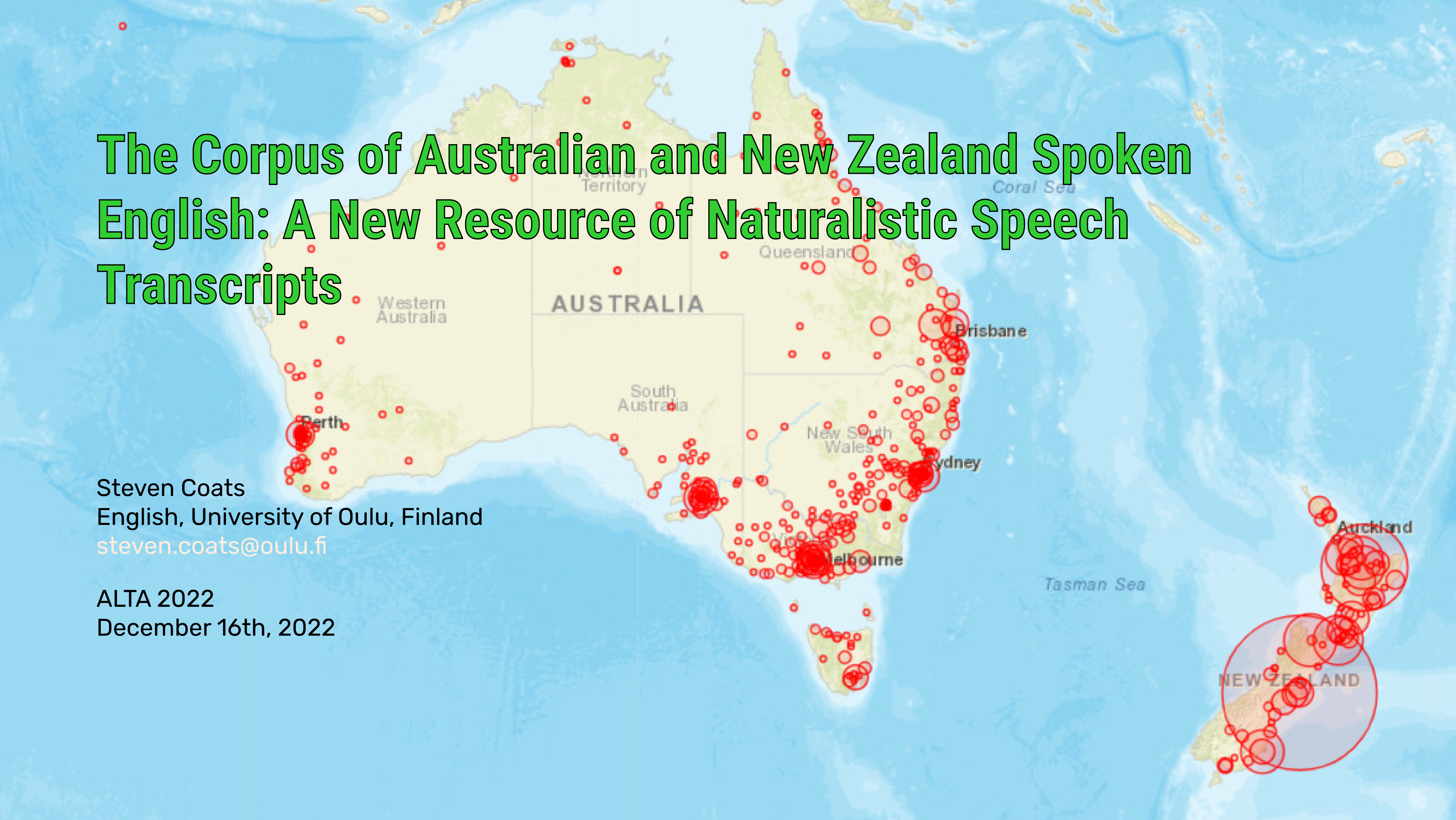
# The Corpus of Australian and New Zealand Spoken English: A New Resource of Naturalistic Speech Transcripts



