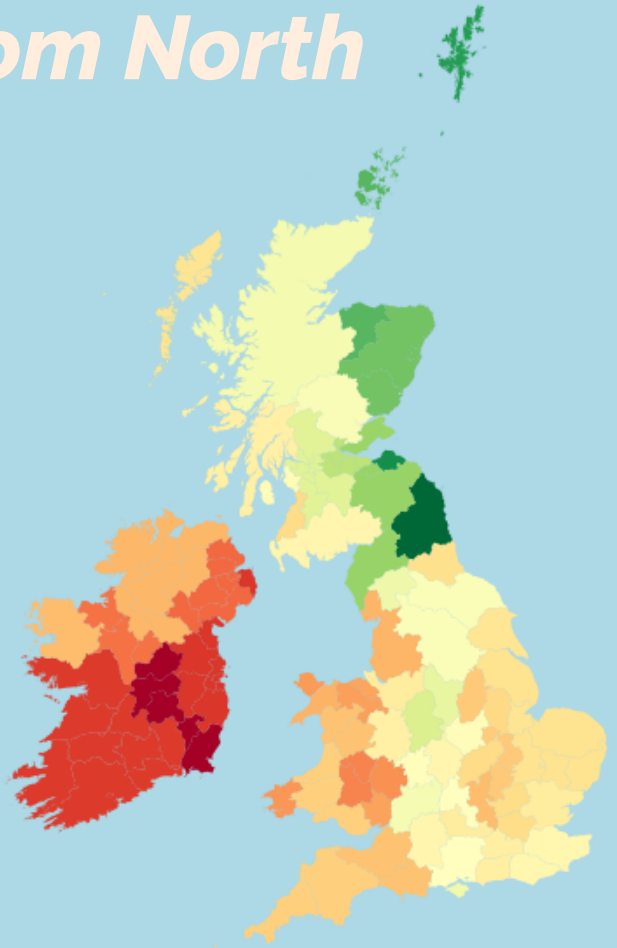
Double modals in YouTube videos from North America and the British Isles

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

Corpus-based and Computational Approaches to Variation Workshop, Helsinki
April 27th, 2022



Outline

1. CoNASE and CoBISE
2. YouTube ASR captions files, data collection and geocoding
3. Methods: Frequency analysis (frequent features), manual inspection/annotation (rare features)
4. Double modals in North America and in the British Isles
5. Caveats, summary

Slides for the presentation are on my homepage at https://cc.oulu.fi/~scoats

Introduction

- Renaissance in corpus-based study of English varieties (Nerbonne 2009; Szmrecsanyi 2011, 2013; Grieve et al. 2019)
- Available corpora of transcribed spoken English (Anderwald & Wagner 2007; Corbett 2014; Corrigan et al. 2012; Du Bois et al. 2000-2005; Kallen & Kirk 2007) are small or lack a broad geographic focus; size may make it difficult to find some features
- CoNASE: 1.25b token corpus of 301,846 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts (Coats forthcoming a)
- CoBISE: 112m token corpus of 38,680 ASR transcripts (Coats forthcoming a)
- Correspond to more than 166,000 hours of video from more than 3,000 YouTube channels of local councils and other government entities in locations in the US, Canada, England, Scotland, Wales, Northern Ireland, and the Republic of Ireland
- Freely available for research use; download from the Harvard Dataverse here and here

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 and CoBISE: target YT ASR captions

