

The Corpus of British Isles Spoken English (CoBISE)

A New Resource of Contemporary British and Irish Speech

Steven Coats¹

¹*University of Oulu, English, Faculty of Humanities, 90100 Oulu, Finland*

Abstract

Corpora of transcribed regional speech are important for the study of dialects of English, but relatively few large corpora of transcribed naturalistic speech from the United Kingdom and Ireland exist. This paper presents the The Corpus of British Isles Spoken English (CoBISE), 112-million-word corpus of Automatic Speech Recognition (ASR) transcripts of YouTube videos from channels of councils and other government entities in the UK and Ireland. Transcripts are linked to publicly-available videos, so the corpus can also serve as a starting point for the study of multimodal phenomena. The paper describes the methods used for identifying relevant channels and the scripting pipeline for data collection and processing. Because ASR transcripts contain errors, analyses undertaken using the corpus should employ methods suitable for dealing with “noisy data”. Two possible approaches are described: for frequent phenomena, appropriate feature selection and use of robust classification models, and for rare phenomena, manual inspection of the audio/video data.

Keywords

corpus linguistics, spoken language, dialectology, British English, Irish English, Scottish English, Welsh English, YouTube

1. Introduction


New methodological approaches [1, 2, 3] and new sources of data have invigorated the study of regional language variation in the British Isles in recent years, with data from spoken language corpora [4] and social media [5] providing new insights into local, regional, and national patterns of lexical and grammatical variation in UK Englishes. Despite this, existing resources may be insufficient for capturing contemporary spoken language variation from a broad geographic perspective: many are either focused on local or national varieties (e.g. the NECTE/DECTE corpora for Newcastle and the Tyneside [6], the Irish component of the International Corpus of English [7], or the Scottish Corpus of Texts and Speech [8]), lack sufficient geographical granularity for the reliable identification of regional or local dialect features [9], or are not large enough to capture the range of syntactic variation in contemporary speech. This paper introduces a new resource: the Corpus of British Isles Spoken English (CoBISE, <https://cc.oulu.fi/~scoats/CoBISE.html>), a 112-million-word corpus of 38,680 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts, corresponding


The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022

✉ steven.coats@oulu.fi (S. Coats)

🌐 <https://cc.oulu.fi/~scoats> (S. Coats)

🆔 0000-0002-7295-3893 (S. Coats)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

to more than 12,801 hours of video, from 494 YouTube channels of local councils or other institutions of local governance in 453 locations in England, Scotland, Wales, Northern Ireland, and the Republic of Ireland. Similar to the Corpus of North American Spoken English,¹ many of the transcripts are records of public council meetings (see also [10]).

This paper summarizes the methods used for data collection, processing, and geolocation of the channels sampled in the corpus. Because CoBISE consists of ASR transcripts, it is “noisy” data, containing errors. Nevertheless, due to its size and the preponderance of accurately transcribed forms, it can be used to extract reliable linguistic signals for a wide range of relatively frequent phenomena. Because the transcripts are from videos viewable by anyone with access to the internet, phenomena of interest can also be examined and manually verified in the corresponding videos—the paper provides an example of how this can be done for a low-frequency feature. Finally, the structure of the corpus facilitates the creation of corpora of video or audio data with a simple pipeline of download and conversion scripts, opening up the possibility for semi-automated analysis of (for example) acoustic or visual aspects of communication. While the resource has been created for the study of linguistic and communicative phenomena, it may also be of interest as a source of data for text-mining based studies within the broader context of digital humanities and social sciences, for example in disciplines such as political science, sociology, media studies, or cultural studies.

2. Data Collection

The data collection process for CoBISE consisted of a three-step procedure. First, relevant channels were identified (YouTube channels of local government entities). Next, identifier metadata and transcripts were accessed through YouTube’s public-facing server. Finally, downloaded transcripts were filtered and processed (removal of non-relevant material, geocoding, conversion of .vtt transcripts, PoS tagging), mostly using procedures already described [11, 12].

