Steven Coats

Dialect Corpora from YouTube

Abstract: This paper introduces two new large corpora comprised of YouTube Automatic Speech Recognition (ASR) transcripts of the speech of videos from geographically localized channels in the United States, Canada, and the British Isles, a promising resource for more in-depth study of regional language variation in spoken English. The procedure used to create the corpora bypasses the web API for YouTube, instead relying on web scraping and open-source scripts or software for the automatic identification and downloading of suitable channel content as well as dealing with the rate-limiting issues that arise thereby. In order to assess the accuracy of downloaded transcripts, word frequency statistics are compared for ASR and manual transcripts of city council meetings of Philadelphia, Pennsylvania, USA, and a transcript classification task is undertaken using vector-based distributed representations of transcript content. Despite errors, corpora of ASR transcripts may prove useful for the characterization and study of regional language variation, particularly when analytical techniques are employed that are relatively robust to low-frequency phenomena.

1 Introduction

Large corpora of geographically localized speech transcripts are an important resource for the analysis of regional variation in English (Szmrecsanyi 2011), but despite the appearance of new corpora in recent years and the proliferation of corpus-based methods for linguistic analysis, particularly in the UK (Busse 2018), relatively few corpora of regionally-located speech exist for North America or the British Isles. Considering the time and resources required for manual transcription of audio and video data, advances in Automatic Speech Recognition (ASR) present opportunities for the creation of corpora of orthographic transcripts that may be useful for corpus linguistic-based research into variation in spoken language. Corpus creation from ASR transcripts, however, raises new methodological issues pertaining to data access and to transcript accuracy. Obtaining ASR transcripts, for example from YouTube, in volumes sufficient for the creation of a

Acknowledgements: Thanks to two anonymous reviewers for suggestions to a draft version of the manuscript and to Finland's Centre for Scientific Computing for access to computing resources.

https://doi.org/10.1515/9783111017433-005

geographically representative corpus may present difficulties: Access to data via YouTube's web API (Application Programming Interface) is by default limited, and web scraping can result in IP blocking, limiting the researcher's ability to access data.

Although ASR algorithms can achieve accuracy levels comparable to those of human transcribers for recordings with high acoustic fidelity or for specific transcription tasks (Chiu et al. 2018; Xiong et al. 2017), and ASR transcripts may be accurate enough for certain types of transcript-based analysis (Ziman et al. 2018), the accuracy of ASR transcripts of naturalistic speech is typically lower, and has been judged to be insufficient for some corpus creation projects. McEnery, for example, discussing the methods used for the creation of the spoken portion of the BNC2014 corpus, found ASR to be "not at all helpful" (2018: 11); the project instead utilized a team to manually transcribe audio data recorded on mobile telephones.

Nevertheless, not all research projects will have the time and resources necessary for large-scale manual transcription. While a corpus of ASR transcripts, which typically contain a certain amount of "noise" (i.e., textual errors), may be unsuitable for analyses of (for example) rare lexical items, it may, given sufficient size, still be useful for a range of linguistic analyses, including a broad range of language processing tasks that can support such analyses, for example topic modelling, content summarization, or word-vector-based approaches. The usefulness of noisy transcripts for such tasks is a result of the law of large numbers: For a given feature, if a sufficient proportion of transcriptions are accurate, the resulting signal in a corpus will be strong enough to make reliable predictions, despite the existence of inaccurate transcriptions of that feature.

Starting from the premise that ASR transcripts will indeed be useful for a variety of analyses of regional English in North America and the British Isles, despite inaccuracies, this paper is organized as follows: First, an overview of some previous work on ASR transcripts is provided. Then, the procedure used for the creation of corpora of geographically localized ASR transcripts from YouTube is presented; two corpora (one for the United States and Canada and one for the United Kingdom and Ireland) are described. In Section 4, two preliminary analyses are conducted: ASR transcripts for a subset of the material (40 transcripts totalling ~500,000 words) are compared to manual transcripts of the same videos in terms of word error rate (WER). Then, word embeddings are used to create a language model from a subset of the North American corpus; word vectors are used to predict the regional provenance of unknown speech transcripts from California or New York and to visualize state-level similarity in lexis. The results are discussed and possible directions for future work are presented in the final sections.

2 Previous Research

The accuracy of ASR transcripts has increased in recent years due to the use of sophisticated machine learning models and large amounts of training data (Chiu et al. 2018; Halpern et al. 2016; Liao/McDermott/Senior 2013; Sainath et al. 2015; Xiong et al. 2017). Ziman et al. (2018) found that Google's speech-to-text service offers high accuracy in terms of word identification and timing. An ASR-based system used to create transcripts of sessions of the Japanese parliament is reported to have accuracy of up to 95% (Kawahara 2012). Ranchal et al. (2013) analysed the use of automatic captioning with IBM's ViaScribe and Hosted Transcription Service for 19 hours of university lectures, finding that error rates ranged from 45%, for spontaneous real-time transcription of speech using an untrained model, to 9.1%, when input parameters of the acoustic signal were carefully prepared and the speech model trained in advance with acoustic data from a specific lecturer. Tatman (2017) found YouTube English ASR captions to be generally accurate, but that accuracy can also depend on speaker gender and dialect.

