Corpora of Automatic Speech Recognition Transcripts for the Study of Variation in English: Syntactic and Phonetic Perspectives

Steven Coats
English, University of Oulu, Finland
steven.coats@oulu.fi
**PAC 2023**
April 12th, 2023

Background

- Renaissance in corpus-based study of English varieties (Nerbonne 2009; Szmrecsanyi 2011, 2013; Grieve et al. 2019)
- Most large existing corpora consist of written language sourced from the web and social media
- Corpora of transcribed speech may have limited availability, are small in size, or lack sufficient geographical granularity to make inferences about regional distributions of features in speech

Corpus	Location	# 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 (spoken)	Australia	~600k	Cassidy et al. 2012
Spoken BNC2014	UK	~10m	Love et al. 2017; Brezina et al. 2018

- 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
- Big data analysis and visualization of grammar and phonetics

Example video
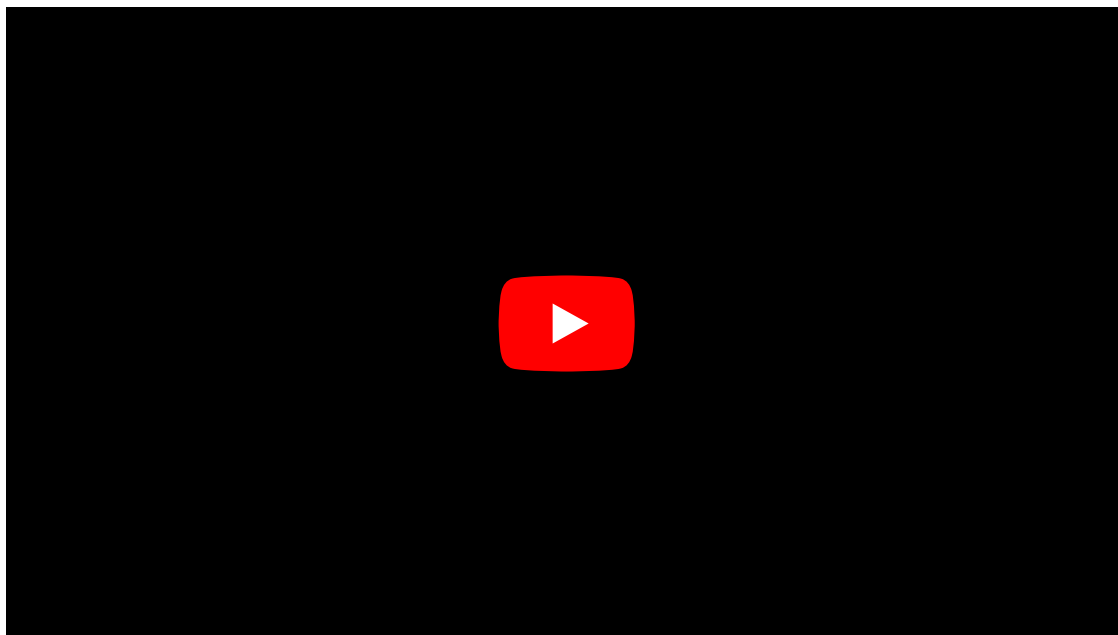
WebVTT file

```
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 captions files

- Videos can have multiple captions files: user-uploaded captions, auto-generated captions created using automatic speech recognition (ASR), or both, or neither

- User-uploaded captions can be manually created or generated automatically by 3rd-party ASR software

- Auto-generated captions are generated by YT's speech-to-text service