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

ALTA 2022
December 16th, 2022

# Background

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

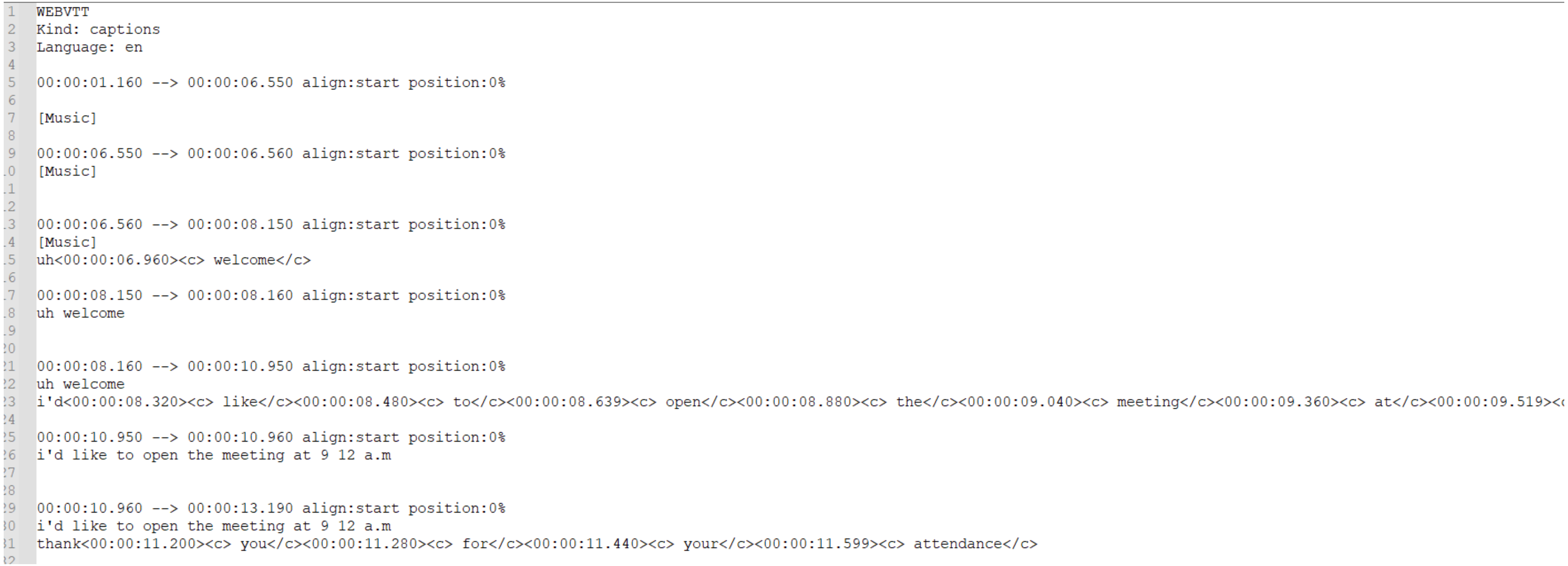
| Corpus | Location(s) | nr_words | Reference |
|---|---|---|---|
| ICE-Aus | Australia | ~600k | Cassidy et al. 2012 |
| Monash Corpus | Melbourne | ~96k | Bradshaw et al. 2010 |
| Griffith Corpus | Brisbane | ~32k | Cassidy et al. 2012 |
| Wellington Corpus | NZ | ~1m | Holmes et al. 1998 |
| ONZE Corpus | NZ | ? | Gordon et al. 2007 |

- 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



```
WEBVTT
Kind: captions
Language: en

00:00:01.160 --> 00:00:06.550 align:start position:0%

[Music]

00:00:06.550 --> 00:00:06.560 align:start position:0%
[Music]


00:00:06.560 --> 00:00:08.150 align:start position:0%
[Music]
uh<00:00:06.960><c> welcome</c>

00:00:08.150 --> 00:00:08.160 align:start position:0%
uh welcome


00:00:08.160 --> 00:00:10.950 align:start position:0%
uh welcome
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><

00:00:10.950 --> 00:00:10.960 align:start position:0%
i'd like to open the meeting at 9 12 a.m


00:00:10.960 --> 00:00:13.190 align:start position:0%
i'd like to open the meeting at 9 12 a.m
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>
```

