The Corpus of Australian and New Zealand Spoken English: A new resource of naturalistic speech transcripts

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

Abstract

The Corpus of Australian and New Zealand Spoken English (CoANZSE) is a 190-million-word corpus of Automatic Speech Recognition (ASR) transcripts from YouTube channels of local councils and other governmental bodies in 472 locations in Australia and New Zealand. CoANZSE can be used to examine grammar and syntax in Australian and New Zealand spoken English, and because tokens are word-timed and transcripts are linked to videos, it can serve as the starting point for phonetic or multi-modal studies. Two exploratory analyses demonstrate differences between Australia and New Zealand in the relative frequencies of double modals, a rare non-standard syntactic feature, and show that transcripts from Australia and New Zealand can be distinguished on the basis of common lexical items.

1 Introduction and Background

The study of regional grammatical variation in English has been stimulated by new methodological approaches (e.g., Nerbonne, 2009; Szmrecsanyi, 2011) and new sources of data in recent years, with corpus-based statistical analyses coming to the forefront, often utilizing textual data from the Web and social media platforms (e.g., Grieve et al., 2019; Hovy and Purschke, 2018; Dunn, 2019). These studies have provided new insights into the structure and distribution of varieties of English, but corpus-based empirical studies of regional patterns of grammatical variation in contemporary English speech remain few. Corpora of transcribed speech may be focused on specific locations, or may not exhibit sufficient geographic granularity for reliable inferences about regional patterns. Some speech corpora are unsuitable for analyses of contemporary language phenomena as they contain mostly transcripts of speech from older speakers recorded in the middle of the 20th century. Most corpora of transcribed speech are not large enough to capture rare features in grammar and syntax (e.g., Corrigan et al., 2012; Greenbaum, 1998; Du Bois et al., 2000-2005; Anderwald and Wagner, 2007).

The widespread use of Automatic Speech Recognition (ASR) by conferencing and video streaming or sharing sites has made it possible to create large corpora of geo-located naturalistic speech, opening up new possibilities for in-depth studies of variation in English. This paper introduces the Corpus of Australian and New Zealand Spoken English (CoANZSE),[1] a 190-million-word corpus of 56,815 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts, corresponding to more than 24,000 hours of video, from 482 YouTube channels of local councils or other institutions of local governance in 472 locations in Australia and New Zealand. In the following, some existing Australian and New Zealand speech corpora are introduced, then the methods used to create CoANZSE are briefly described. Two example exploratory analyses are provided: the syntactic features of double modals is identified in the transcripts, and a classifier is used to distinguish Australian from New Zealand transcripts. ASR transcripts contain errors, so methods of analysis must be robust for use with "noisy data". The summary notes a few possibilities for future work with CoANZSE and similar data.

For Australia and New Zealand, several corpora of transcribed speech exist. The Australian National Corpus (Cassidy et al., 2012) includes speech transcripts from the Australian component of the International Corpus of English (Greenbaum, 1996), the Monash Corpus of Spoken Australian English (Bradshaw et al., 2010), and the Griffith Corpus of Spoken Australian English (Haugh and Chang, 2013). The geographical coverage of these corpora, however, is inconsistent: the Monash Corpus consists mainly of transcripts of Melbourne speakers,

[1] https://cc.oulu.fi/~scoats/CoANZSE.html

and the Griffith Corpus of Brisbane speakers. In terms of corpus size, existing Australian corpora of speech transcripts are mostly not large enough for research into patterns of syntactic variation.

Several corpora of English speech transcripts have been created from the speech of New Zealanders. The spoken component of the New Zealand International Corpus of English (ICE-NZ) comprises approximately 600,000 words, mainly recorded in the 1990s. The Wellington Corpus of Spoken New Zealand English (Holmes et al., 1998), approximately 1 million words in size, contains transcripts of formal and informal speech, also collected mostly in the 1990s. The Origins of New Zealand English Corpus (Gordon et al., 2007) comprises transcripts of recordings of older New Zealand speakers made by New Zealand Radio in the middle of the 20th century, in addition to recordings made by researchers in the 1990s and 2000s. Transcript corpora from Australia and New Zealand have been used for a wide range of studies, but regional variation in grammar and syntax has not been a consistent focus of research attention, due both to geographical sampling and corpus size considerations.

