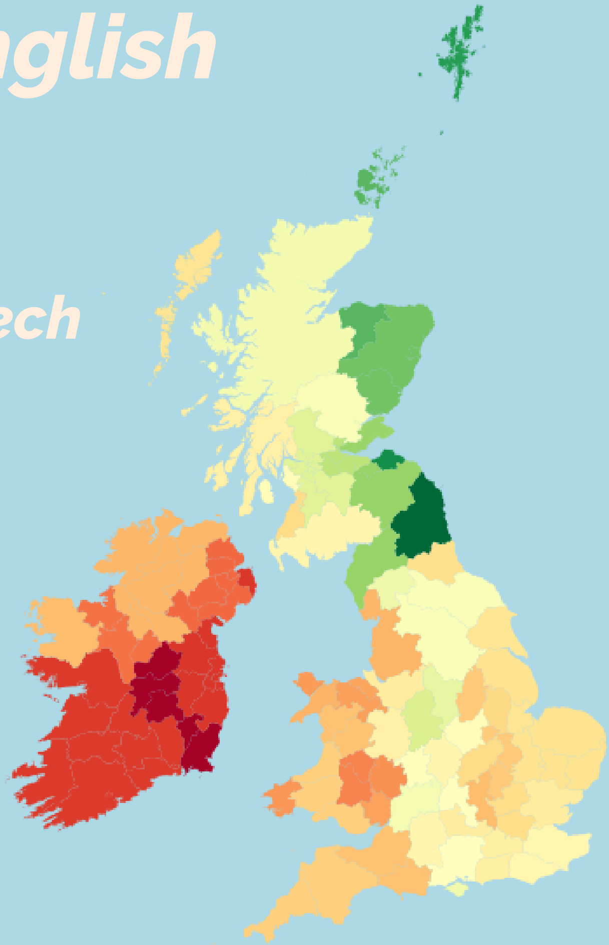
The Corpus of British Isles Spoken English (CoBISE)

A New Resource of Contemporary British and Irish Speech



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

DHNB22 Conference, Uppsala
March 17th, 2022

Outline

1. Introduction
2. Data collection and processing
3. Transcript accuracy and corpus use cases
4. Example: Manual inspection/annotation of specific features
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 British and Irish English (Anderwald & Wagner 2007; Corbett 2014; Corrigan et al. 2012; Kallen & Kirk 2007) are mostly text or focused on specific countries/regions; size may make it difficult to find some features
- CoBISE: 112m word corpus of 38,680 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts (Coats forthcoming a)
- > 12,801 hours of video from 495 YouTube channels of local councils and other government entities in 453 locations in England, Scotland, Wales, Northern Ireland, and the Republic of Ireland
- Created using procedures similar to those for CoNASE
- Freely available for research use; download from the Harvard Dataverse