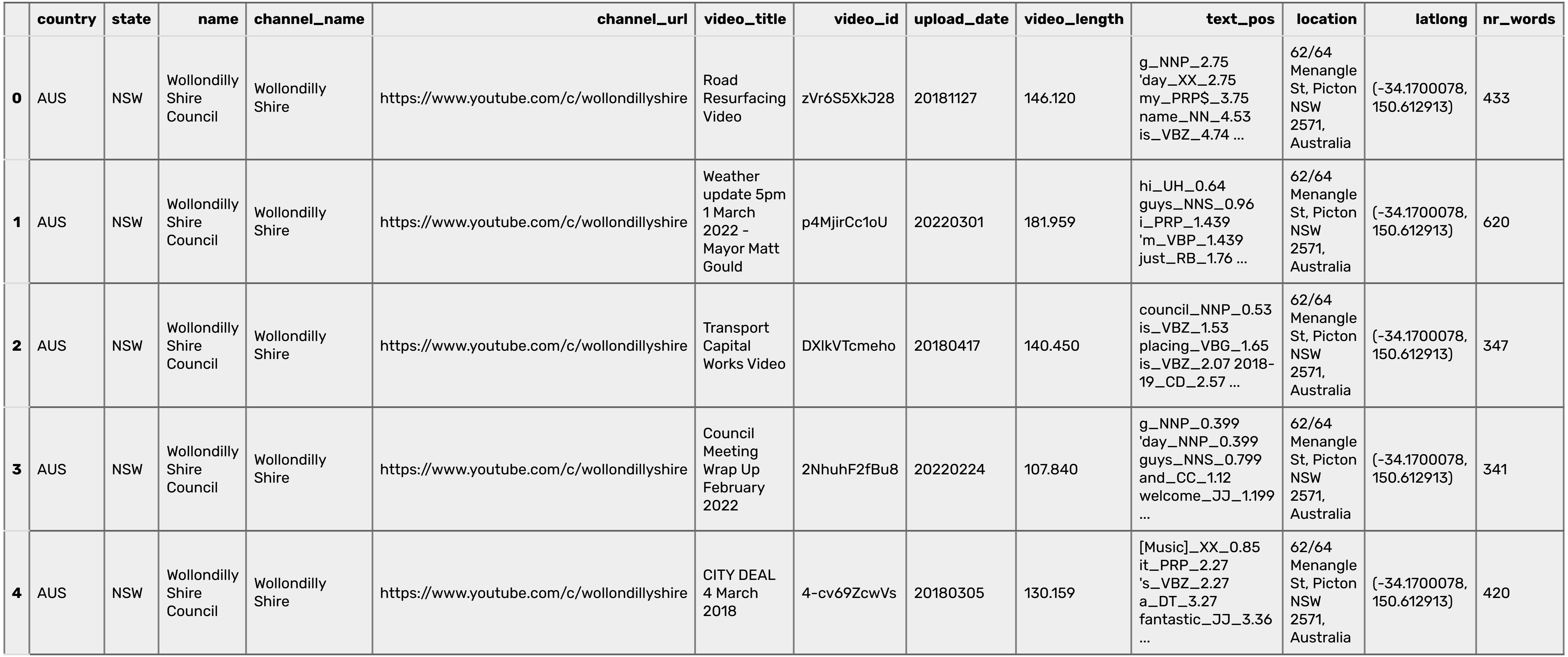
# YouTube ASR Corpora

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

- 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 2022b)
- CoGS: 50.5m tokens, 39.5k transcripts, 1,308 locations (Coats in review)
- CoANZSE: 190m tokens, 57k transcripts, 482 locations

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 | DXlkVTcmeho | 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