2 Data and methods

Lists of councils, shires, and other administrative units were obtained from state, territorial and national government websites in Australia and New Zealand: 157 from New South Wales, 78 from Victoria, 69 from South Australia, 178 from Western Australia, 21 from Northern Territory, 77 from Queensland, 29 from Tasmania, and 9 from the Australian Capital Territory. A list of 78 councils was retrieved for New Zealand.

Of these 696 local government entities, 578 had web pages, which were were then scraped for links to YouTube channels. The procedure returned 515 YouTube channels, of which 482 contained video content. Channels were manually checked to ensure they corresponded to the linked municipality. Latitude-longitude coordinates were retrieved by inputting the street address listed on the corresponding web page to a geo-coding script. Locations of the sampled channels are shown in Figure 1.

All available ASR transcripts were retrieved from the targeted channels with a Python script, using functions in the yt-dlp[2] library. A custom script parsed transcripts and appended word-timing information; part-of-speech tagging was undertaken using SpaCy's en_core_web_sm model.[3] Table 1 shows corpus size by state/territory in terms of channel, transcript, and word count as well as the corresponding aggregate video length.

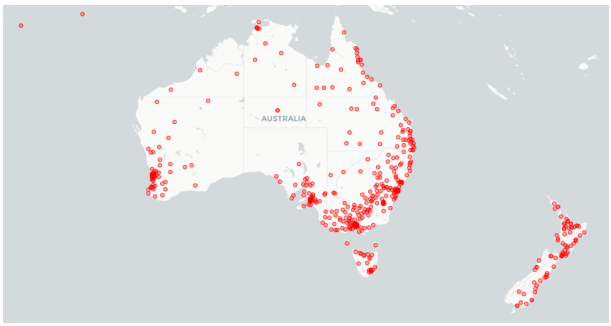

Figure 1: CoANZSE channel locations

Many of the transcripts in CoANZSE record meetings, but the transcripts in the corpus are from many other video types as well, such as interviews, informational and public service videos, vlogs, public readings, and other content types.

3 Exploratory analyses

CoANZSE may be useful for many kinds of linguistic analysis, including regional analyses of grammar and syntax and discourse studies of the content of (for example) public meetings. Because the underlying video and audio data are available, scripting pipelines can be set up that extract targeted content for acoustic or multi-modal analysis. Two preliminary, exploratory analyses are noted below.

3.1 Double modals

The syntactic feature of double modals (e.g., *I might could help you with that*; cf. standard English *I could help you with that* or *I might help you with that*), traditionally held to be restricted to speech in the Southern US and the Northern British Isles, is attested as absent for Australian English (Kortmann and Lunkenheimer, 2013). Recent work using naturalistic data, however, shows that the feature has a broader geographical extent than previously thought (Coats, 2022, In review). A preliminary search for double modals in CoANZSE resulted in 3,119 hits; the first approximately 400 of these were examined in their original videos in order to remove false positives. This exploratory query showed a large number of Australian double modals to be authentic naturalistic usages (Fig. 2).

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

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

Table 1: Corpus Size by Country Location

Location	Channels	Videos	Words	Length (h)
Australian Capital Territory	8	650	915,542	111.79
New South Wales	114	9,741	27,580,773	3,428.87
Northern Territory	11	289	315,300	48.72
New Zealand	74	18,029	84,058,661	10,175.80
Queensland	58	7,356	19,988,051	2,642.75
South Australia	50	3,537	13,856,275	1,716.72
Tasmania	21	1,260	5,086,867	636.99
Victoria	78	12,138	35,304,943	4,205.40
Western Australia	68	3,815	8,422,484	1,063.78
Total	482	56,815	195,528,896	24,030.82

CoANZSE data may therefore be able to provide researchers with a more realistic starting point for analyses of the geographical distribution of grammatical and syntactic features in spoken English in Australia and New Zealand. From a theoretical perspective, instead of a model in which a given feature is held to be categorically present (or absent) for a pre-defined language variety, CoANZSE data may show that syntactic features of English can be found in naturalistic speech in many locations: the question of their use is "in many cases a matter of statistical frequency rather than the presence or absence of a feature" (Kortmann, 2010, p. 843).