Focus on regional and local council channels

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

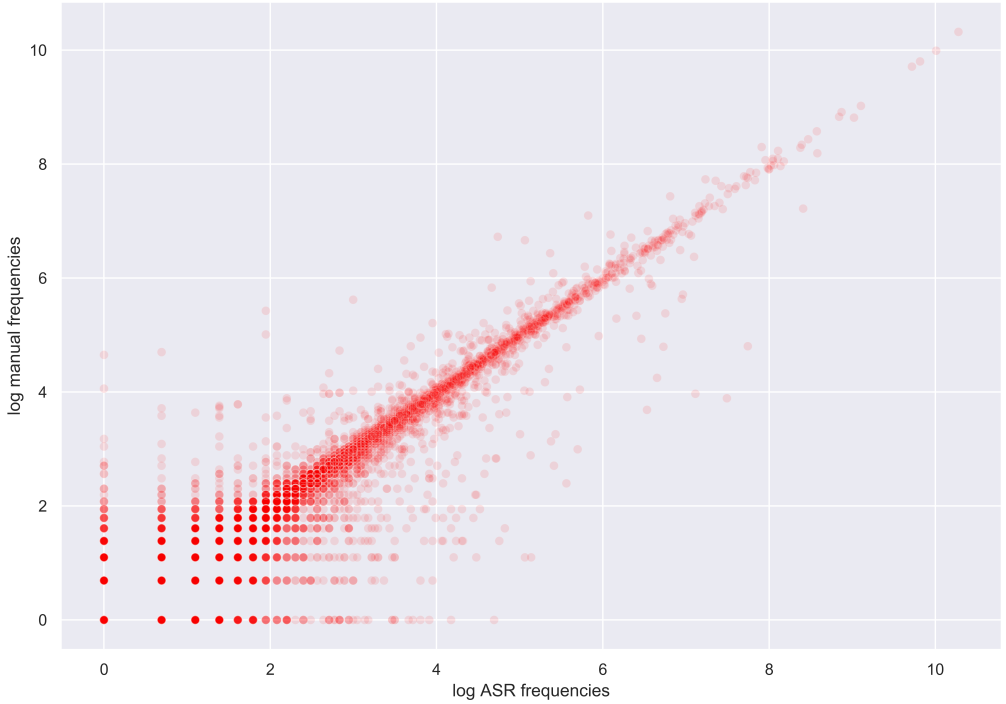
- Speaker place of residence (cf. videos collected based on place-name search alone)
- Topical contents and communicative contexts comparable

Data collection and processing

- Identification of relevant channels (YouTube API, searches of public-facing server, lists of councils with YT channels)
- Inspection of returned channels to remove false positives
- Retrieval of ASR transcripts using YouTube-DL
- VPN or Tor to circumvent IP blocking
- Geocoding: String containing council name + channel name + country location to Google's geocoding service
- PoS tagging with SpaCy (Honnibal et al. 2019)

Transcript accuracy

- ASR transcripts contain errors (WER ~22%)
- High-frequency phenomena: signal of correct transcriptions will be stronger (Agarwal et al. 2009) → classifiers
- Low-frequency phenomena: manually inspect corpus hits



Example analysis: Double modals

- Non-standard rare syntactic feature in the British Isles, North America, and elsewhere (Montgomery & Nagle 1994; Coats 2022)
- **Will you can help me with this?**
- Occurs only in the American Southeast and in Scotland/Northern Ireland/Northern England?
- Most studies based on non-naturalistic data with limited geographical scope (*LAMSAS*, *LAGS*, Murray 1873; Wright 1898-1905; Anderwald & Wagner 2007; Kallen & Kirk 2007; Smith et al. 2019)