Channels were identified by sending search queries for the names of 413 sub-regional administrative areas (generally counties or equivalent administrative bodies) in the UK and Ireland to YouTube’s search page; additional channels were identified from online lists of local government authorities maintained by the UK and Ireland governments. Results were manually checked to remove non-government channels or non-UK/Ireland channels from places with the same names (e.g. Boston in Massachusetts, USA instead of Norfolk, UK or Ipswich in Queensland, Australia, instead of Suffolk, UK).

Transcripts were collected with scripts based on the open-source program YouTube-DL in Python, routed through the Tor service to circumvent IP restrictions.² A script removed non-ASR or automatically-translated non-English transcripts and those with fewer than 50 words. Geocoding of channel locations was undertaken by sending the channel name and country location to Google’s geocoding API;³ results were manually checked and corrected if necessary. Part-of-speech tagging with the Penn Treebank tagset was undertaken with spaCy [13].⁴ Tokens

¹<https://cc.oulu.fi/~scoats/CoNASE.html>.

²<https://github.com/ytdl-org/youtube-dl/>, <https://www.torproject.org/>.

³<https://developers.google.com/maps/documentation/geocoding/overview>.

⁴Some corpus creation scripts are available at <https://github.com/stcoats>.

in the corpus have the format `token_POS_10.0`, where `token` is the transcribed lexical item, `POS` the part-of-speech tag, and `10.0` the time offset from the start of the corresponding video. The corpus is structured as a table in which each transcript is assigned a single row; columns indicate country, the name of the channel from which the transcript was downloaded, the id code of that channel, the title of the video, the video's id code, the length of the video in seconds, the street address of the authority that is responsible for the channel, the number of words in the transcript, the PoS-tagged and timed text of the transcript, and the latitude-longitude coordinates of the channel location. The publicly available version of the corpus⁵ has been additionally altered in order to comply with Fair Use provisions of copyright law: every 200 tokens, 10 words have been removed and replaced with the @ symbol. Table 1 shows the size of the corpus by country location as number of sampled channels and videos, number of word tokens, and aggregate length in hours of the videos for which transcripts were downloaded.

Table 1
Corpus Size by Country Location

Country Location	Channels	Videos	Words	Length (h)
England	358	23,630	72,854,319	8,518.39
Northern Ireland	11	1,925	6,533,359	774.17
Republic of Ireland	28	2,525	6,264,276	680.81
Scotland	77	8,112	17,094,334	1,843.38
Wales	20	2,465	8,800,264	982.66

3. Transcript Accuracy and Corpus Use

ASR transcripts of naturalistic speech are inaccurate, with recent systems showing word error rates (WER) in the range of 0.2-0.5 for naturalistic conversational speech [14, 15]. Many factors can affect WER: audio recording quality, speech fluency or lack thereof, use of out-of-vocabulary words such as proper nouns, technical terms, slang, or dialect words, as well as properties of the speech signal related to individual characteristics, including regional accent, speech rate, pitch, and other prosodic features [16]. Calculation of the WER for CoBISE has not been undertaken, as it would require a large sample of ground-truth (manually prepared) transcripts, but can be estimated based on the average WER of 0.22 found for a sample of transcripts from Philadelphia, USA, from the CoNASE corpus [12]; a semi-manual analysis using data from CoBISE found that 27.6% of 1,154 manually-examined search hits contained an ASR error [10]. Accuracy rates for ASR can be lower for regional varieties of English such as Scottish English/Scots or Indian English, compared to Southern UK or American English [17, 18, 19], if models have been trained using data from Southern UK and Standard American speakers.

Noisy data such as ASR transcripts can nevertheless be used to draw accurate inferences about lexical, grammatical, and pragmatic feature use in naturalistic conversation, given sufficient sample sizes. Agarwal et al. [20], for example, found that noisy data such as randomly introduced spelling errors or inaccurate ASR transcripts do not significantly affect text classification tasks

⁵<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UGIIWD>.