- CoNASE, CoANZSE, CoBISE: target YT ASR captions

YouTube ASR Corpora

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

- CoNASE: 1.25b token corpus of 301,846 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts (Coats 2023)
- CoBISE: 112m tokens, 452 locations, 38,680 ASR transcripts (Coats 2022b)
- CoANZSE: 190m tokens, 482 locations, 57k transcripts (Coats 2022b)

Also CoGS: 50.5m tokens, 1,308 locations, 39.5k transcripts (Coats in review)

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

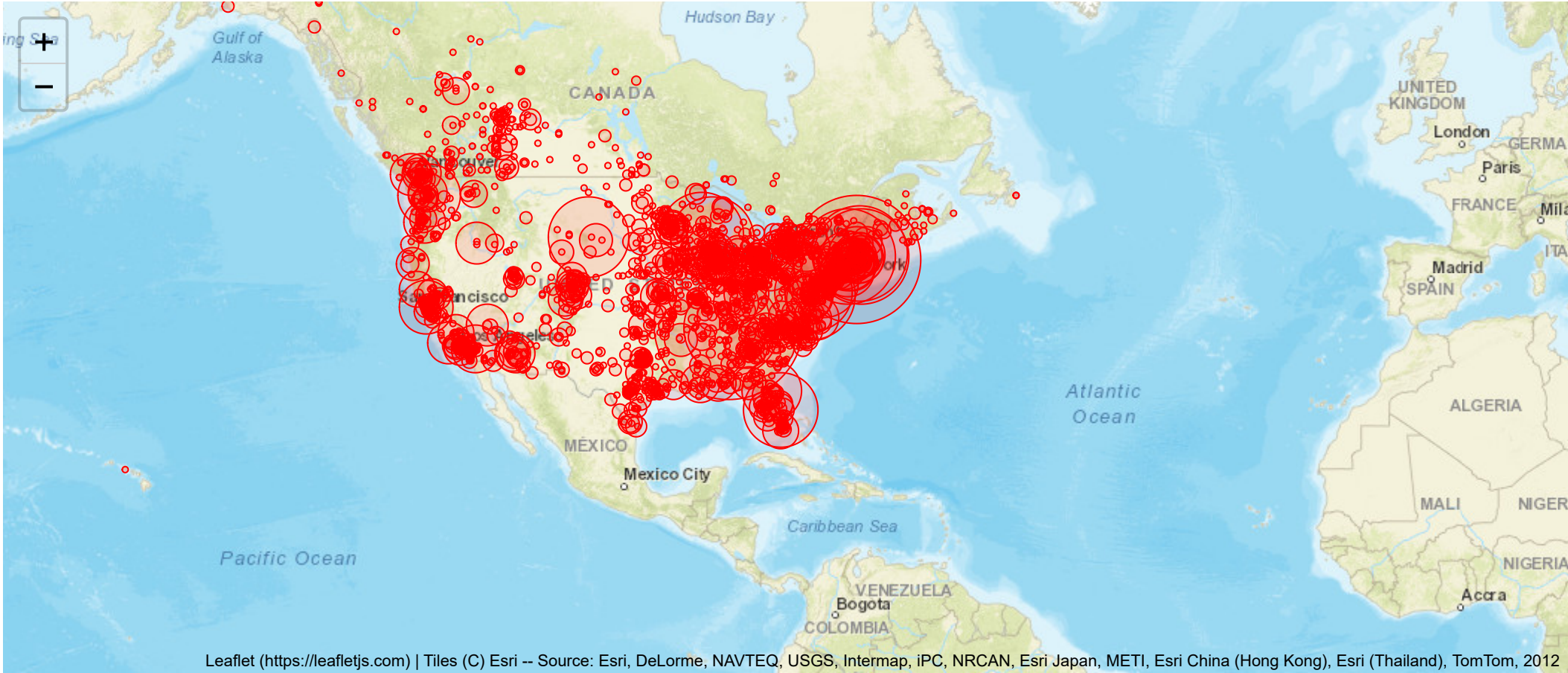# Data collection and processing

- 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)

CoNASE

State	Channels	Videos	Words	Length (h)	State	Channels	Videos	Words	Length (h)	State	Channels	Videos	Words	Length (h)
Alabama	27	2827	10,581,345	1,315.67	Michigan	90	9832	51,293,982	6,079.47	Texas	155	21,330	44,736,009	5,789.44
Alaska	6	451	1,854,654	248.37	Minnesota	80	8666	31,366,468	3,661.89	Utah	21	2,561	7,766,782	940.21
Arizona	35	6356	26,393,272	3,063.73	Mississippi	18	1448	2,613,901	346.07	Vermont	3	94	131,558	16.62
Arkansas	14	986	6,748,658	882.77	Missouri	53	5093	15,094,086	1,946.43	Virginia	42	9,209	34,806,149	4,059.67
California	211	18278	83,915,246	10,146.57	Montana	3	145	926,229	143.2	Washington	51	6,178	28,949,403	3,387.77
Colorado	56	8802	36,551,218	4,299.68	Nebraska	16	677	2,487,171	312.51	W. Virginia	6	101	196,479	25.86
Connecticut	25	3731	24,549,746	3,010.04	Nevada	5	2,759	6,110,915	638.06	Wisconsin	83	9,514	45,983,568	5,744.59