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 regex of two-tier combinations, plus forms with intervening pronouns, auxiliary verbs, negations

```
import re
hits = []
for i,x in cobise_df.iterrows():
    pat1 = re.compile("((\\w+_\\S+_\\S+\\s){3}'+x[0]+'_\\w+_\\S+ '+x[1]+'n?_\\w+_\\S+(\\w+_\\S+_\\S+\\s){3})",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

	Country	Channel	DM	Regex_hit	link	type	notes
1	England	Fylde Council	would will	would not will not	https://youtu.be/VogFB5X_1UM?t=490	fp o1	
2	England	Fylde Council	may would	may would	https://youtu.be/6UudWIe_wYM?t=1733	t	"the wording, which it may would be better to read"
3	England	Fylde Council	could can	could he can	https://youtu.be/udvH0BtQ2ls?t=960		
4	England	Fylde Council	may would	may I would	https://youtu.be/V3YFSetBxgM?t=2184		
5	England	Fylde Council	should can	should you can	https://youtu.be/6erR7ZuYtjc?t=97		
6	England	Fylde Council	'll can	'll can	https://youtu.be/zA7LkTk0Vt4?t=1447	t	"and we'll can see how"
7	England	Fylde Council	would can	would we can 't	https://youtu.be/f1_x5C1ttCk?t=1702		
8	England	Fylde Council	should can	should you can	https://youtu.be/7AsXVW1vako?t=119	sr d	