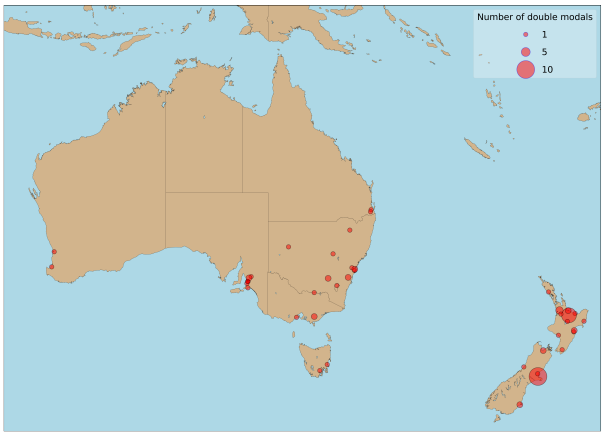

Figure 2: Verified double modal locations

3.2 Lexical distinctiveness

In order to test the hypothesis that Australian and New Zealand varieties of spoken English can be distinguished in CoANZSE, a simple machine learning model was created using Scikit-learn (Pedregosa et al., 2011). A sample of 10,000 randomly-selected CoANZSE transcripts was converted to term frequency-inverse document frequency (tf-idf) matrices using the 500 most common words in these transcripts, then trained using a linear support vector machine (Joachims, 1998) with 80% of the Australian and New Zealand transcripts, using parameters optimized with the GridSearchCV method in Scikit and balanced class weights. The model then predicted the country labels for the test data (1,359 Australian and 641 New Zealand transcripts). Model accuracy is summarized in Table 2.

The overall model accuracy of 0.80 suggests that there may be different usage patterns for common lexical items in discourse in Australian and New Zealand spoken English varieties. This preliminary finding, however, needs more thorough linguistic investigation. One approach would be to undertake a multi-dimensional analysis, using regular expressions to explore the frequencies of a number of grammatical and syntactic phenomena.

4 Caveats

Although the accuracy of ASR transcription systems continues to increase, transcripts of naturalistic speech contain errors due to factors such as audio recording quality, speech fluency or lack thereof, use of out-of-vocabulary words, slang, or dialect words, strong regional accent, or prosodic features (Aksënova et al., 2021). For a subset of CoANZSE videos, both ASR and manually-uploaded transcript files can be retrieved from YouTube; calculating the word error rate (WER) on the basis of these shared transcripts resulted in a value of 0.14, after careful filtering. YouTube transcripts are not diarized (i.e. have no indication of speaker turns), so they are not suitable "out-of-the-box" for analyses of language phenomena on the basis of social or demographic speaker traits. Two

Table 2: Binary classification results

Label	Precision	Recall	F1	Support	Accuracy
Australia	0.82	0.90	0.86	1359	0.80
New Zealand	0.74	0.59	0.66	641	

basic approaches for use of CoANZSE and similar data can be taken: First, the method of manual verification of targeted linguistic phenomena, utilized for the preliminary analysis of double modals noted above, can be done quickly because the transcripts are word-timed and linked to videos. This approach allows the analyst to identify and filter out transcript errors, as well as annotate additional features that may be of interest (for example, some speaker demographic traits). In a large-scale approach, a focus on relatively frequent features and broad geographical granularity will help to mitigate the effects of transcript errors, which would be outweighed by the greater frequency of correct transcriptions (Agarwal et al., 2007).

5 Summary and Outlook

CoANZSE is a large corpus of spoken English from Australia and New Zealand comprising ASR transcripts of YouTube videos uploaded by local councils and other local government entities. Two exploratory analyses using CoANZSE data attest use of double modals in naturalistic speech and show that Australian and New Zealand transcripts can be distinguished on the basis of their different rates of use of common words.

There are many possibilities for future work with CoANZSE data. Because the underlying video and audio recordings of CoANZSE transcripts are available, a script pipeline can be set up to retrieve video or audio excerpts for features of interest, which can then be analyzed using common tools such as ffmpeg and Praat. Such an approach permits, for example, the semi-automatic analysis of acoustic and prosodic properties of speech such as formant frequencies or pitch contours; video data retrieved using a scripting pipeline approach could be used for corpus-based analysis of multi-modal aspects of communication.

