CoANZSE Audio: Creation of an Online Corpus for Linguistic and Phonetic Analysis of Australian and New Zealand Englishes

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

Abstract

CoANZSE Audio is a searchable online version of the *Corpus of Australian and New Zealand Spoken English*, a 195-million-word collection of geo-located YouTube transcripts of local government channels. In addition to the part-of-speech-tagged and lemmatized transcript data, CoANZSE Audio provides access to almost all of the underlying audio, as well as to forced alignments of the audio with transcript content, in Praat's TextGrid format. This paper describes the methods used to create the corpus from open-source tools and the architecture of the CoANZSE Audio website. Two possible linguistic analyses based on CoANZSE Audio data are described: use of double modals, a rare syntactic feature, and raising of the mid front vowel /ɛ/ in New Zealand English. CoANZSE Audio can be considered to be among the first large, free, fully searchable online corpora containing data suitable for acoustic phonetic analyses in addition to lexical, grammatical, and discourse properties of Australian and New Zealand Englishes.

Keywords: Corpus Phonetics, YouTube, Australian English, New Zealand English, BlackLab

1. Introduction and Background

In recent years increasing availability of naturalistic language data has facilitated research into varieties of English. Large corpora and text archives have been created from web and social media data, (for example Davies and Fuchs, 2015; Dunn, 2019), enabling detailed studies of regional distributions of English lexis, grammar, and syntax. Although some resources include audio transcripts, relatively few corpora of English contain audio recordings as well as transcribed content (for an overview see Liberman, 2019). Audio corpora that are available tend to be relatively small in scale, in the range of tens to hundreds of hours of recorded audio, and are mostly not searchable or accessible online. Many language resources suitable for phonetic analysis are only accessible via a paid license.

At the same time, recent technological and societal developments have resulted in increased access to potentially valuable sources of linguistic data such as naturalistic audio recordings and transcripts, facilitating the collection of large corpora of multimedia data for linguistic research purposes. In addition, legislative frameworks and institutional actors have begun to emphasize the importance of FAIR principles (findable, accessible, interoperable, and reusable) for data collection and management (Wilkinson et al., 2016); the European Union and other jurisdictions increasingly allow reuse of copyrighted data in the context of research and education, especially for data generated by public entities such as governmental bodies.[1] In Australia and New Zealand as well, state and local governments have begun to emphasize the fundamental openness and reusability of data.[2]

This article presents CoANZSE Audio (Coats, 2023c), a large, searchable online collection of more than 20,000 hours of geolocated naturalistic transcripts and audio from Australia and New Zealand. The resource builds upon the existing CoANZSE corpus (Coats, 2022a, 2022b) by providing access to audio data and forced alignment files in addition to transcript content, via BlackLab, a powerful Apache Lucene-based search engine developed at the Dutch Language Institute.[3] Access to CoANZSE Audio is free of charge. The resource thus represents one of the few large combined corpora of transcripts, audio, and alignments of English that is fully open for research and educational purposes.

The rest of the article is organized as follows: Section 2 describes the current situation with regards to availability of corpora of combined transcript and audio data and briefly notes the principal resources for the study of Australian and New Zealand Englishes. Section 3 describes the methods used to collect audio data, force align the audio with the transcripts, and prepare the material for the CoANZSE Audio website. Section 4

[1] See, e.g., Kupietz et al. (2018) for use of language data in German-language corpora.

[2] See, e.g. https://data.nsw.gov.au/nsw-government-open-data-policy

[3] http://inl.github.io/BlackLab

LREC-COLING 2024, pages 3407–3412

presents two potential analyses using data from the resource, and Section 5 notes several caveats and provides a summary and an outlook for future developments for the website.

2. Comparable Resources

Most searchable online corpora of English are text-only: the extensive and widely-used corpora at english-corpora.org, for example, comprise textual content scraped from websites as well as digitized versions of physical texts, but do not provide access to audio data (e.g., COCA; Davies, 2008–). Language resources that include speech transcripts, audio data, and, in some cases, alignments, are maintained by organizations such as the Linguistic Data Consortium,[4] the European Language Resources Association,[5] or the Bavarian Archive for Speech Signals.[6] These organizations mostly make language resources available via paid licenses, an approach which may limit accessibility for academic researchers.[7] A move towards providing open online access to transcripts, audio, and alignments for linguistic research purposes is evident (e.g., from the United States, the Corpus of Regional African American Language, with 160 hours of transcribed speech; Kendall and Farrington, 2023), but as of 2023, few searchable, open-access resources are available that contain transcript, audio, and alignment data.