Delaware	3	148	242,073	25.45	N.H.	11	1,305	10,913,552	1,469.04	Wyoming	7	251	2,638,963	348.39
District of Columbia	3	242	261,209	32.9	New Jersey	88	6,982	29,523,334	3,977.57	Alberta	95	6,623	21,239,251	2,497.45
Florida	89	17625	64,647,923	7,468.48	New Mexico	14	1,895	6,750,477	883.1	British Columbia	102	10,002	26,853,481	3,246.83
Georgia	49	5487	18,565,796	2,421.53	New York	97	8,037	37,560,959	4,856.87	Manitoba	20	3,286	2,771,200	318.21
Hawaii	1	152	123,617	15.42	N. Carolina	97	11,357	46,231,979	5781.4	New Brunswick	8	382	2,347,141	278.05
Idaho	11	1547	8,747,885	1,012.14	N. Dakota	10	768	3,616,363	442.05	Newfoundland and Labrador	2	108	186,070	29.99
Illinois	151	14243	54,613,612	6,725.31	Ohio	97	7,647	33,695,476	4,268.46	Northwest Territories	3	32	21,404	3.27
Indiana	46	4017	12,958,084	1,643.88	Oklahoma	19	1,977	5,271,339	643.35	Nova Scotia	11	332	1,229,149	148.38
Iowa	43	7516	24,286,940	3,072.57	Oregon	38	2,769	15,675,898	1,992.84	Nunavut	1	6	1,230	0.23
Kansas	35	4444	19,862,293	2,504.08	Pennsylvania	74	6,984	32,571,217	3,970.32	Ontario	112	8,404	45,970,092	5,774.59
Kentucky	26	4965	17,834,978	2,092.75	Rhode Island	7	822	3,195,777	530.94	Prince Edward Island	6	753	777,772	95.87
Louisiana	16	2018	10,500,407	1,221.96	S. Carolina	24	3,894	8,716,589	1115.2	Quebec	6	166	486,265	60.29
Maine	12	819	5,879,165	797.01	S. Dakota	12	1,819	18,619,258	2,172.97	Saskatchewan	10	663	895,143	103.12