A tantalizing possibility for CoANZSE data is to shed light on the possible development of regional varieties of English within Australia and New Zealand in terms of pronunciation (Cox and Palethorpe, 2019), lexis, and grammar. For Australia, previous studies have mostly maintained that little regional variation is evident, at least in grammar or syntax, a situation usually held to result from the relatively young age of the variety (Murray and Manns, 2020). As noted by Burridge, however, the necessary components for regional diversification, namely "time, physical/social distance and the processes of linguistic change" (2020, p. 185), are in place in the broader Australian English speech community.

Finally, because CoANZSE contains transcripts of public meetings and content broadcast by local government entities, its content may prove to be useful for discourse analyses of a broad range of contemporary political and cultural phenomena such as environmental issues, migration, elections, or other topics.

Widespread use of video streaming and sharing sites and ASR transcription have in recent years opened up new sources of data for the empirical study of language. It is hoped that the CoANZSE resource will allow researchers to gain new insights into the current status of English in Australia and New Zealand and thus further our understanding of ongoing the development and diversification of the language.

References

Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. 2007. How much noise is too much: A study in automatic text classification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 3–12.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, Online. Association for Computational Linguistics.

Liselotte Anderwald and Suzanne Wagner. 2007. The Freiburg English Dialect Corpus: Applying corpus-linguistic research tools to the analysis of dialect data. In Joan C. Beal, Karen P. Corrigan, and Hermann Moisl, editors, *Creating and digitizing language corpora volume 1: Synchronic databases*, pages 35–53. Palgrave Macmillan, Houndmills, Basingstoke.

Julie Bradshaw, Kate Burridge, and Michael Clyne.

2010. The Monash Corpus of Spoken Australian English. In *Proceedings of the 2008 Conference of the Australian Linguistics Society*. Australian Linguistic Society.

Kate Burridge. 2020. History of Australian English. In Louisa Willoughby and Howard Manns, editors, *Australian English reimagined: Structure, features and developments*, pages 175–192. Routledge.

Steve Cassidy, Michael Haugh, Pam Peters, and Mark Fallu. 2012. The Australian national corpus: National infrastructure for language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3295–3299, Istanbul, Turkey. European Language Resources Association (ELRA).

Steven Coats. 2022. *Naturalistic double modals in North America*. American Speech.

Steven Coats. In review. Double Modals in contemporary British and Irish Speech.

Karen P. Corrigan, Isabelle Buchstaller, Adam Mearns, and Hermann Moisl. 2012. The Diachronic Electronic Corpus of Tyneside English.

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.

John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. Santa Barbara Corpus of Spoken American English.

Jonathan Dunn. 2019. *Global syntactic variation in seven languages: Toward a computational dialectology*. Frontiers in Artificial Intelligence, Section Language and Computation.

Elizabeth Gordon, Margaret Maclagan, and Jennifer Hay. 2007. The ONZE corpus. In Joan C. Beal, Karen P. Corrigan, and Hermann Moisl, editors, *Creating and digitizing language corpora volume 2: Diachronic databases*, pages 82–104. Palgrave Macmillan, Houndmills, Basingstoke.

Sidney Greenbaum, editor. 1996. *Comparing English worldwide: The International Corpus of English*. Clarendon Press.

Sidney Greenbaum. 1998. A proposal for an international computerized corpus of english. *World Englishes*, 7(3):315.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence*, 2.

Michael Haugh and Wei-Lin Melody Chang. 2013. Collaborative creation of spoken language corpora. In Tim Greer, Yuriko Kite, and Donna Tatsuki, editors, *Pragmatics and Language Learning, Volume 13*, pages 133–159. National Foreign Language Resource Center, University of Hawaii.

Janet Holmes, Bernadette Vine, and Gary Johnson. 1998. Guide to the Wellington Corpus of Spoken New Zealand English.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394.

Thorsten Joachims. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Bernd Kortmann. 2010. Areal variation in syntax. In Peter Auer and Jürgen E. Schmidt, editors, *Language and space: An international handbook of linguistic variation, volume 1, theories and methods*, pages 837–64. de Gruyter Mouton.

Bernd Kortmann and Kerstin Lunkenheimer, editors. 2013. *The Electronic World Atlas of Varieties of English (eWAVE)*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Lee Murray and Howard Manns. 2020. Lexical and morphosyntactic variation in australian english. In Louisa Willoughby and Howard Manns, editors, *Australian English reimagined: Structure, features and developments*, pages 120–133. Routledge.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3:175–198.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Benedikt Szmrecsanyi. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora*, 6(1):45–76.