For Australia, the Australian National Database of Spoken Language (Millar et al., 1994) was collected during the 1990s in order to create a resource of English spoken in Australia. The materials, held at Macquarie University in Sydney, are not available online. The AusTalk project (Estival et al., 2014; Cassidy et al., 2017), created from 2011–2016, comprises recordings of reading tasks and spontaneous speech of 1,000 persons from 15 Australian locations. As of late 2023, some of the material is available online,[8] and the project is slated to be transferred to the Linguistic Data Commons of Australia (LDaCA), a centralized resource for linguistic research data. LDaCA's search interface provides document-level access to some resources but not to, as of early 2024, AusTalk.[9] Several other projects have collected Australian speech recordings in the context of sociolinguistic research.[10] Notable is the Sydney Speaks project, comprising over 130 hours of recordings from the 1970s/80s and the 2010s/20s (see e.g., Grama et al., 2021), for which, however, the data is not yet publicly available.

For New Zealand, the most important corpus resource containing audio recordings is the Origins of New Zealand English project (ONZE; Gordon et al., 2007), which comprises recordings made in the middle of the 20th century by New Zealand's National Broadcasting Service and since the 1990s by academic researchers. An audiovisual corpus of sociolinguistic interviews and monologues pertaining to the Christchurch earthquakes of 2010 and 2011, the UC QuakeBox Corpus, has also been utilized to study the phonetics of New Zealand English (Clark et al., 2016; Walsh et al., 2013), and some of the underlying audio and video data is available online.[11] ONZE is not publicly available online, as of 2024.

Analyses of Australian and New Zealand English speech have been undertaken using AusTalk and ONZE materials. For example, Leung et al. (2022) considered fundamental frequency in terms of age, sex, and geographical location for 379 speakers using excerpts of speech from the AusTalk project. Speaker sex was a significant predictor of F0, but age and geographical location were not. Cox et al. (2019) analyzed vowel formants in monophthongs in AusTalk reading list recordings of speakers from Perth, Sydney, Melbourne, and Adelaide, finding some incipient regional variation. A shift in the quality of front vowel monophthongs in New Zealand English has been noted in previous research (Watson et al., 1998, 2000; Gordon et al., 2004). Recent studies of New Zealand vowels using ONZE and UC QuakeBox Corpus data have shed light on the dynamics of these changes (Hay et al., 2015; Brand et al., 2021; Black et al., 2023).

Overall, although a number of resources containing audio recordings and transcripts exist for research into the linguistic properties of English in Australia and New Zealand, including in terms of their acoustic properties, the underlying data has not yet, for the most part, been made freely available online. The CoANZSE Audio website may therefore offer a useful complement to existing resources for the study of these varieties, as well as exemplifying the current turn in linguistics towards collection and reuse of publicly available data.

3. CoANZSE Data Collection

3.1. CoANZSE Transcripts and Audio

Data was collected for CoANZSE from the YouTube channels of local governments in Aus-

[4] https://www.ldc.upenn.edu

[5] https://http://www.elra.info

[6] https://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html

[7] An academic license to an ELRA resource containing 199 hours of Australian English speech, for example, costs more than €43,000.

[8] https://app.alveo.edu.au/catalog/austalk

[9] https://www.ldaca.edu.au

[10] See, e.g., https://rsha.cass.anu.edu.au/events/workshop-language-corpora-australia.

[11] https://quakestudies.canterbury.ac.nz/

State/territory	FLAC files	TextGrid files
ACT	43,752	41,253
NSW	1,299,949	1,299,949
NT	6,628	6,471
QLD	950,084	837,297
SA	643,866	629,524
TAS	240,453	233,910
VIC	1,624,830	1,595,737
WA	386,898	377,016
Total	5,196,460	4,993,881

Table 1: Number of audio and alignment files by Australian state/territory

tralia and New Zealand. Channels were identified on the basis of lists provided by national, state, and territory governments (for details see Coats, 2022b and the references therein); the open-source Python library yt-dlp was used to retrieve automatic speech recognition (ASR) transcripts.[12] Transcripts were annotated with part-of-speech tags using SpaCy's en_core_web_sm model.[13] Audio data was retrieved for the videos indexed in the corpus by cutting individual ASR transcripts into 20-word chunks and retrieving the audio from YouTube's servers, using the timing tags of the first and last word in the corresponding chunk. Audio was retrieved in the format with the highest available quality, generally 130kbit/s and .m4a; audio files were converted to FLAC using ffmpeg. Audio and text chunks were aligned with the Montreal Forced Aligner (McAuliffe et al., 2017), version 2.0,[14] using the default English acoustic model (english_us_arpa v2.0.0a), trained on a subset of the librispeech dataset (Panayotov et al., 2015). Outputs, as TextGrid files, were rendered using ARPAbet symbols on the basis of the CMU Pronunciation Dictionary (Weide et al., 1998).