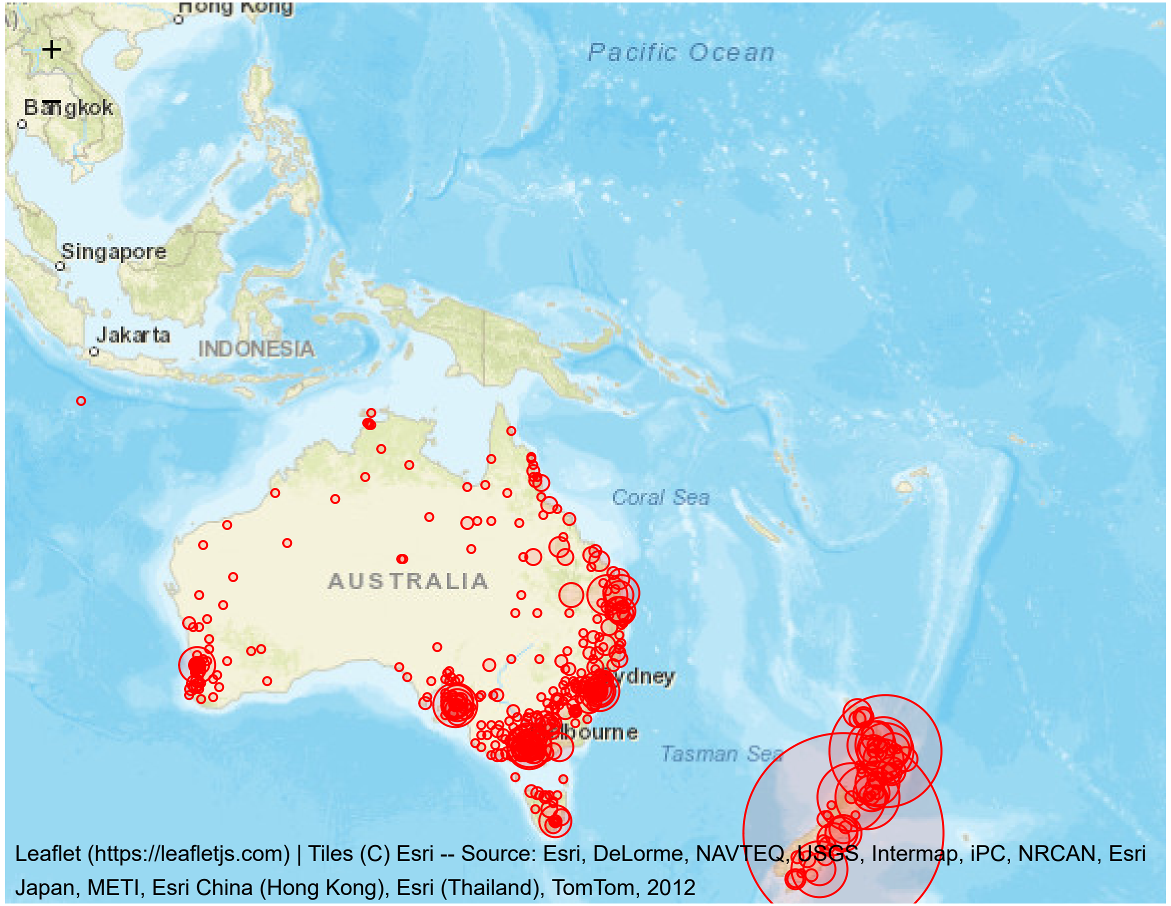
- Speaker place of residence (cf. videos collected based on place-name search alone)
- Topical contents and communicative contexts comparable
- In most jurisdictions government content is in the public domain

# 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

Circle size corresponds to channel size in number of words



Leaflet (https://leafletjs.com) | Tiles (C) Esri -- Source: Esri, DeLorme, NAVTEQ, USGS, Intermap, iPC, NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), TomTom, 2012

# 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 2022a)
    - *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 linguistic atlas interviews, surveys administered mostly in American Southeast and North of Britain)
- More widely used in North America and the British Isles than previously thought (Coats 2022a, Coats in review)
- Little studied in Australian and New Zealand speech