Focus on regional and local council channels

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

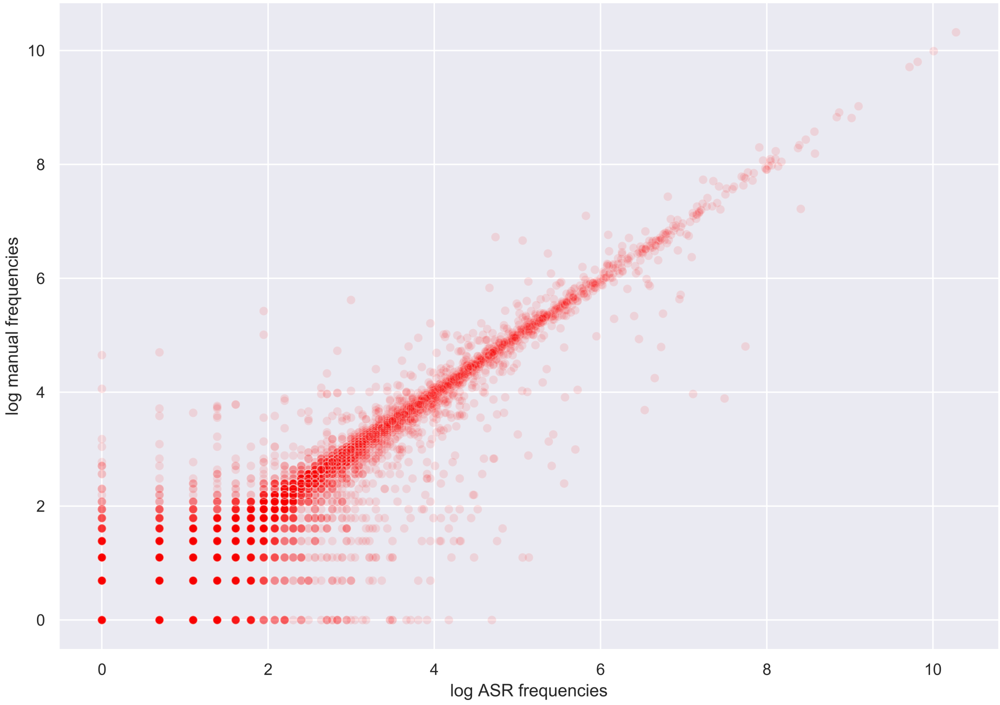
- Speaker place of residence (cf. videos collected based on place-name search alone)
- Topical contents and communicative contexts comparable

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
- Download all available ASR transcripts as .vtt files using YouTube-DL
- Use Tor to circumvent IP blocking by YT
- Remove transcripts < 50 words and those that are not ASR
- String containing council name + channel name + country location to Google's geocoding service
- Check results, correct if necessary
- PoS tagging with SpaCy (Honnibal et al. 2019)

Transcript accuracy

- ASR transcripts contain errors
- Given a minimum accuracy level, for high-frequency phenomena the signal of correct transcriptions will be stronger (Agarwal et al. 2009); for low-frequency phenomena one can manually inspect corpus hits



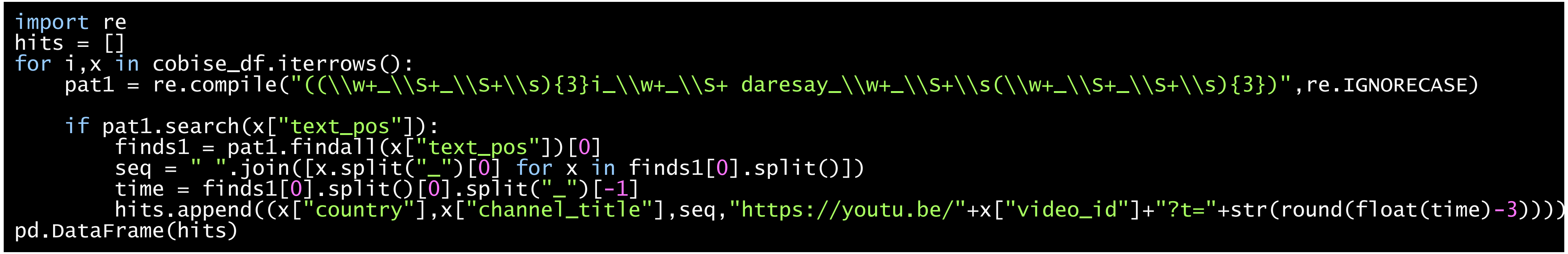
Corpus use cases and size

- Regional language (dialectology): e.g. syntax, mood and modality
- Pragmatics: Turn-taking, politeness markers
- Script pipeline: Use corpus to identify areas/speakers/words/phonemes of interest, get videos, convert to audio (FFMpeg), automated formant extraction/vowel quality analysis on a large scale

Country	Channels	Videos	Words	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

Script: Generating a table for manual inspection of 'I daresay'

- Pseudo-modal with interesting grammatical properties
- Used in spoken language, quite rare in written language (excepting dialogue)