Not all transcripts in CoANZSE have audio and forced alignment data. Some videos were removed from YouTube or made private in the time between collection of the transcripts (mid-2022) and the audio (mid-2023). For other transcript files, the forced alignment pipeline did not result in a usable TextGrid file, likely due to errors in the transcripts and/or poor audio quality. Table 1 shows the proportion of transcripts for which audio and forced alignments are available, as of October 2023.[15]

[12] https://github.com/yt-dlp

[13] https://spacy.io/usage/models

[14] https://montreal-forced-aligner.readthedocs.io/en/latest/index.html

[15] As of October 2023, the alignment of the New Zealand data is still ongoing. The process is expected to be completed by November 2023.

3.2. Coanzse.org Website

The core functionality of the CoANZSE Audio website is a customized version of BlackLab (de Does et al., 2017), an open-source search platform based on Apache Lucene and developed at the Dutch Language Institute. The website, audio, and TextGrid files are hosted on servers at Finland's Centre for Scientific Computing.[16]

Basic search functionality for CoANZSE Audio allows the user to search for words, lemmas, and part-of-speech tags as well as the metadata fields country, state or territory, council name, channel name, channel url, document ID (i.e., YouTube's 11-character unique identifier), location, date of upload, and transcript length. Audio and TextGrid files can be downloaded for individual search hits or for all the hits displayed on a search results page. Concordance lines for all results of a search can be downloaded as a CSV file. Figure 1 shows the extended search interface.

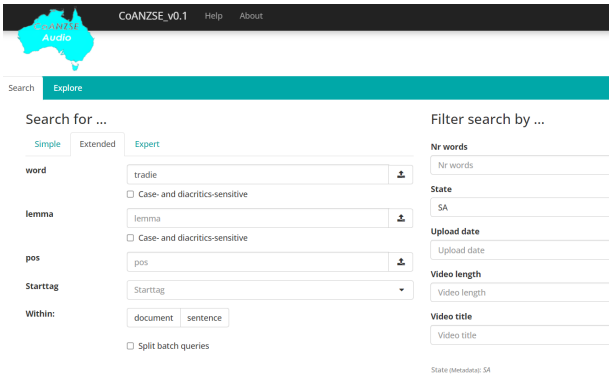

Figure 1: CoANZSE Audio website extended search interface

Corpus Query Language can be used for detailed searches of multi-element combinations of word tokens, lemmas, part-of-speech tags, and wildcards. Results can be filtered based on channel location – for example, for a given utterance, a user can easily access all instances from New South Wales or from greater Perth by selecting the corresponding states, channels, or councils. For each result, the corresponding audio clip can be listened to in the browser or downloaded locally, as can the TextGrid file (Figure 2).

The display options allow the user to find lexical and grammatical items for which there may different regional usage frequencies. For example, in Australia, the term "eastern states" is used in Western Australia, and to a lesser extent in South Australia, to refer to Victoria, New South Wales, and Queensland. Figure 3 captures the intermediate position of South Australia, both geographically and in the relative frequency of this item.

[16] https://csc.fi

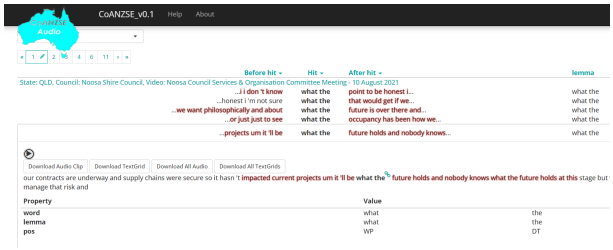

Figure 2: CoANZSE Audio results page

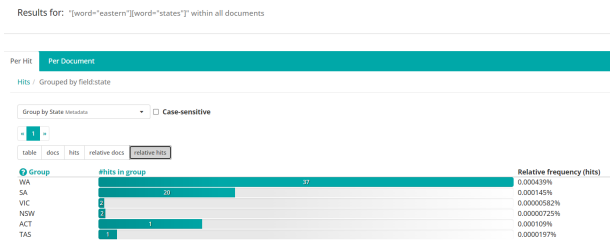

Figure 3: Relative frequencies of ”eastern states”

In total, CoANZSE Audio provides access to more than 195 million words of searchable, part-of-speech-tagged transcript content, nearly 10 million audio files, and a similar number of forced alignment files: in total, over 6 TB of text and audio data.