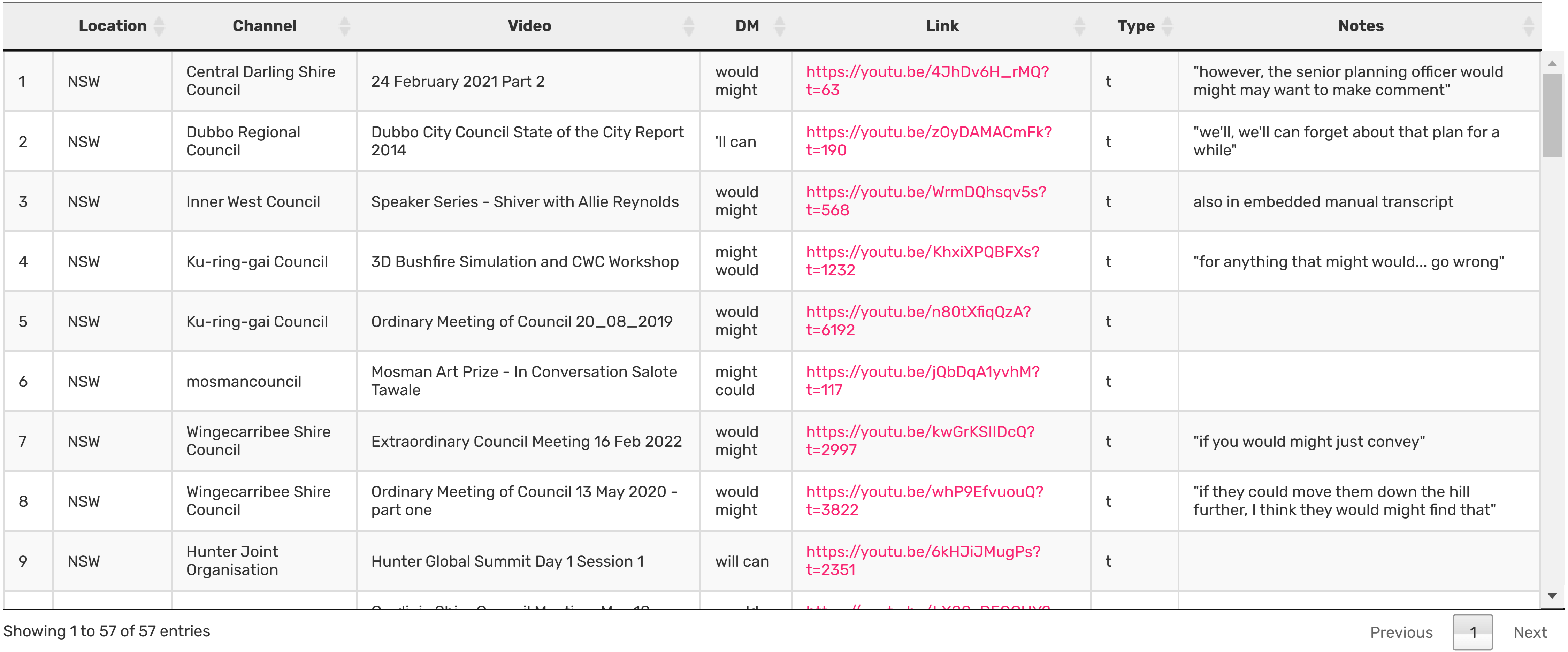
# Script: Generating a table for manual inspection of double modals