9	England	Fylde Council	can will	can will	https://youtu.be/APNxUQP3Zok?t=1718	t	"that business rating can will neither increase nor" Scottish accent
10	England	Fylde Council	could would	could would	https://youtu.be/X3rY_QDk5kA?t=492		
11	England	Fylde Council	could can	could we can	https://youtu.be/ded7W7m7id4?t=15		
12	England	Fylde Council	would can	would can	https://youtu.be/PiSnXSuI8tQ?t=447	fp a1 o1	

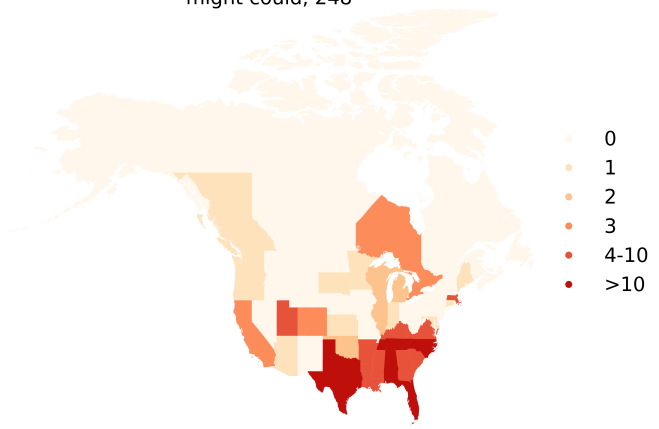
Showing 1 to 45 of 101 entries

Previous 1 2 3 Next

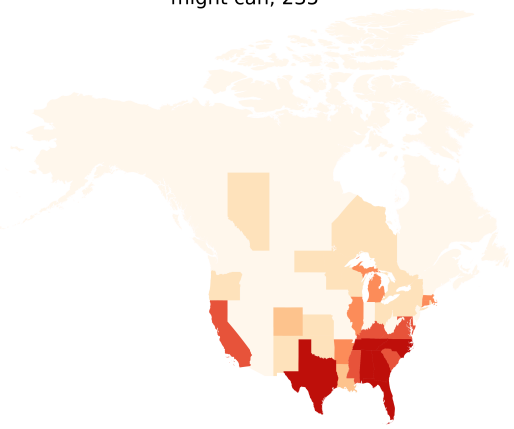
Double modals

- 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)

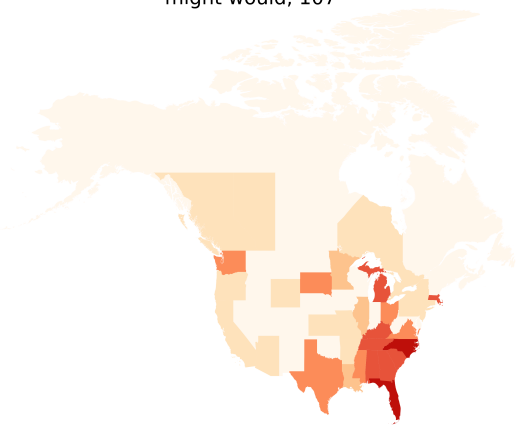
might could, 248



might can, 235

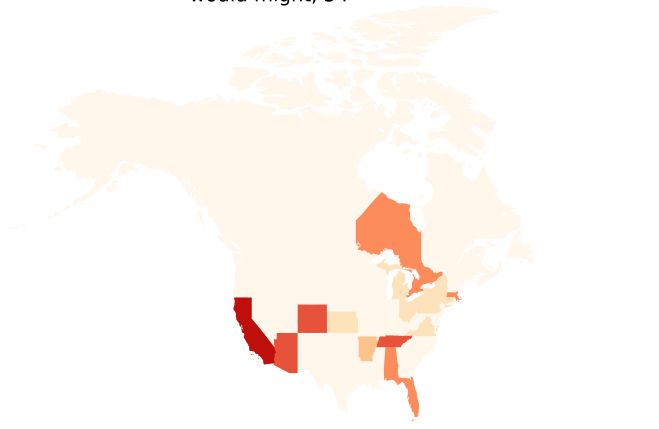


might would, 107

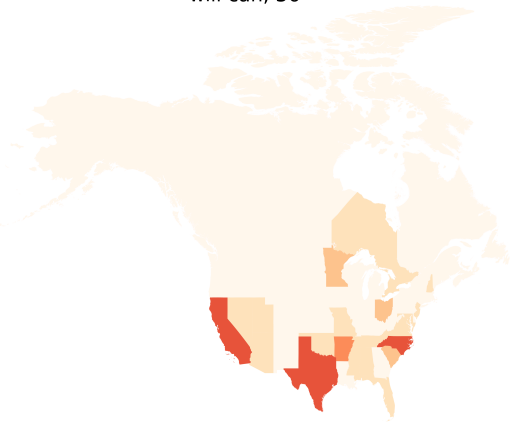


0
1
2
3
4-10
>10

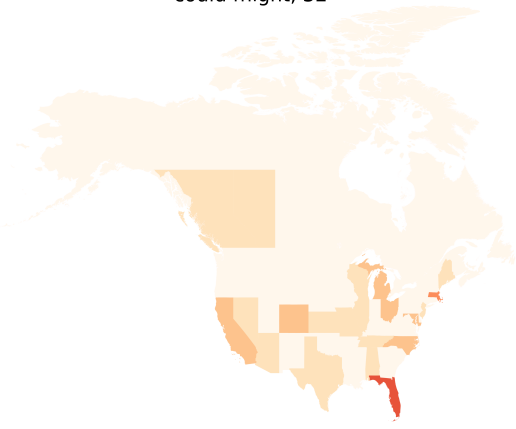
would might, 54

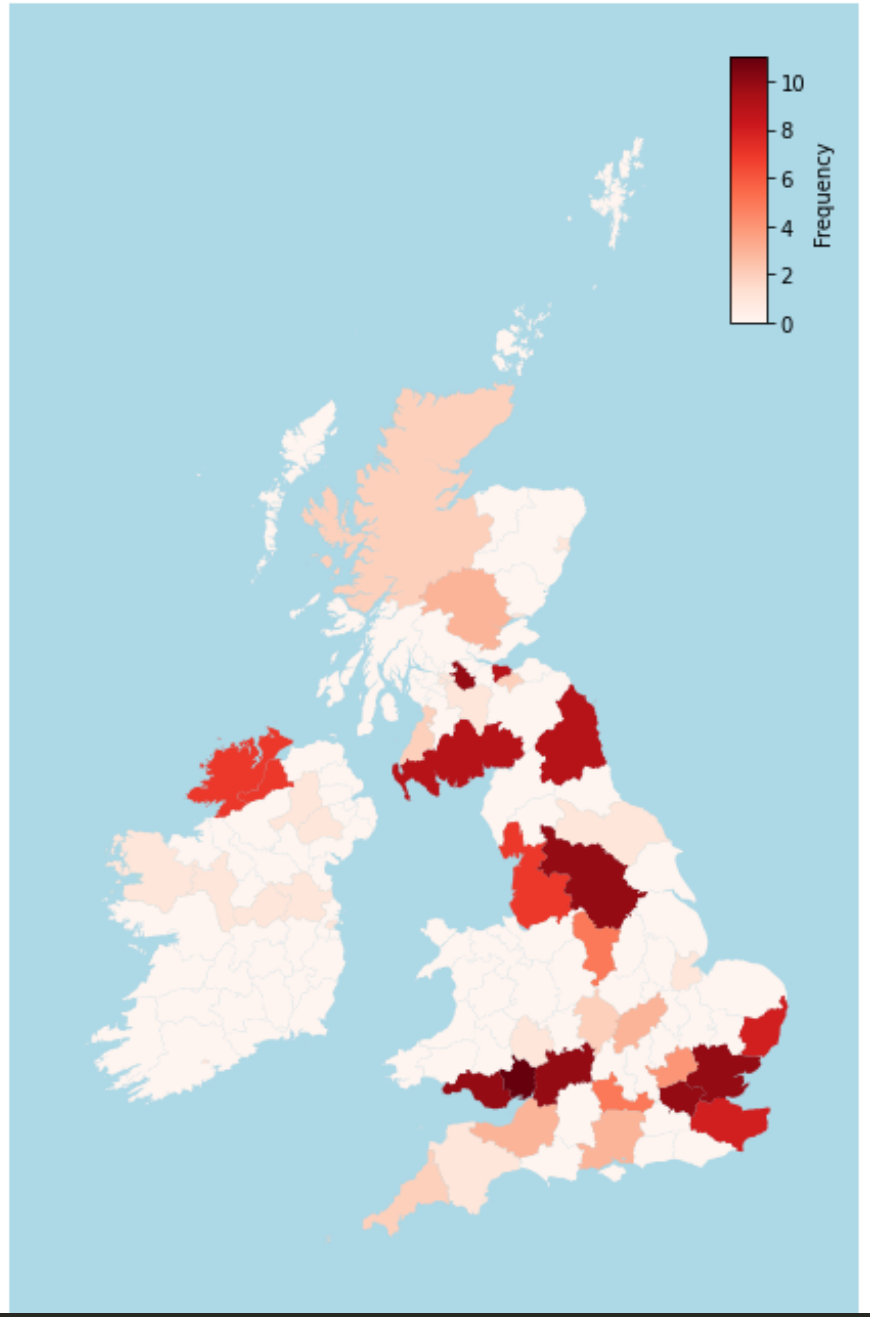


will can, 36



could might, 32




10
8
6
4
2
0
Frequency

A few caveats

- Meetings of local government not representative of speech in general
- ASR errors, quality of transcript related to quality of audio as well as dialect features (Tatman 2017; Meyer et al. 2020; Markl & Lai 2021)

Summary and outlook

- Large corpora of ASR transcripts from YouTube channels of local governments in the US, Canada, Britain, and Ireland (coming soon: Australia/NZ, 190m tokens, Germany, 56m tokens)
- Useful for corpus studies of spoken language, dialectology, pragmatics
- Double modals are more widespread than has previously been documented

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.

Coats, S. In review. Double modals in contemporary British and Irish Speech.

Coats, S. Forthcoming a. Dialect corpora from YouTube. *Proceedings of ICAME41*. De Gruyter.

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.

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.

References II

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.

Murray, J. 1873. *The dialect of the southern counties of Scotland: Its pronunciation, grammar, and historical relations.* London: Asher & Co.

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

Smith, J., D. Adger, B. Aitken, C. Heycock, E. Jamieson & G. Thoms. 2019. *The Scots Syntax Atlas*. University of Glasgow.

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.

Wright, J. 1898–1905. *The English dialect dictionary* (6 volumes). London: Henry Frowde.