4. Potential Use Cases

CoANZSE Audio can potentially be used to explore a wide range of research questions pertaining to linguistic and sociocultural language variation. As it contains audio and TextGrid files for hundreds of millions of naturalistic, ecologically valid individual phones, one potential application for the data is to build upon the strong foundation of existing work in phonetics in Australia and New Zealand in order to investigate regional variation in acoustic properties of speech. In addition, CoANZSE Audio data enables analyses of grammatical, lexical, and cultural phenomena. A rare grammatical phenomenon, the use of double modals, and a phonetic feature, raising of a monophthong, exemplify use cases for the resource.

4.1. Double modals

Double modals (e.g., *we would might consider it*) are a rare, non-standard feature of spoken language syntax known from Northern Britain and Ireland and the Southern United States, but until recently, had not been investigated in Australian or New Zealand Englishes (Coats, 2023b). The feature can be easily examined and verified in CoANZSE Audio via input of specific modal combinations or the part-of-speech tag sequence MD MD in the extended search field. Cameron and Coats (2023) found the feature to be a consistent, although quite rare feature of naturalistic speech in both Australia and New Zealand, findings that may potentially shed light on the nature of syntactic variation in English as well as the process of new dialect formation (Trudgill, 2004).

4.2. New Zealand monophthong raising

Formant values of monophthongs have been shown to have shifted in New Zealand English in the latter half of the 20th century (Watson et al., 1998, 2000; Hay et al., 2015; Brand et al., 2021; Black et al., 2023): the DRESS vowel, for example, has raised and the KIT vowel has lowered in recent New Zealand English, whereas the DRESS vowel for some speakers in Victoria, Australia, has moved in the opposite direction. Data from CoANZSE Audio can be used to assess the contemporary status of this and other vowels.

5. Caveats, Summary, and Future Developments

CoANZSE Audio provides access to hundreds of millions of words force-aligned transcript and audio data; it represents not only a potentially useful resource for the investigation of linguistic and discourse properties of English in Australia and New Zealand, but also an example of how naturalistic speech harvested from online sources can organized and made available for linguistic research purposes.

Despite this, CoANZSE Audio data are not suitable for the investigation of every research question. Because the corpus transcripts and audio files are undiarized (i.e., there is no indication of changes in speaker turn) and there is no metadata pertaining to demographic or social traits of individual speakers, raw CoANZSE Audio data is better suited to to the analysis of large-scale aggregate variation rather than variation that correlates with individual speaker traits. Nevertheless, analyses using traditional sociolinguistic variables such as age, educational level, sex/gender, or ethnic identity may be possible after manual annotation of these categories: because search hits (and file downloads) contain YouTube video identifiers, source videos can be checked by an analyst and the coresponding categories added to the data. CoANZSE transcripts are from ASR transcripts and contain errors due to poor audio quality, out-of-vocabulary items, or non-standard pronunciations. The part-of-speech tag annotations in the resource also contain inaccuracies. For aggregate analyses of frequent phenomena, these errors are unlikely to confound potential results, but for analyses of infrequent phenomena, the data need to be

manually inspected before they are interpreted.[17] Finally, the forced alignments generated by the Montreal Forced Aligner and made available as search results may contain errors due to a variety of issues, including audio quality, background noise or music, transcript errors, voice quality, or other factors.

Australian and New Zealand Englishes are varieties that may be in the process of undergoing regional differentiation (Cox and Palethorpe, 2001, 2019) in terms of phonetic values. CoANZSE Audio opens new possibilities for the study of grammatical and phonetic variation in Australian and New Zealand Englishes. It represents an important step towards freely accessible large-scale, fully searchable, geolocated audio corpora, in partial fulfilment Liberman's 2019 remark that audio recordings will "undoubtedly become more generally available over the next dozen years, facilitating detailed corpus-based studies of allophonic and prosodic variation" (2019: 102). CoANZSE Audio provides researchers with access to a large trove of naturalistic, "ecologically valid" phonetic material with which the contemporary state of Australian and New Zealand Englishes can be analyzed. Whether these varieties are currently undergoing regional differentiation is one of the major research questions that this resource may be able to help address.

6. Bibliographical References

Joshua Wilson Black, Jennifer Hay, Lynn Clark, and James Brand. 2023. The overlooked effect of amplitude on within-speaker vowel variation. *Linguistics Vanguard*, 9(1):173–189.

James Brand, Jen Hay, Lynn Clark, Kevin Watson, and Márton Sóskuthy. 2021. Systematic co-variation of monophthongs across speakers of New Zealand English. *Journal of Phonetics*, 88.