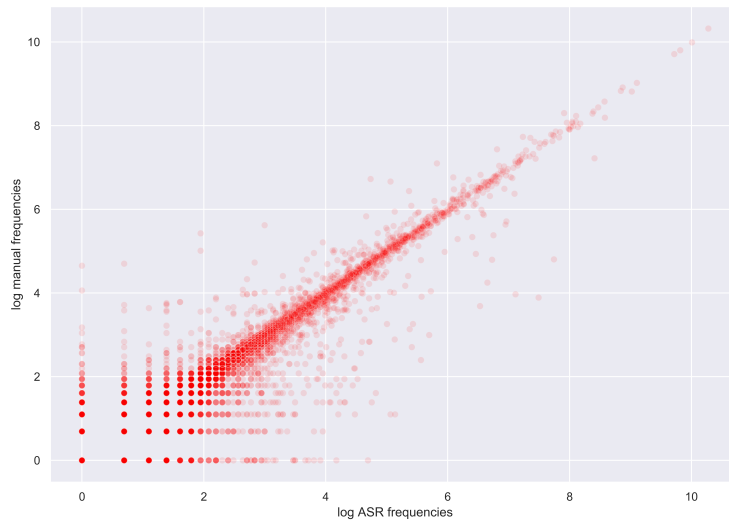


Figure 1: Log word frequency (ASR) vs. Log word frequency (manual transcripts), 14,433 word types from 41 Philadelphia/USA council meeting transcripts

using bag-of-words approaches, even when up to 70% of the words in training texts contain errors, due to the fact that for a given word, as long as the proportion of correct transcriptions is higher, the signal will be stronger in the data.

For common lexical types, frequencies in large ASR transcript corpora are unlikely to deviate significantly from those in corpora comprised of manual transcripts of the same recordings. Figure 1 shows, for 42 Philadelphia City Council meetings for which both ASR and manual transcripts were obtained, the logarithm of frequency for the 14,433 word types that occur at least once in both transcript types (see [12] for details). 96.5% of word types have frequencies that are not significantly different at $\alpha = .05$, according to a log-likelihood test. CoBISE data is likely to exhibit a similar pattern and therefore may prove useful “out of the box” for large-scale descriptive analyses in which common lexical items or relatively frequent grammatical constructions are considered. Given the robustness of noisy ASR data for classification tasks demonstrated by Agarwal et al., it may also be possible to use CoBISE data in predictive models that employ machine learning algorithms such as linear support vector machines [21].

For infrequent phenomena and/or analyses in which precision is required, manual annotation can be used to verify transcript texts. The design of the corpus makes it possible to link every instance of a particular utterance to the URLs of the corresponding videos at the moment of utterance, allowing the analyst to check the accuracy of transcripts and to mark up utterances with speaker or contextual features that may be of interest.

Figure 2 schematically illustrates the procedure for creating a table with search hits for *I daresay*: A regular expression is used to search the corpus and generate a table showing the locations, channels, search hits, and links to the videos at the times of utterance. The analyst

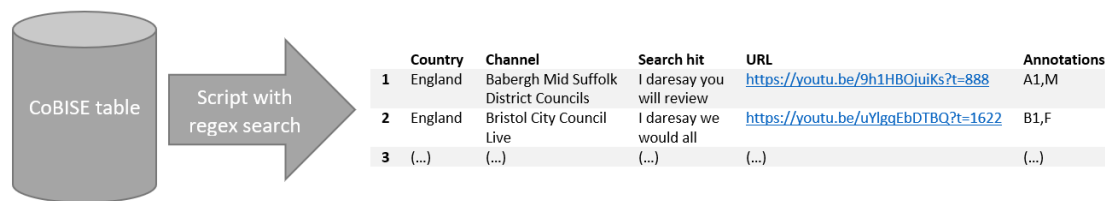


Figure 2: Procedure for Manual Search and Annotation for *I daresay*

can then sequentially listen to the utterances, adding annotations that indicate (for example) ASR errors or contextual features such as conversation type, apparent speaker gender, or other categories that may be relevant for an ensuing analysis.

