons / PD](#)

- **Commonwealth tradition:** Named after English publisher Thomas Curson Hansard; standard in the UK, Australia, New Zealand, Canada, Singapore, and elsewhere
- **"Substantially verbatim":** Official transcripts omit repetitions, false starts, and redundancies, while regularizing grammar
- **Linguistic limitation:** Not ideal for some linguistic research questions; can directly alter the **interpretation of discourse** (illustrated in the next examples)

From speech to the official record: Grammar



Llew O'Brien • House of Representatives
17 August 2026 • 12 seconds

ASR

"Indeed, if every international obligation Australia had ever undertaken **was** swept away and only one could remain, for me it would be the one that protects the fundamental human rights of our children."

Hansard

"Indeed, if every international obligation Australia had ever undertaken **were** swept away ..."

- Evidence consistent with the ongoing shift away from subjunctive "were" in Australian English (cf. Filppula, 2022; Vaughan & Mulder, 2014)

From speech to the official record: Managing participation



The Speaker • House of Representatives
17 August 2026 • 18 seconds

ASR

"She's done that. So we're going to... **No, resume your seat. Resume your seat.** We're going to deal with this properly and efficiently and effectively."

Hansard

"She's done that. We're going to deal with this properly, efficiently and effectively."

- Hansard omits the directive, the repetition, and the interruption of the Speaker's own formulation

Why preserve these features?

Linguistic authenticity:

- Parliamentary corpora are records of actual human speech, not only repositories of political decisions
- Spoken grammar (e.g. *was* vs. subjunctive *were*) captures ongoing variation and language change which editorial regularization may erase

Turn interaction and decision-making:

- Interactional details (directives, interruptions, floor management) reveal how political differences and agreements emerge in real time and are essential for understanding how legislative debate actually unfolds

Background: Parliamentary corpora and AU Hansard

CLARIN infrastructure

- **ParlaMint:** European parliamentary transcripts with metadata (Erjavec et al., 2023) and **ParlaSpeech:** Aligned audio for some Slavic languages (Ljubešić et al., 2025)

Australian Hansard research

- ***Australian Diachronic Hansard Corpus:*** Aggregated Hansard transcripts 1940s–2010s, used to study (e.g.) colloquialization and style shifts (Collins & Smith, 2021; Kruger & Smith, 2018)
- Comparison of a small corpus of manually transcribed parliament recordings with the official Hansard records (Kotze et al., 2023)
- **RAPID** (Reusable and Accessible Public Interest Documents) project; Metapragmatics of "unparliamentary language" over a century (Hames et al., 2025)

Goals of AU ParlaSpeech pilot

- **Verbatim speech:** Retain non-standard usages and turn structures/content, also self-repairs, interjections, hesitations
- **Synchronized audio & alignment:** Word-level timing from WhisperX
- **Speaker identities & metadata:** Speaker attribution from Hansard; MP information and electoral demographics
- cf. Asonitis et al. (2026): pairs AU/NZ audio to Hansard for ASR training

Data sources

Parliament of Australia – ParView

(<https://aph.gov.au>):

- House sitting audio, delivered through HLS media streams
- Sitting dates and ParView recording IDs
- Hansard XML files: Speaker names and roles, party and electorate metadata, and speech-turn structure
- Archived WebVTT live captions

OpenAustralia:

- Mirror of Parliament's Hansard XML

Wikidata:

- Supplementary speaker metadata (date of birth, gender, education)

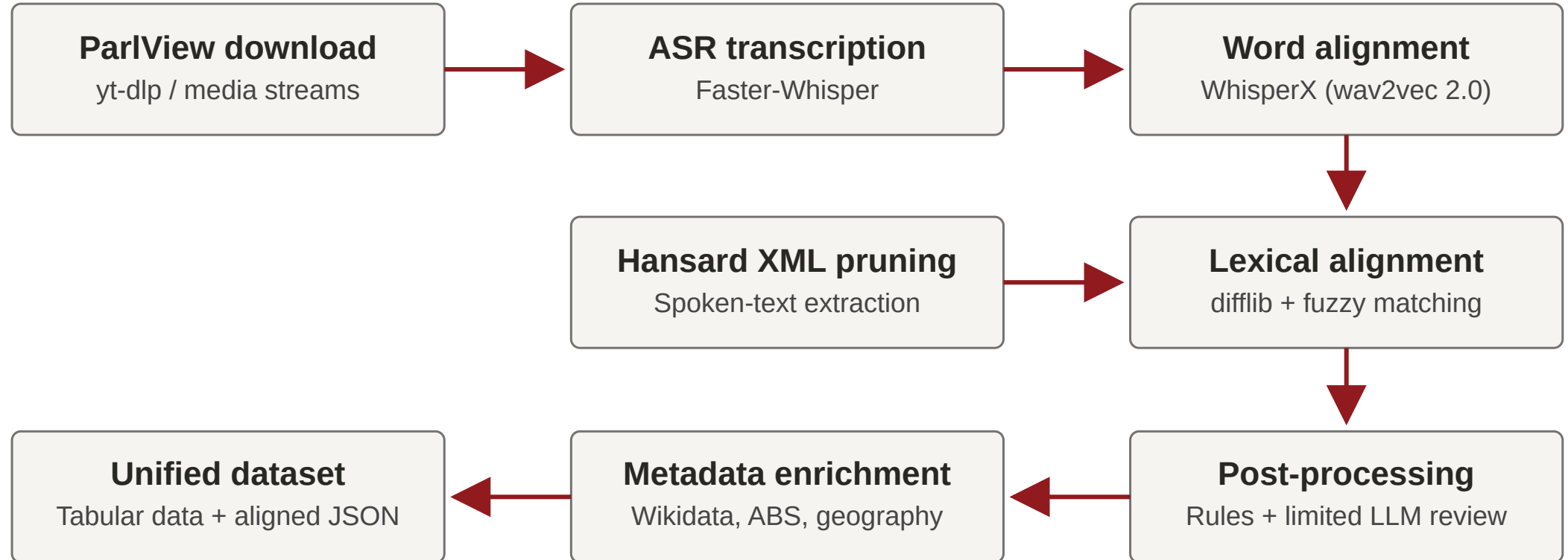
Digital Atlas of Australia:

- Constituency latitude-longitude coordinates

Australian Bureau of Statistics:

- Constituency median weekly household income

The proof-of-concept pipeline



ASR and Hansard transcript matching

1. Download the recording, transcribe with faster-whisper large-v3 (Radford et al., 2023)
2. WhisperX (Bain et al., 2023) supplies word-level timing
3. Convert the ASR and Hansard texts into normalized word streams
4. Use **diffli**b to identify matching stretches
5. Transfer Hansard speaker labels to aligned ASR segments using matched content words
6. Use fuzzy turn matching to assign ASR segments not resolved by the global alignment
7. Refine the assignments with contextual and parliamentary procedural rules

Proof of concept: a final Gemini 2.5 Flash pass reviewed assignments in local Hansard context and could split mixed-speaker segments.

Metadata enrichment

Linked information

- **Hansard:** speaker, party, electorate, parliamentary role
- **Wikidata API:** date of birth, gender, institutions attended
- **Digital Atlas of Australia:** constituency boundaries and calculated centroids
- **Wikidata SPARQL** → **ABS:** electorate identifiers and census income

Possible extensions: age at recording, birthplace, qualifications, previous occupation, parliamentary experience; richer electorate demographics.

Preliminary extensions

- Better matching: combine lexical and acoustic evidence
- Hansard-supported onomastic correction of plausible name errors

Acoustic speaker matching: method and preliminary comparison

- Raw faster-whisper segments sometimes overlap speakers

Method

1. Use **pyannote** to group speech into acoustic speaker clusters
2. Use ASR–Hansard matches to attach names to parts of each cluster
3. Name a cluster when one speaker consistently dominates the matched evidence
4. Extend that name to other speech in the cluster

- Adding heuristics to the acoustic matching method (speech of the Clerk, some Speaker utterances) should improve this

Results on three additional sittings: Percent of ASR segments assigned to a speaker

Sitting	Hansard text only	Acoustic + Hansard	Agreement where both named
17 Aug.	99.1%	94.6%	94.9%
18 Aug.	98.6%	94.2%	95.5%
20 Aug.	97.8%	92.7%	95.3%

Hansard-supported name correction

- Extract names, places, and institutions from Hansard (spaCy *en_core_web_trf*)
- Locate the corresponding ASR span using matching words on both sides
- Require a fuzzy orthographical match
- Save a separate corrected layer and a record of each substitution

First name-correction test: 17 August

Original ASR	Hansard-supported replacement
Greg Kavanagh	Greg Cavanagh
Hillsville	Healesville
Unadera	Unanderra
Sam Ray	Sam Rae
CSRIO's	CSIRO's
TEXA	TEQSA

49 substitutions, 45 distinct pairs, across approximately **58,000 ASR words**.

Verbatim speech: Keeping disfluencies

Disfluencies as cognitive and interactional signals:

- Filled pauses (*um*, *uh*) and self-repairs can
 - Signal planning delay and upcoming complexity (Clark & Fox Tree, 2002; Arnold et al., 2003; Levelt, 1983; Bortfeld et al., 2001)
 - Mark disagreement or resistance (Kendrick & Torreira, 2015)
 - Be analyzed as sociolinguistic variables subject to regional variation and diachronic change (Fruehwald, 2016; Tottie, 2011; Wieling et al., 2016)

Verbatim ASR possibilities:

- ParlaSpeech-Pause: use Wav2Vec2-BERT to add filled pauses
- Verbatim models such as *CrisperWhisper2* (Wagner et al., 2026)
- Fine-tuned Whisper on Conversation Analysis (CA) data to "keep the ums" (Coats, forthcoming)