- 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 x in modals:
  for i,y in coanzse_df.iterrows():
      pat1 = re.compile("("+x[0]+"_\\w+_\\S+\\s+"+x[1]+"_\\w+_\\S+\\s)",re.IGNORECASE)
      finds = pat1.findall(y["text_pos"])
      if finds:
          for z in finds:
            seq = z.split()[0].split("_")[0].strip()+" "+z.split()[1].split("_")[0].strip()
            time = z.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)
```

- The script creates a URL for each search hit at a time 3 seconds before the targeted utterance
- In the resulting data frame, each utterance can be annotated after examining the targeted video sequence
- Filter out non-double-modals (clause overlap, speaker self-repairs, ASR errors)

# 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/kwGrKSIIDcQ?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/6kHJiJMugPs?t=2351 | t | |

Showing 1 to 57 of 57 entries

# Training a classifier on the basis of common word types

- Simple machine-learning classifiers using SVM, logistic regression, or other algorithms can distinguish between Australian and NZ transcripts on the basis of the 500 most common words in CoANZSE

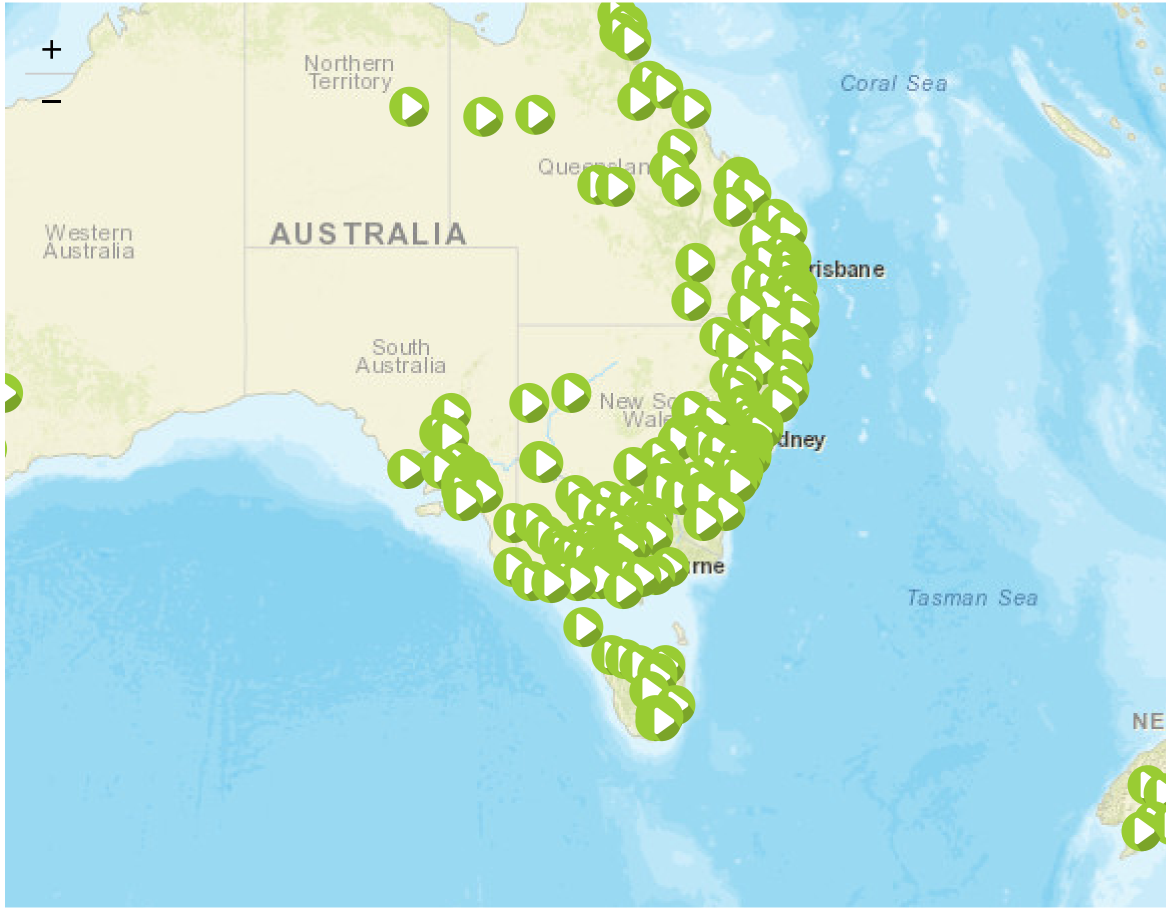
| | Precision | Recall | F1 | Support | Accuracy |
|---|---|---|---|---|---|
| Australia | 0.82 | 0.90 | 0.86 | 1,359 | 0.80 |
| New Zealand | 0.74 | 0.59 | 0.66 | 641 | |

# Pipeline for acoustic analysis

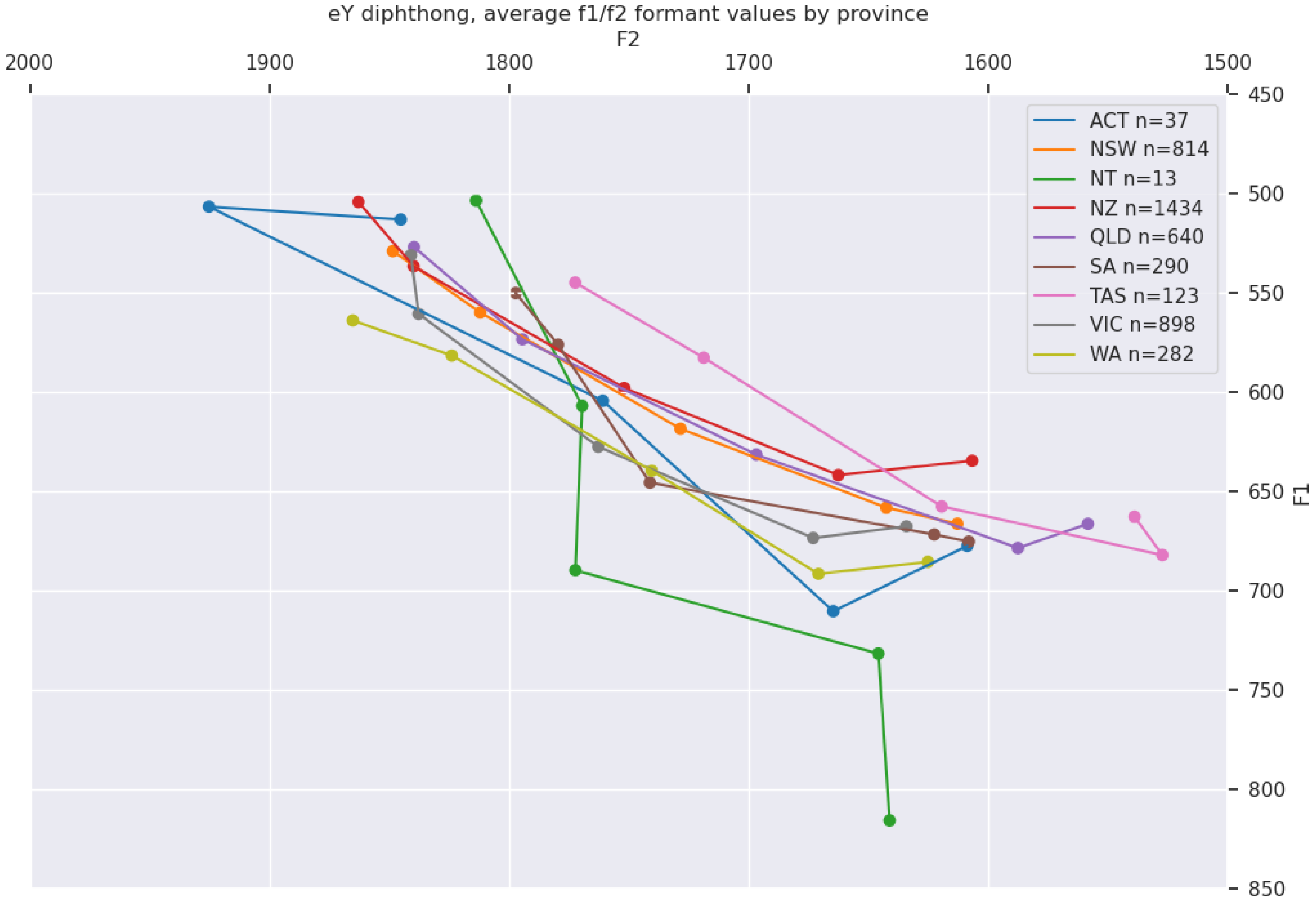
- Regular expressions to target specific words/phrases in the corpus
- Extract audio span containing the targeted item(s) from YT stream
- Feed audio and transcript excerpt to forced aligner
- Extract desired sounds/acoustic phenomena

# Extracted *today* tokens

A selection of *today* realizations from CoANZSE videos



