

CoANZSE: The Corpus of Australian and New Zealand Spoken English



Steven Coats

English, University of Oulu, Finland

steven.coats@oulu.fi

Computational Thinking in the Humanities Online Workshop

September 1st, 2022

Background

- Renaissance in corpus-based study of English varieties (Nerbonne 2009; Szmrecsanyi 2011, 2013; Grieve et al. 2019)
- Many available corpora of transcribed spoken English are small or lack sufficient geographical granularity to make inferences about regional distributions of features

Corpus	Location(s)	nr_words	Reference
FRED	Britain	~2.5m	Anderwald & Wagner 2007
SCOTS Corpus	Scotland	~1m	Corbett 2014
NECTE/DECTE	Newcastle/Tyneside	~315k	Corrigan et al. 2012
Santa Barbara Corpus	US	~249k	Du Bois et al. 2000-2005
ICE-Ireland	Ireland	~600k	Kallen & Kirk 2007
ICE-Aus	Australia	~600k	Cassidy et al. 2012
Monash Corpus	Melbourne	~96k	Bradshaw et al. 2010
Griffith Corpus	Brisbane	~32k	Cassidy et al. 2012

- Automatic Speech Recognition (ASR) transcripts are available online for speech from specific locations
- Videos from local councils and other government entities can be harvested to create large corpora

Example video

Maranoa Regional Council - Ordinary Meeting - 24 Novemb...



WebVTT file

```
1 WEBVTT
2 Kind: captions
3 Language: en
4
5 00:00:01.160 --> 00:00:06.550 align:start position:0%
6
7 [Music]
8
9 00:00:06.550 --> 00:00:06.560 align:start position:0%
10 [Music]
11
12
13 00:00:06.560 --> 00:00:08.150 align:start position:0%
14 [Music]
15 uh<00:00:06.960><c> welcome</c>
16
17 00:00:08.150 --> 00:00:08.160 align:start position:0%
18 uh welcome
19
20
21 00:00:08.160 --> 00:00:10.950 align:start position:0%
22 uh welcome
23 i'd<00:00:08.320><c> like</c><00:00:08.480><c> to</c><00:00:08.639><c> open</c><00:00:08.880><c> the</c><00:00:09.040><c> meeting</c><00:00:09.360><c> at</c><00:00:09.519><c>
24
25 00:00:10.950 --> 00:00:10.960 align:start position:0%
26 i'd like to open the meeting at 9 12 a.m
27
28
29 00:00:10.960 --> 00:00:13.190 align:start position:0%
30 i'd like to open the meeting at 9 12 a.m
31 thank<00:00:11.200><c> you</c><00:00:11.280><c> for</c><00:00:11.440><c> your</c><00:00:11.599><c> attendance</c>
32
```

YouTube ASR Corpora

US, Canada, England, Scotland, Wales, Northern Ireland, the Republic of Ireland, Australia, New Zealand, and Germany

- **CoNASE**: 1.25b token corpus of 301,846 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts (Coats forthcoming a)
- **CoBISE**: 112m tokens, 452 locations, 38,680 ASR transcripts (Coats forthcoming b)
- **CoGS**: 50.5m tokens, 39.5k transcripts, 1,308 locations (Coats in review b)
- **CoANZSE**: 190m tokens, 57k transcripts, 482 locations (Coats in review b)

Freely available for research use; download from the Harvard Dataverse ([CoNASE](#), [CoBISE](#), [CoGS](#), [CoANZSE](#))

Data format

	country	state	name	channel_name	channel_url	video_title	video_id	upload_date	video_length	text_pos	location	latlong	nr_words
0	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Road Resurfacing Video	zVr6S5Xkj28	20181127	146.120	g_NNP_2.75 'day_XX_2.75 my_PRP\$ 3.75 name_NN_4.53 is_VBZ_4.74 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	433
1	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Weather update 5pm 1 March 2022 - Mayor Matt Gould	p4MjirCc1oU	20220301	181.959	hi_UH_0.64 guys_NNS_0.96 i_PRP_1.439 'm_VBP_1.439 just_RB_1.76 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	620
2	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Transport Capital Works Video	DXIkVTcmeho	20180417	140.450	council_NNP_0.53 is_VBZ_1.53 placing_VBG_1.65 is_VBZ_2.07 2018-19_CD_2.57 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	347
3	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Council Meeting Wrap Up February 2022	2NhuhF2fBu8	20220224	107.840	g_NNP_0.399 'day_NNP_0.399 guys_NNS_0.799 and_CC_1.12 welcome_JJ_1.199 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	341
4	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	CITY DEAL 4 March 2018	4-cv69ZcwVs	20180305	130.159	[Music]_XX_0.85 it_PRP_2.27 's_VBZ_2.27 a_DT_3.27 fantastic_JJ_3.36 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	420