Maryland	32	7373	34,009,832	4,100.84	Tennessee	33	7,194	43,286,858	5,127.52	Yukon	7	159	257,171	30.48
Massachusetts	44	17596	11,517,230	14,682.19										

CoNASE channel locations

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

CoBISE

Country	Channels	Videos	Tokens	Length (h)
England	324	23,657	72,879,173	8,518.39
Northern Ireland	10	1,898	6,508,505	774.17
Republic of Ireland	26	2,525	6,264,276	680.81
Scotland	75	8,135	17,111,396	1,845.35
Wales	18	2,465	8,800,264	982.66
Total	453	38,680	111,563,614	12,801.38

CoBISE channel locations

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
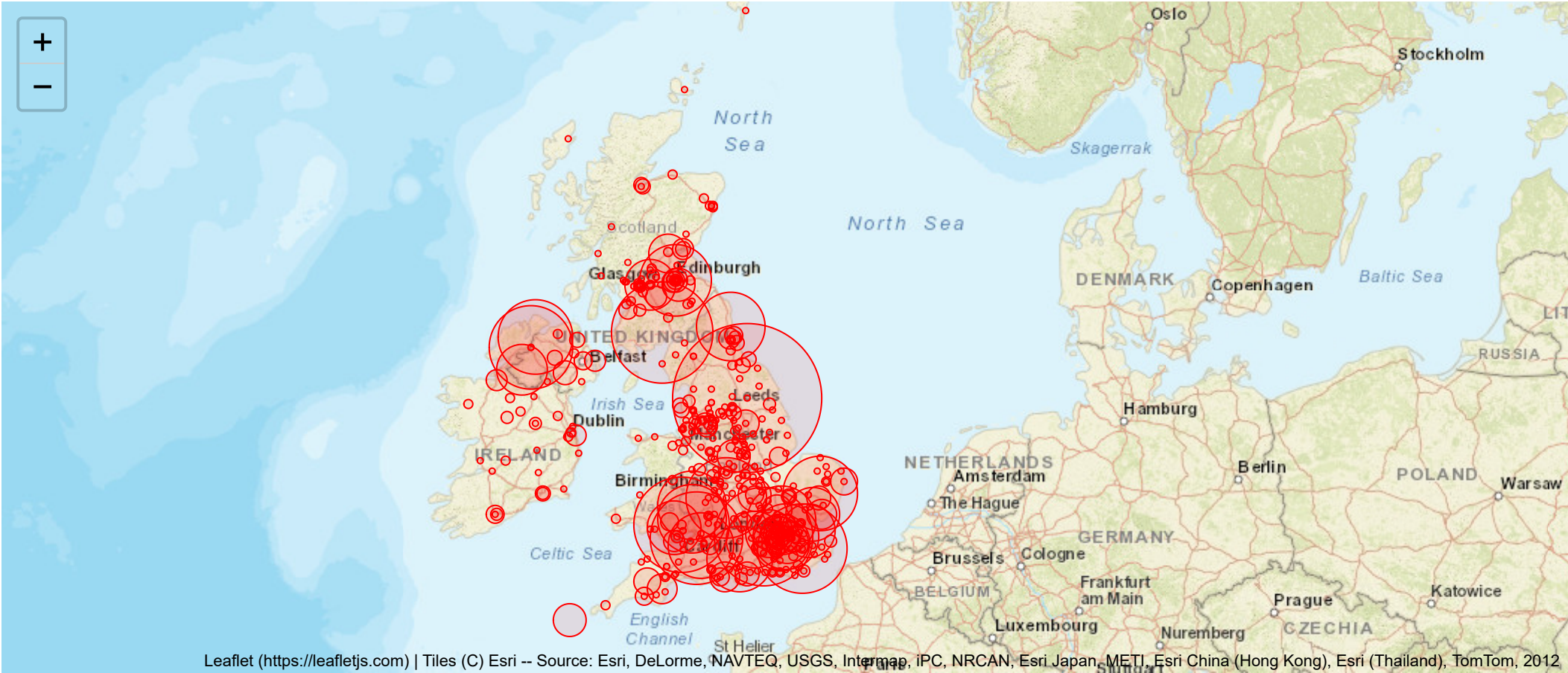
Territory	nr_channels	nr_videos	nr_words	video_length (h)
Australian Capital Territory	8	650	915,542	111.79
New South Wales	114	9,741	27,580,773	3,428.87
Northern Territory	11	289	315,300	48.72
New Zealand	74	18,029	84,058,661	10,175.80
Queensland	58	7,356	19,988,051	2,642.75
South Australia	50	3,537	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

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
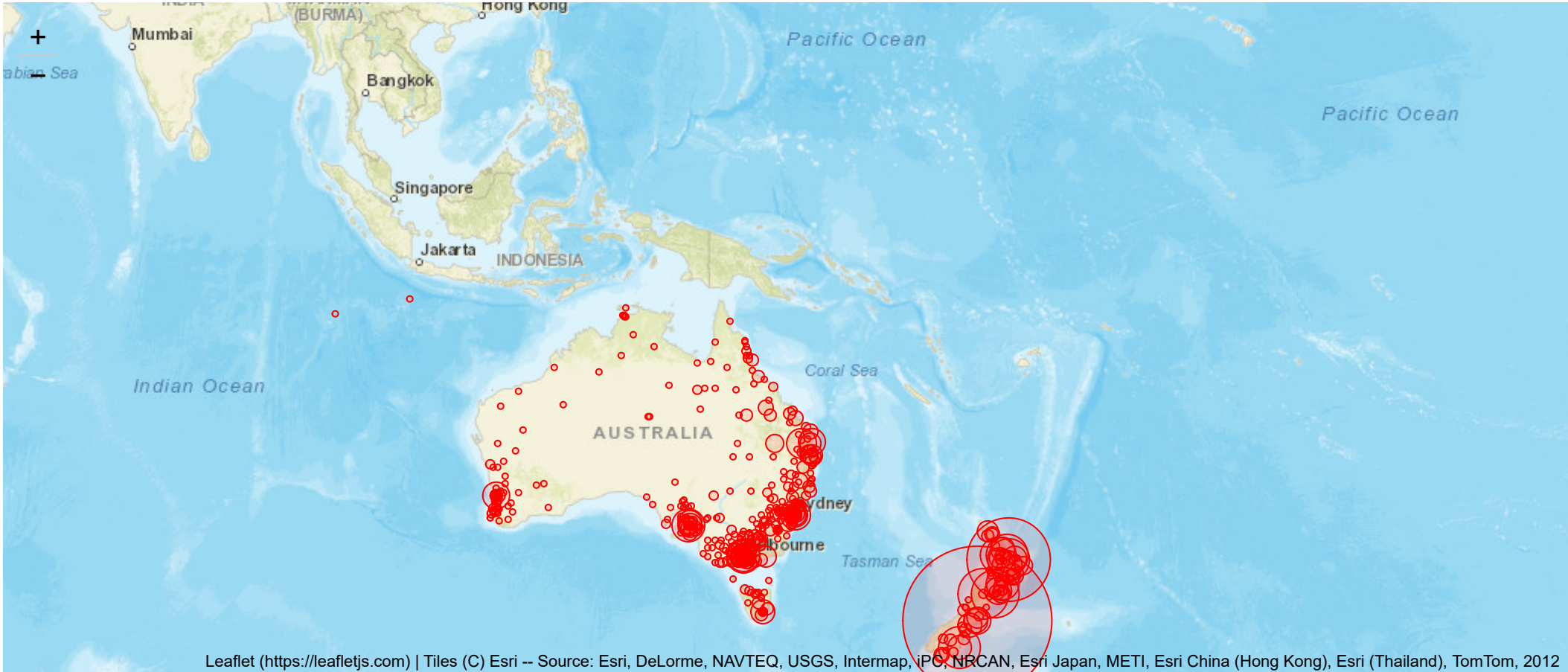
Corpus use cases: Syntax/grammar/pragmatics

- Regional variation in syntax, mood and modality
- Lexical items
- Contractions
- Hortatives/commands/interjections
- Pragmatics: Turn-taking, politeness markers
- Multidimensional analysis à la Biber
- Typological comparison at country/state/regional level

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