# Average eɪ diphthong



# 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., Godbole, S., Punjani, D., & Roy, S. (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.

Bradshaw, J., Burridge, K., & Clyne, M. (2010). The Monash Corpus of Spoken Australian English. In L. de 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). Double modals in contemporary British and Irish Speech.

Coats, S. (Forthcoming). Dialect corpora from YouTube. In B. Busse, N. Dumrukcic, & I. Kleiber (Eds.), *Lanugage and linguistics in a complex world*. De Gruyter.

Coats, S. (2022a). Naturalistic double modals in North America. *American Speech*.

Coats, S. (2022b). The Corpus of British Isles Spoken English (CoBISE): A new resource of contemporary British and Irish speech. In K. Berglund, M. La Mela, & I. Zwart (Eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference, Uppsala, Sweden, March 15–18, 2022*, 187–194. Aachen, Germany: CEUR.

Gordon, E., Maclagan, M. & Hay, J. (2007). The ONZE corpus. In J. C. Beal, K. P. Corrigan, & H. Moisl (Eds.) *Creating and digitizing language corpora volume 2: Diachronic databases*, 82–104.Palgrave Macmillan.

Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2.

# References II

Holmes, J., Vine, B., & Johnson, G. (1998). *Guide to the Wellington Corpus of Spoken New Zealand English*.

Honnibal, M. et al. (2019). Explosion/spaCy v2.1.7: Improved evaluation, better language factories and bug fixes.

Markl, N. & Lai, C. (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., Rauchenstein, L., Eisenberg, J. D., & Howell, N. (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. & Nagle, S. J. (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.