4. Example Analysis

This method has been used in [22] and in [10] to verify naturalistic usages of double modals, a rare non-standard syntactic feature of some regional varieties of spoken English in the British Isles, North America, and elsewhere [23, 24].⁶ Because double modals are mainly absent from text corpora and quite rare in speech, even in varieties in which they are known to occur, knowledge of the geographical extent of the feature has been based on limited data, and in the British Isles, the feature has been thought to occur exclusively in Scotland, Northern Ireland, and Northern England. Using the regular expression search and manual annotation approach described above, however, showed that double modals can be found in naturalistic speech from throughout the UK and Ireland. Figure 3 shows that in Britain, the relative frequency of double modals is somewhat higher in the North of England and Scotland, but the feature also occurs in speech from the English Midlands and South and from Wales.

5. Conclusion and Summary

CoBISE, a large corpus of naturalistic speech created from ASR transcripts of videos uploaded by councils and other government entities in the UK and Ireland, may be useful for research in dialectology, sociolinguistics, phonetics, or pragmatics, as well as digital humanities and social sciences. Despite ASR errors in the transcripts, frequent lexical items, collocations, or lexical bundles leave a reliable signal in the corpus, and manual verification and annotation methods can be used to investigate rare lexical, discourse, or syntactic features, such as double modals. Like the related CoNASE corpus, data in CoBISE is linked to publicly-available videos from which the audio and video signals can easily be extracted, opening up new opportunities for corpus-based studies of acoustic or visual properties of speech and interaction. A further possibility for CoBISE data would be to investigate pragmatic or discourse phenomena such as turn-taking, markers of politeness, expressions of consternation, or self-repairs. In coming years, the accuracy of ASR algorithms will likely continue to improve, and more and more

⁶Use of two modal auxiliary verbs within a single verbal phrase, for example *Will you can help me with this?*

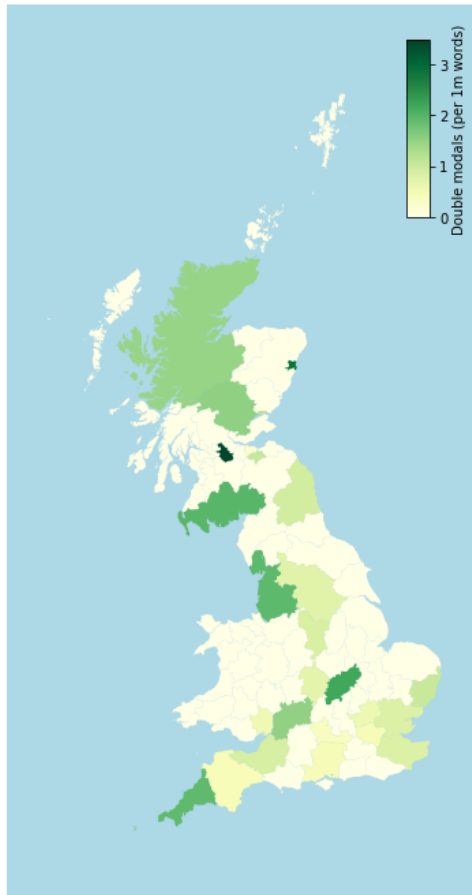


Figure 3: Relative frequency of double modals per million words in CoBISE data

speech data will become available for researchers interested in regional variation in speech. CoBISE, as a curated collection of ASR transcripts, represents an early stage in this development, and it is hoped that the resource will offer researchers in linguistics and interaction studies useful material for the investigation of naturalistic speech in the British Isles.

Acknowledgments

Thanks are due to Finland's Centre for Scientific Computing (<https://csc.fi>) for providing access to computing and storage resources.

References

- [1] J. Nerbonne, Data-driven dialectology, *Language and Linguistics Compass* 3 (2009) 175–198.