Towards an Australian ParlaSpeech: what remains to be done

1. Evaluation (primary priority)

- **Verbatim gold standard:** Human-verified reference transcripts to quantify actual ASR Word Error Rate (WER) and disfluency capture
- **Attribution audit:** Ground-truth evaluation of acoustic clustering, diarization boundaries, and speaker assignment precision

2. Corpus scaling

- Scale beyond the 1-sitting proof of concept to full multi-year legislative coverage

3. ParlaSpeech layers & verbatim ASR

- **ParlaSpeech enrichments:** UD syntax, sentiment, and topic classification
- **Interoperable formats:** Parla-CLARIN TEI-XML, Praat TextGrids, BlackLab/NoSketchEngine concordancers

Thank you!

Proof-of-concept dataset:

<https://huggingface.co/datasets/stcoats/au-parliament-poc-turns-2026-03-30>

References

- Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). *Disfluencies signal theee, um, new information*. *Journal of Psycholinguistic Research*, 32(1), 25–36.
- Asonitis, A., Lanzendörfer, L. A., Berdoz, F., & Wattenhofer, R. (2026). *WorldSpeech: A multilingual speech corpus from around the world*. *arXiv preprint arXiv:2605.09167*.
- Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). *WhisperX: Time-accurate speech transcription of long-form audio*. *Proceedings of Interspeech 2023*, 4489–4493.
- Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). *Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender*. *Language and Speech*, 44(2), 123–147.
- Clark, H. H., & Fox Tree, J. E. (2002). *Using uh and um in spontaneous speaking*. *Cognition*, 84(1), 73–111.
- Coats, S. (forthcoming). Keeping the ums: Conversation-analysis-informed verbatim ASR for web speech. *Proceedings of the 13th Web as Corpus Workshop (WAC-XIII)*.
- Collins, P., & Smith, A. (2021). *Diachronic register change in Australian English parliamentary debates*. In P. Collins & A. Smith (Eds.), *Varieties of English in the Indo-Pacific* (pp. 147–172). Springer.
- Erjavec, T., et al. (2023). *The ParlaMint corpora of parliamentary proceedings*. *Language Resources and Evaluation*, 57, 415–448.
- Filppula, M. (2022). *The variable fortunes of the were-subjunctive in varieties of English*. In S. Lucek & C. P. Amador-Moreno (Eds.), *Expanding the landscapes of Irish English research* (pp. 54–64). Routledge.
- Fruehwald, J. (2016). *Filled pause choice as a sociolinguistic variable*. *University of Pennsylvania Working Papers in Linguistics*, 22(2), 41–49.
- Hames, S., Haugh, M., & Musgrave, S. (2025). “How is that unparliamentary?”: The metapragmatics of ‘unparliamentary’ language in the Australian Federal Parliament. *Lingua*, 320, 103932.
- Kendrick, K. H., & Torreira, F. (2015). *The timing and construction of preference: A quantitative study*. *Discourse Processes*, 52(4), 255–289.
- Kotze, H., Korhonen, M., Smith, A., & van Rooy, B. (2023). *Salient differences between Australian oral parliamentary discourse and its official written records: A comparison of 'close' and 'distant' analysis methods*. In M. Korhonen, H. Kotze, & J. Tyrkkö (Eds.), *Exploring language and society with big data: Parliamentary discourse across time and space* (pp. 54–88). John Benjamins.
- Kruger, H., & Smith, A. (2018). *Colloquialization versus densification in Australian English: A multidimensional analysis of the Australian Diachronic Hansard Corpus (ADHC)*. *Australian Journal of Linguistics*, 38(3), 293–328.
- Levelt, W. J. M. (1983). *Monitoring and self-repair in speech*. *Cognition*, 14(1), 41–104.
- Ljubešić, N., Rupnik, P., & Koržinek, D. (2025). *The ParlaSpeech collection of automatically generated speech and text datasets from parliamentary proceedings*. In *Speech and Computer* (pp. 137–150). Springer.
- OpenAustralia Foundation. (n.d.). *OpenAustralia: Australian parliamentary Hansard and XML archive*.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision*. *Proceedings of ICML 2023*, 28492–28518.
- Tottie, G. (2011). *Uh and um as sociolinguistic markers in British English*. *International Journal of Corpus Linguistics*, 16(2), 173–197.
- Vaughan, J., & Mulder, J. (2014). *The survival of the subjunctive in Australian English: Ossification, indexicality and stance*. *Australian Journal of Linguistics*, 34(4), 486–505.
- Wagner, L., Zusag, M., & Thallinger, B. (2026). *Transcription policy as a latent variable: Activating controllable verbatim ASR with word-level timing*. *arXiv preprint arXiv:2607.18934*.
- Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). *Variation and change in the use of hesitation markers in Germanic languages*. *Language Dynamics and Change*, 6(2), 199–234.