Focus on regional and local council channels

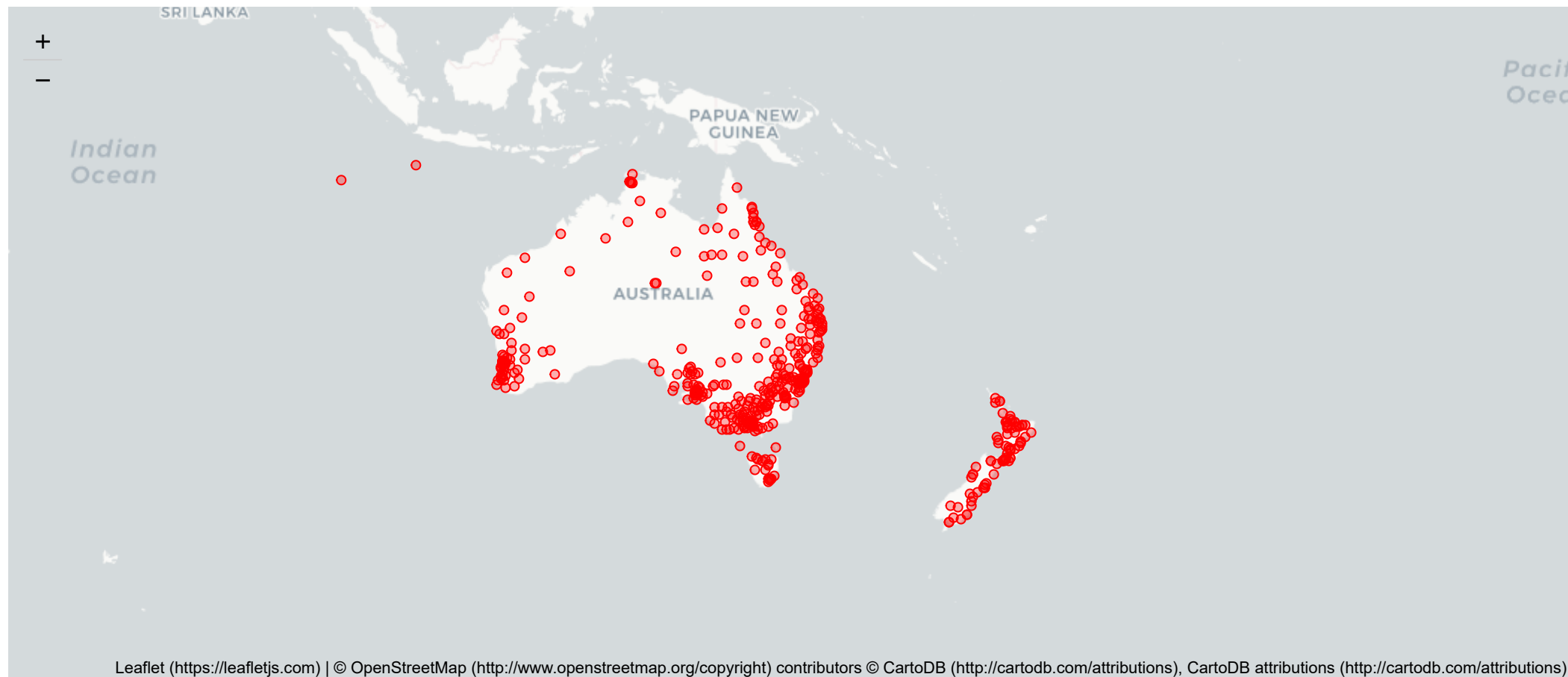
Many recordings of meetings of elected councillors: advantages in terms of representativeness and comparability

- Speaker place of residence (cf. videos collected based on place-name search alone)
- Topical contents and communicative contexts comparable
- In most jurisdictions government content is in the public domain