```python
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

Previous | 1 | Next

Pipeline for acoustic analysis (work in progress)

- Regular expressions to target specific words/phrases in the corpora
- Extract audio spans containing the targeted item(s) from YT stream
- Feed audio and transcript excerpt to forced aligner
- Extract desired sounds
- Measure acoustic phenomena of interest (formants, voice onset time, pitch, etc.)

Example: Excerpt from a council meeting in Gallatin, Tennessee (https://www.youtube.com/watch?v=yzjGnz_Rs7I)

Pipeline for acoustic analysis: Vowel formants

For each transcript/video in the collection:

- Regular expressions to search for words with [eɪ]
- yt-dlp to download audio segments in a window around the target word
- Feed the segments (audio and corresponding transcript segment) to the Montreal Forced Aligner; output is Praat TextGrids
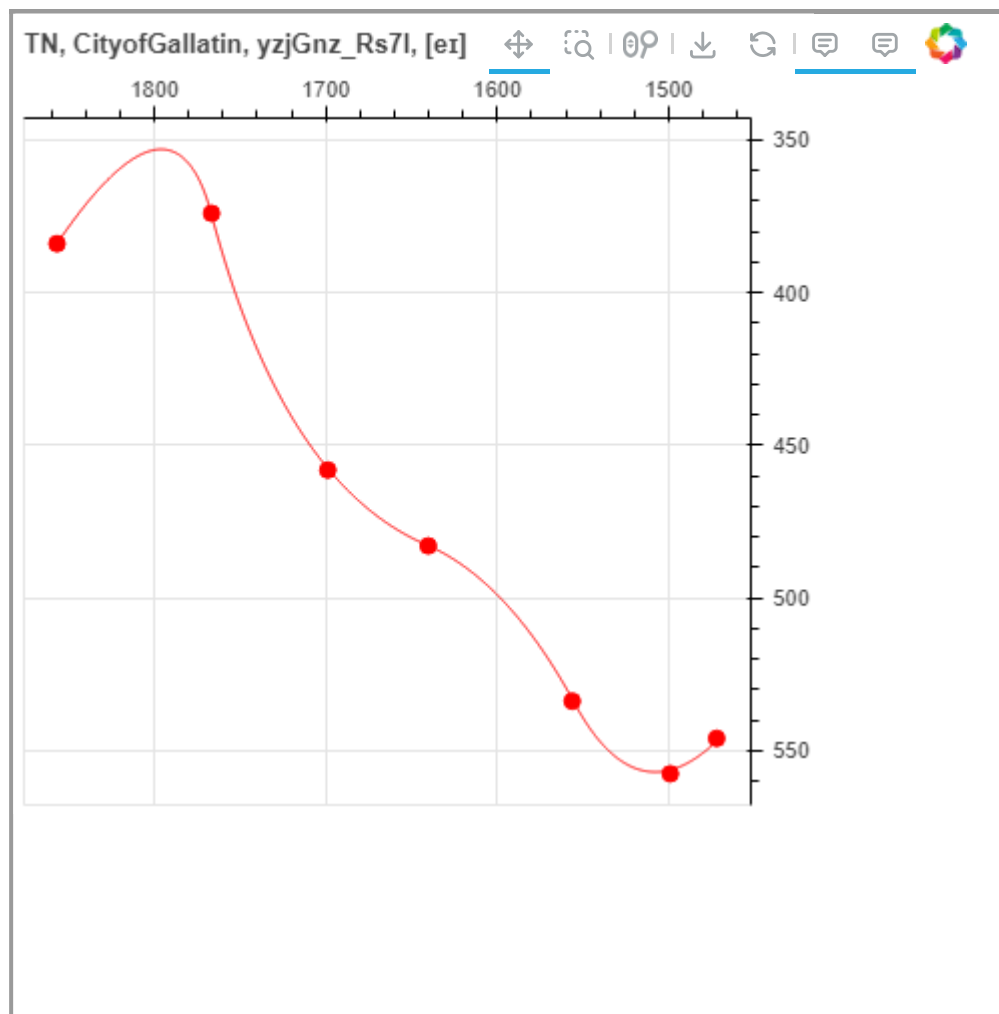- Select vowel(s) of interest using TextGrid timings and Parselmouth (Python port of Praat functions)