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

Table

	Country	Channel	Regex_hit	link
1	England	Babergh Mid Suffolk District Councils	I think and I daresay you will review	https://youtu.be/9h1HBOjuiKs?t=888
2	England	Bristol City Council Live	the resources although I daresay we would all	https://youtu.be/uYlgqEbDTBQ?t=1622
3	England	Bristol City Council Live	Steve Pierce which I daresay some of you	https://youtu.be/5duBHnj3IPA?t=1777
4	England	Cambridgeshire County Council	put support now I daresay most other councillors	https://youtu.be/bDmodpBb0vU?t=5132
5	England	City of York Council	you very much I daresay there might be	https://youtu.be/6P03L8C5KS4?t=1941
6	England	City of York Council	much and then I daresay there may be	https://youtu.be/TFcWdkaQ6Js?t=356
7	England	City of York Council	their own volition I daresay that choice has	https://youtu.be/lPRHzXqJjEA?t=1005
8	England	City of York Council	months ago and I daresay might be here	https://youtu.be/SsYxaxz3opw?t=1656
9	England	City of York Council	more substantive but I daresay he may feed	https://youtu.be/6PK_GieQSro?t=2976
10	England	IWCouncil	of Wight and I daresay members of the	https://youtu.be/p3zKKB7-el8?t=448
11	England	Maidstone Council	such proposals and I daresay we will continue	https://youtu.be/x_g3mmcB_o0?t=2157
12	England	Maidstone Council	that surrounding area I daresay they would probably	https://youtu.be/b5g0mqKdH_U?t=3178
13	England	Maidstone Council	the night so I daresay we ought not	https://youtu.be/nfEWdS6tEUE?t=1949

Showing 1 to 42 of 42 entries

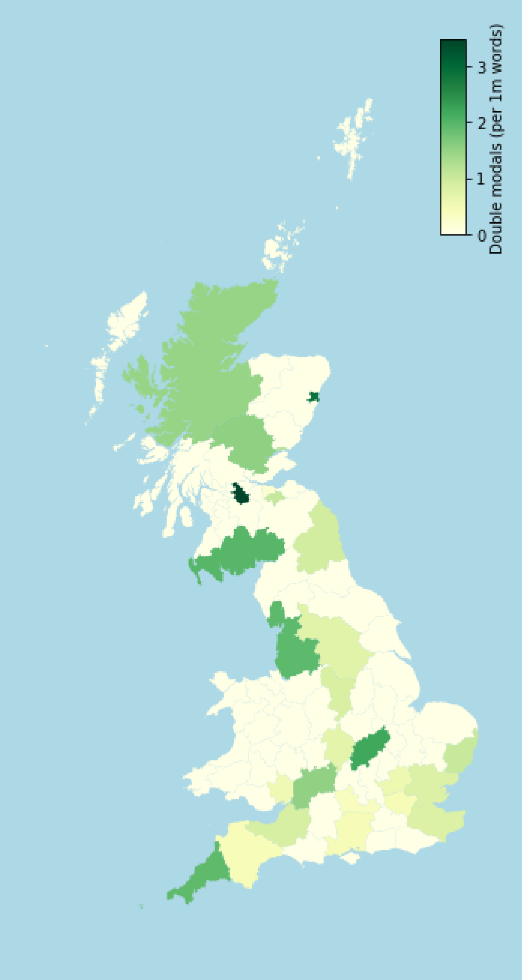
Previous 1 Next

Example analysis: Double modals

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

Double modals

- Regular-expression-search and manual annotation approach
- Double modals can be found in in Scotland, N. Ireland, and N. England, but also in the English Midlands and South and in Wales (Coats in Review)



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 corpus of automatic speech-to-text transcripts from YouTube channels of local governments in Britain and Ireland
- Useful for corpus studies of spoken language, dialectology, pragmatics
- Freely available!

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. Forthcoming b. Naturalistic double modals in North America. *American Speech*.

Coats, S. 2019. A corpus of regional American language from YouTube. In: C. Navarretta, M. Agirrezabal & B. Maegaard (Eds.), *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference, Copenhagen, Denmark, March 6–8, 2019*, 79–91. CEUR-WS.

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*.

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.