- [2] B. Szmrecsanyi, Corpus-based dialectometry: A methodological sketch, *Corpora* 6 (2011) 45–76.
- [3] B. Szmrecsanyi, *Grammatical variation in British English dialects: A study in corpus-based dialectometry*, Cambridge University Press, Cambridge, UK, 2013.
- [4] L. Anderwald, S. Wagner, 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*, Palgrave Macmillan, Houndmills, Basingstoke, 2007, pp. 35–53.
- [5] J. Grieve, C. Montgomery, A. Nini, A. Murakami, D. Guo, Mapping lexical dialect variation in British English using Twitter, *Frontiers in Artificial Intelligence* 2 (2019). doi:10.3389/frai.2019.00011.
- [6] K. P. Corrigan, I. Buchstaller, A. Mearns, H. Moisl, *The Diachronic Electronic Corpus of Tyneside English*, 2012. URL: <https://research.ncl.ac.uk/decte>.
- [7] J. Kallen, J. Kirk, 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*, Palgrave Macmillan, Houndmills, Basingstoke, 2007, pp. 121–162.
- [8] J. Corbett, Syntactic variation: Evidence from the Scottish Corpus of Text and Speech, in: R. Lawson (Ed.), *Sociolinguistics in Scotland*, Palgrave Macmillan, Houndmills, Basingstoke, 2014, pp. 258–276.
- [9] V. Brezina, R. Love, K. Aijmer, *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*, Routledge, New York, 2018, pp. 3–9.
- [10] S. Coats, *Double Modals in contemporary British and Irish Speech* (In review).
- [11] S. Coats, 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, DHN '19, CEUR-WS, Aachen, Germany, 2019*, pp. 79–91. URL: http://ceur-ws.org/Vol-2364/7_paper.pdf.
- [12] S. Coats, *Dialect corpora from YouTube*, in: *Proceedings of ICAME41, De Gruyter, Forthcoming*.
- [13] M. Honnibal, 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, *Explosion/spaCy v2.1.7: Improved evaluation, better language factories and bug fixes*, 2019. doi:10.5281/zenodo.3358113.
- [14] J. Y. Kim, C. Liu, R. A. Calvo, K. McCabe, S. C. R. Taylor, B. W. Schuller, K. Wu, *A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech*, 2019. arXiv:1904.12403.
- [15] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, S. Goel, *Racial disparities in automated speech recognition*, *Proceedings of the National Academy of Sciences* 117 (2020) 7684–7689. doi:10.1073/pnas.1915768117.
- [16] A. Aksënova, D. van Esch, J. Flynn, P. Golik, *How might we create better benchmarks for speech recognition?*, in: *Proceedings of the 1st Workshop on Benchmarking: Past,*

- Present and Future, Association for Computational Linguistics, Online, 2021, pp. 22–34. doi:10.18653/v1/2021.bppf-1.4.
- [17] R. Tatman, Gender and dialect bias in YouTube’s automatic captions, in: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 53–59. doi:10.18653/v1/W17-1606.
- [18] N. Markl, C. Lai, 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, Online, 2021, pp. 34–40. URL: <https://aclanthology.org/2021.hcinlp-1.6>.
- [19] J. Meyer, L. Rauchenstein, J. D. Eisenberg, N. Howell, 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, pp. 6462–6468. URL: <https://aclanthology.org/2020.lrec-1.796>.
- [20] S. Agarwal, S. Godbole, D. Punjani, S. Roy, How much noise is too much: A study in automatic text classification, in: Seventh IEEE International Conference on Data Mining (ICDM 2007), 2007, pp. 3–12. doi:10.1109/ICDM.2007.21.
- [21] V. Laippala, J. Egbert, D. Biber, A.-J. Kyröläinen, Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents, *Language Resources and Evaluation* 55 (2021) 757–788. doi:10.1007/s10579-020-09519-z.
- [22] S. Coats, Naturalistic double modals in North America, *American Speech* (2022). doi:10.1215/00031283-9766889.
- [23] B. A. Fennell, R. R. Butters, Historical and contemporary distribution of double modals in english, in: E. W. Schneider (Ed.), *Focus on the USA: Varieties of English around the world*, John Benjamins, Amsterdam, 1996, pp. 265–288.
- [24] M. B. Montgomery, S. J. Nagle, Double modals in Scotland and the Southern United States: Trans-atlantic inheritance or independent development?, *Folia Linguistica Historica* 14 (1994) 91–108.