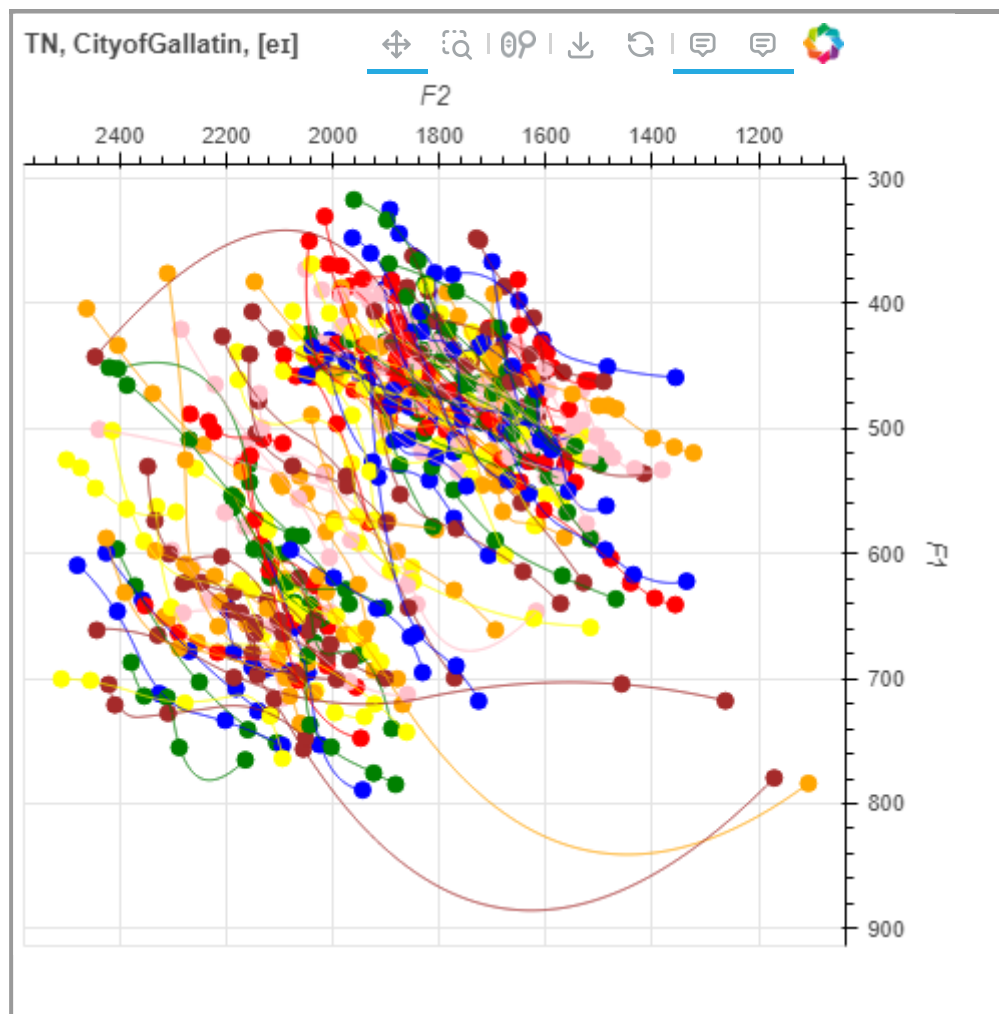`have a great d`**`ay`**` on that`   [eɪ]
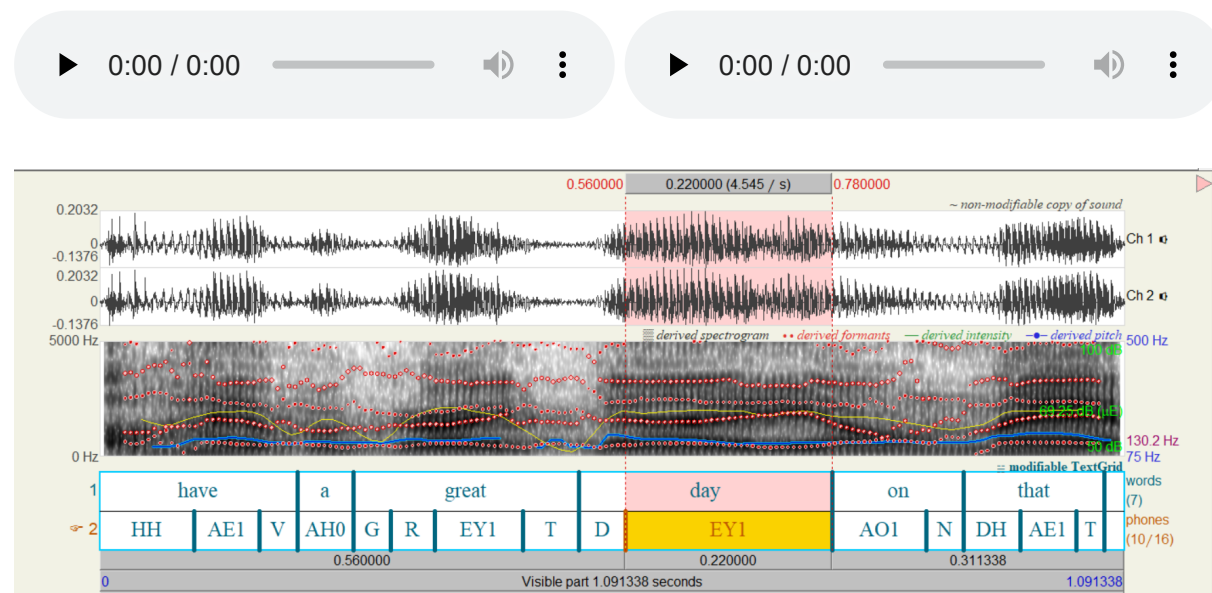
Formants: F1/F2 values for a single utterance