Data collection and processing

- Identification of relevant channels (lists of councils with web pages -> scrape pages for links to YouTube)
- Inspection of returned channels to remove false positives
- Retrieval of ASR transcripts using **YT-DLP**
- Geocoding: String containing council name + address + country location to Google's geocoding service
- PoS tagging with SpaCy (Honnibal et al. 2019)

CoANZSE channel locations



CoANZSE corpus size by country/state/territory

Territory	nr_channels	nr_videos	nr_words	video_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

Example analysis: Double modals

- Non-standard rare syntactic feature (Montgomery & Nagle 1994; Coats 2022)
- **I might could help you with this**
- Occurs only in the American Southeast and in Scotland/Northern England/Northern Ireland?
- Most studies based on non-naturalistic data with limited geographical scope (data from LAMSAS, LAGS, surveys administered mostly in American Southeast and North of Britain)
- More widely used in North America and the British Isles than previously thought (Coats 2022, Coats in review)
- Not yet reported in Australian or NZ speech

Script: Generating a table for manual inspection of double modals

- Base modals *will, would, can, could, might, may, must, should, shall, used to, 'll, ought to, oughta*
- Script to generate regexes of two-tier combinations

```
import re
hits = []
for i,x in coanzse_df.iterrows():
    pat1 = re.compile("("+x[0]+"_\\w+_\\s+\\s"+x[1]+"_\\w+_\\s+\\s)", re.IGNORECASE)

    if pat1.search(x["text_pos"]):
        finds1 = pat1.findall(x["text_pos"])[0]
        seq = " ".join([x.split("_")[0] for x in finds1[0].split()])
        time = finds1[0].split()[0].split("_")[-1]
        hits.append((x["country"],x["channel_title"],seq,"https://youtu.be/"+x["video_id"]+"?t="+str(round(float(time)-3))))
pd.DataFrame(hits)
```

Excerpt from generated table

	Location	Channel	Video	DM	Link	Type	Notes
1	NSW	Central Darling Shire Council	24 February 2021 Part 2	would might	https://youtu.be/4JhDv6H_rMQ?t=63	t	"however, the senior planning officer would might may want to make comment"
2	NSW	Dubbo Regional Council	Dubbo City Council State of the City Report 2014	'll can	https://youtu.be/zOyDAMACmFk?t=190	t	"we'll, we'll can forget about that plan for a while"
3	NSW	Inner West Council	Speaker Series - Shiver with Allie Reynolds	would might	https://youtu.be/WrmDQhsqv5s?t=568	t	also in embedded manual transcript
4	NSW	Ku-ring-gai Council	3D Bushfire Simulation and CWC Workshop	might would	https://youtu.be/KhxiXPQBFXs?t=1232	t	"for anything that might would... go wrong"
5	NSW	Ku-ring-gai Council	Ordinary Meeting of Council 20_08_2019	would might	https://youtu.be/n80tXfiqQzA?t=6192	t	
6	NSW	mosmancouncil	Mosman Art Prize - In Conversation Salote Tawale	might could	https://youtu.be/jQbDqA1yvHM?t=117	t	
7	NSW	Wingecarribee Shire Council	Extraordinary Council Meeting 16 Feb 2022	would might	https://youtu.be/kwGrKSIIcQ?t=2997	t	"if you would might just convey"
8	NSW	Wingecarribee Shire Council	Ordinary Meeting of Council 13 May 2020 - part one	would might	https://youtu.be/whP9EfvuouQ?t=3822	t	"if they could move them down the hill further, I think they would might find that"
9	NSW	Hunter Joint Organisation	Hunter Global Summit Day 1 Session 1	will can	https://youtu.be/6kHJjMugPs?t=2351	t	

Showing 1 to 57 of 57 entries

Previous 1 Next

Finding features

- Regular-expression-search and manual annotation approach
- Double modals can be found in the US North and West and in Canada; in Scotland, N. Ireland, and N. England, but also in the English Midlands and South and in Wales (Coats in Review)
- Also in Australia and New Zealand!

A few caveats

- Videos of local government not representative of speech in general
- ASR errors (mean WER after filtering ~14%), quality of transcript related to quality of audio as well as dialect features (Tatman 2017; Meyer et al. 2020; Markl & Lai 2021)
 - Low-frequency phenomena: manually inspect corpus hits
 - High-frequency phenomena: signal of correct transcriptions will be stronger (Agarwal et al. 2009) → classifiers

Thank you!