Steve Cassidy, Dominique Estival, and Felicity Cox. 2017. Case study: the AusTalk corpus. *Handbook of Linguistic Annotation*, pages 1287–1301.

Lynn Clark, Helen MacGougan, Jennifer Hay, and Liam Walsh. 2016. "Kia ora. This is my earthquake story". Multiple applications of a sociolinguistic corpus. *Ampersand*, 3:13–20.

Steven Coats. 2022b. The Corpus of Australian and New Zealand Spoken English: A new resource of naturalistic speech transcripts. In *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, pages 1–5.

Steven Coats. 2023a. Dialect corpora from YouTube. In *Language and linguistics in a complex world*, pages 79–102. De Gruyter.

Steven Coats. 2023b. Double modals in contemporary British and Irish speech. *English Language & Linguistics*, page 1–26.

Felicity Cox and Sallyanne Palethorpe. 2001. Vowel change: synchronic and diachronic evidence. In *English in Australia*, pages 17–44. John Benjamins.

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*, volume 450.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus (glowbe). *English World-Wide*, 36(1):1–28.

Jess de Does, Jan Niestadt, and Katrien Depuydt. 2017. Creating research environments with BlackLab. *CLARIN in the Low Countries*, pages 245–257.

Jonathan Dunn. 2019. Global syntactic variation in seven languages: Toward a computational dialectology. *Frontiers in Artificial Intelligence*, 2:15.

Dominique Estival, Steve Cassidy, Felicity Cox, and Denis Burnham. 2014. AusTalk: an audio-visual corpus of Australian English.

Elizabeth Gordon, Lyle Campbell, Jennifer Hay, Margaret Maclagan, Andrea Sudbury, and Peter Trudgill. 2004. *New Zealand English: Its origins and evolution*. Cambridge University Press.

Elizabeth Gordon, Margaret Maclagan, and Jennifer Hay. 2007. The ONZE corpus. In *Creating and digitizing language corpora, volume 2: Diachronic Databases*, pages 82–104. Palgrave Macmillan.

James Grama, Catherine E Travis, and Simon Gonzalez. 2021. Ethnic variation in real time: Change in Australian English diphthongs. In *Language Variation–European Perspectives VIII*, pages 291–314. John Benjamins.

Jennifer B. Hay, Janet B. Pierrehumbert, Abby J. Walker, and Patrick LaShell. 2015. Tracking word frequency effects through 130 years of sound change. *Cognition*, 139:83–91.

[17] See the discussion in Coats (2023a).

Marc Kupietz, Harald Lüngen, Paweł Kamocki, and Andreas Witt. 2018. The German reference corpus DeReKo: New developments–new opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yeptain Leung, Jennifer Oates, Viktória Papp, and Siew-Pang Chan. 2022. Speaking fundamental frequencies of adult speakers of Australian English and effects of sex, age, and geographical location. *Journal of Voice*, 36(3):434–e1.

Mark Y Liberman. 2019. Corpus phonetics. *Annual Review of Linguistics*, 5:91–107.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

J Bruce Millar, Julia P Vonwiller, Jonathan M Harrington, and Phillip J Dermody. 1994. The Australian National Database of Spoken Language. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–97–I–100. IEEE.

Cameron Morin and Steven Coats. 2023. Double modals in Australian and New Zealand English. *World Englishes*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Peter Trudgill. 2004. *New-dialect formation: The inevitability of colonial Englishes*. Oxford University Press, USA.

Liam Walsh, Jen Hay, Derek Bent, Jeanette King, Paul Millar, Viktoria Papp, and Kevin Watson. 2013. The UC QuakeBox Project: Creation of a community-focused research archive.

Catherine I Watson, Jonathan Harrington, and Zoe Evans. 1998. An acoustic comparison between New Zealand and Australian English vowels. *Australian Journal of Linguistics*, 18(2):185–207.

Catherine I Watson, Margaret Maclagan, and Jonathan Harrington. 2000. Acoustic evidence for vowel change in New Zealand English. *Language Variation and Change*, 12(1):51–68.

Mark D. Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santo, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

7. Language Resource References

Steven Coats. 2022a. *The Corpus of Australian and New Zealand Spoken English*.

Steven Coats. 2023c. *CoANZSE Audio*. v0.2.

Mark Davies. 2008–. *Corpus of Contemporary American English*.

Tyler Kendall and Charlie Farrington. 2023. *The Corpus of Regional African American Language*. Version 2023.06.

Weide, Robert and others. 1998. *The Carnegie Mellon Pronouncing Dictionary*. Release 0.6.