- 9 measurements per segment in order to get trajectory of vowel sounds

- Retain segments for which at least 5 measurements were possible

# Formants: F1/F2 values for a single location (filtered)

- 9 measurements per segment in order to get trajectory of vowel sounds

- Retain segments for which at least 5 measurements were possible

- This visualization filters out segments that do not have the typical shape of the [eɪ] diphthong

Formants: Values for a single location
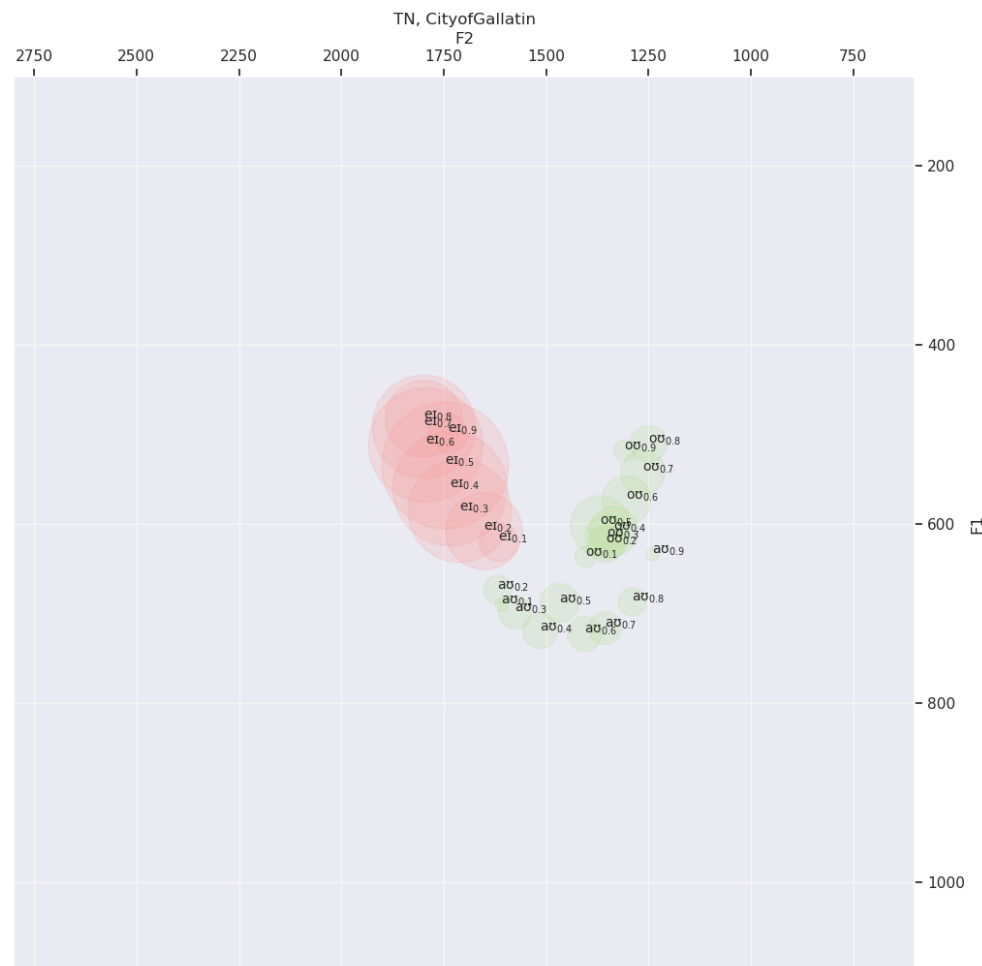
- Circle locations represent the average value for that duration quantile (subscript)
- Circle size is proportional to the number of measurements for that quantile (more likely to get formant values in the middle of the vowel than at the beginning/end)