References

- Agarwal, S., S. Godbole, D. Punjani & S. Roy. 2007. *How much noise is too much: A study in automatic text classification*. In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 3–12.
- Anderwald, L. & S. Wagner. 2007. The Freiburg English Dialect Corpus: Applying corpus-linguistic research tools to the analysis of dialect data. In: J. C. Beal, K. P. Corrigan & H. Moisl (Eds.), *Creating and digitizing language corpora volume 1: Synchronic databases*, 35–53. Palgrave Macmillan.
- Bradshaw, J., Burrige, K. & Clyne, M. (2010). The Monash Corpus of Spoken Australian English. In de L. Beuzeville & P. Peters (Eds.), *Proceedings of the 2008 Conference of the Australian Linguistics Society*.
- Cassidy, S., Haugh, M., Peters, P., & Fallu, M. (2012). The Australian National Corpus: National infrastructure for language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3295–3299. http://www.lrec-conf.org/proceedings/lrec2012/pdf/400_Paper.pdf
- Coats, S. In review a. Double modals in contemporary British and Irish Speech.
- Coats, S. In review b. New Corpora of Geolocated ASR Transcripts from Australia/New Zealand and Germany.
- Coats, S. Forthcoming a. Dialect corpora from YouTube. *Proceedings of ICAME41*. De Gruyter.
- Coats, S. Forthcoming b. The Corpus of British Isles Spoken English (CoBISE) A New Resource of Contemporary British and Irish Speech. *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, Uppsala, Sweden, March 15-18, 2022.
- Coats, S. 2022. *Naturalistic double modals in North America*. *American Speech*.
- Corbett, J. 2014. Syntactic variation: Evidence from the Scottish Corpus of Text and Speech. In: R. Lawson (Ed.), *Sociolinguistics in Scotland*, 258–276. Palgrave Macmillan.
- Corrigan, K. P., I. Buchstaller, A. Mearns & H. Moisl. 2012. *The Diachronic Electronic Corpus of Tyneside English*.
- Du Bois, J. W., W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson & N. Martey. 2000-2005. Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

References II

- Grieve, J., C. Montgomery, A. Nini, A. Murakami & D. Guo. 2019. [Mapping lexical dialect variation in British English using Twitter](#). *Frontiers in Artificial Intelligence* 2.
- Honnibal, M., I. Montani, H. Peters, S. V. Landeghem, M. Samsonov, J. Geovedi, J. Regan, G. Orosz, S. L. Kristiansen, P. O. McCann, D. Altinok, Roman, G. Howard, S. Bozek, E. Bot, M. Amery, W. Phatthiyaphaibun, L. U. Vogelsang, B. Böing, P. K. Tippa, jeannefukumaru, G. Dubbin, V. Mazaev, R. Balakrishnan, J. D. Møllerhøj, wbwseeker, M. Burton, thomasO & A. Patel. 2019. [Explosion/spaCy v2.1.7: Improved evaluation, better language factories and bug fixes](#).
- Kallen, J. & J. Kirk. 2007. ICE-Ireland: Local variations on global standards. In: J. C. Beal, K. P. Corrigan & H. Moisl (Eds.), *Creating and digitizing language corpora volume 1: Synchronic databases*, 121–162. Palgrave Macmillan.
- Markl, N. & C. Lai. 2021. [Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation](#). In: *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, Association for Computational Linguistics, 34–40. Association for Computational Linguistics.
- Meyer, J., L. Rauchenstein, J. D. Eisenberg & N. Howell. 2020. [Artie bias corpus: An open dataset for detecting demographic bias in speech applications](#). In: *Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020*, 6462–6468.
- Montgomery, M. B. & S. J. Nagle. 1994. Double modals in Scotland and the Southern United States: Trans-atlantic inheritance or independent development? *Folia Linguistica Historica* 14, 91–108.
- Nerbonne, J. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3, 175–198.
- Szmrecsanyi, B. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge University Press.
- Szmrecsanyi, B. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora* 6, 45–76.
- Tatman, R. 2017. [Gender and dialect bias in YouTube’s automatic captions](#). In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. Association for Computational Linguistics.