Average F2 values for the [eɪ] nucleus, spatial autocorrelation (8,774,077 tokens)
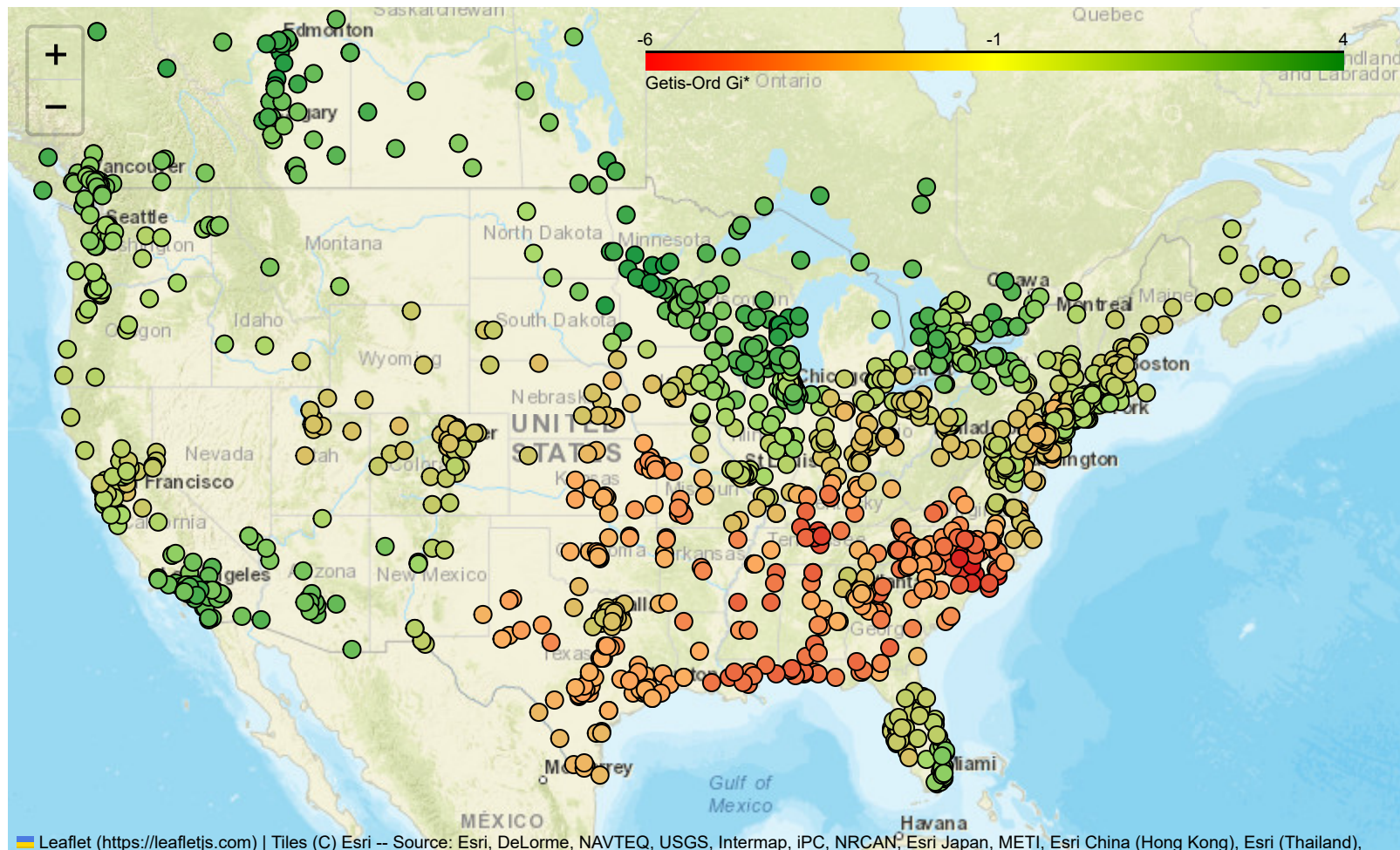
- Locations with at least 100 tokens
- Getis-Ord Gi* values based on a 20-nearest neighbor binary spatial weights matrix

Extracted *today* tokens

A selection of *today* realizations from CoANZSE videos

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
- Need to analyze error rates of forced alignment

Summary and outlook

- Large corpora of ASR transcripts from YouTube channels of local governments
- Naturalistic data, can shed new light on regional language variation
- Possibly useful for corpus studies of spoken language, dialectology, pragmatics, phonetics
- Audio download and forced alignment of larger/semi-complete CoNASE/CoANZSE sample underway
- Regional analysis of vowel formants, pitch

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.

Anderwald, L. & Wagner, S. (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.

Brezina, V., Love, R. & Aijmer, K. (2018). Corpus linguistics and sociolinguistics: Introducing the Spoken BNC2014. In V. Brezina, R. Love & K. Aijmer (Eds.), *Corpus approaches to contemporary British speech: Sociolinguistic studies of the Spoken BNC2014*, 3–9. Routledge.

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. (Forthcoming). Double modals in contemporary British and Irish Speech. *English Language and Linguistics*.

Coats, Steven. (2023a). CoANZSE: The Corpus of Australian and New Zealand Spoken English: A new resource of naturalistic speech transcripts. In Pradeesh Parameswaran, Jennifer Biggs, and David Powers (eds.), *Proceedings of the the 20th Annual Workshop of the Australasian Language Technology Association*, 1–5. Australasian Language Technology Association.

Coats, S. (2023b). Dialect corpora from YouTube. In Beatrix Busse, Nina Dumrukcic, and Ingo Kleiber (Eds.), *Language and linguistics in a complex world*, 79–102. Walter 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. CEUR.

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.

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.

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:*

References II

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

Kallen, J. & Kirk, J. (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.

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. In T. McEnery, R. Love & V. Brezina (Eds.), *Compiling and analysing the Spoken British National Corpus 2014* [= International Journal of Corpus Linguistics 22(3)], 319–44.

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.