Bokhove/Downey (2018) discussed the advantages of using ASR transcripts in research requiring speech transcripts in terms of time and expenditure, compared to manual transcriptions. They analysed the automatic transcripts created by YouTube for three videos: a one-to-one interview of a lecturer at an English university with high audio fidelity, a video of a mathematics lesson for 8th-graders at an American school, and a video of a UK parliamentary inquiry interview with a British Army officer. They found textual similarity rates between 64% and 92% for the YouTube ASR transcripts and manual transcripts.[1] Këpuska/Bohouta (2017) found that Google Cloud's speech-to-text system outperformed Microsoft's ASR service and a system created at Carnegie-Mellon University in terms of WER. Kim et al. (2019) evaluated the performance of several ASR transcription services by calculating WERs for transcripts of medical conversations with Australian medical school students. They found WERs between 0.28 and 0.55, with YouTube showing the lowest rates.

In natural language processing, 'noisy' text has been shown to be useful for a number of analytical tasks. Agarwal et al. (2007) conducted an experiment in which machine learning was used to automatically classify collections of texts using the "bag of words" approach (i.e., on the basis of word frequencies,

1 The method used to measure accuracy was unorthodox: ASR and manual transcripts were compared using the similarity score of the commercial plagiarism detection software *Turnitin*, rather than standard measures such as WER.

but not considering word order). Tests were undertaken in classifier performance after increasing levels of random noise (i.e., spelling errors) had been introduced into the text data. The authors found that the performance of naïve Bayes and Support Vector Machine classifiers remained relatively stable even when noisy data, with errors in 40% of the words, was utilized. Similarly, Eder working with texts in English, German, Polish, Latin and Ancient Greek, found that textual error rates of up to 20% do not significantly affect the results of an authorship attribution task.

Franzini et al. (2018), applying authorship attribution to a corpus of correspondence between Jacob and Wilhelm Grimm, found that error-containing OCR (optical character recognition)-generated texts can serve as a reliable proxy for more accurate manually-keyboarded texts. Pentland et al. (2019) reported on a project that investigates the relationship between ASR transcript accuracy and text classification model performance using transcripts of company earnings call audio files and ASR transcripts of the audio. They reported a relatively high WER of 34% for the ASR transcripts. In follow-up work, they found that when used to train a machine-learning model, manual transcripts and ASR transcripts do not differ substantially in model performance, even for ASR transcripts with relatively high WER values (S. Pentland, pers. comm. of paper under review, 17 November 2020).

Coats (2019) described a method for the creation of corpora from ASR transcripts of local government and community organization channels by using a script to send multiple search terms to YouTube's API, then downloading channel content using open-source tools. Word timings from this data were used to investigate regional variation in speech articulation rate in spoken American English in Coats (2020).

3 YouTube, Data, Channel Identification, and Data Collection

YouTube transcripts are available for download through the site's API or through URLs that are generated automatically when a user accesses a video on the platform's website. The API is a convenient means of accessing transcript (and other) data, but may not be suitable for the creation of larger corpora due to access and rate limitations. Accessing transcripts through a URL and downloading them with the open-source YouTube-DL software (Yen/Remite/Sergey 2020) is an alternative.

3.1 YouTube ASR Transcripts and API

YouTube makes video content and metadata, including speech-to-text captions, available for download through an API (Google Developers 2021). Access to API content is limited by a system that assigns a "quota cost" to each HTTP request sent to YouTube's servers: For example, listing the various types of metadata associated with a specific video or channel has a quota cost of 3, conducting a search of all YouTube content a cost of 100, and downloading a specific transcript a cost of 200 quota points. In the spring of 2019, YouTube reduced the daily default quota for API access to 10,000 quota units (1% of the volume previously available), making the collection of a large number of transcripts via the API less feasible (cf. Coats 2019). Because YouTube content, including transcripts, are stored at publicly available URLs, however, they can be scraped directly from web pages, rather than collected via the API. A web-scraping method, utilizing Python scripts and libraries, was used to collect transcript data in order to create the corpora described below.

3.2 Channel Identification and Data Collection

Two scaping-based approaches were adopted for data collection by using preexisting lists of websites. In the first approach, a large list of local government entities from the U.S. Census Bureau was scraped for websites; these websites were then scraped for links to YouTube channels. In a second approach, an automated browser script sent lists of search terms to YouTube's public web interface (rather than the API). Both of these methods made use of the browser automation tool *Selenium* in Python (Muthukadan 2018).

3.2.1 United States

For the United States, a list of 35,924 websites was extracted from a comprehensive listing of 91,386 local government entities provided by the U.S. Census Bureau (2017). These websites, mostly homepages of cities, towns, school boards, public utility districts, or other administrative entities, were then scraped for links to YouTube channels, resulting in 2,534 channels. After removal of false positives,[2] all

2 Some local government websites are built from templates which include icons that can link to social media such as Facebook, Twitter, and YouTube. If the default templates are not

available English-language ASR transcripts were downloaded from 2,376 channels using YouTube-DL (Yen/Remite/Sergey 2020) routed through the Tor network (see below; Loesing/Murdoch/Dingledine 2010). Exact locations for channels were assigned using a geocoder by passing a string consisting of the Census Bureau entity name, the YouTube channel name, and the city and state location to a geocoder (Esmukov et al. 2018). Channels with the same location (for example, city government and city school district channels resolved to the same street address) were then merged. Tokenization of the 322,677 individual transcript files was undertaken with Spacy (Honnibal 2019). Transcripts with fewer than 100 words, as well as transcripts with textual features indicating they were not generated by the YouTube ASR algorithm and transcript files without individual word timings were removed,[3] resulting in a corpus of 270,931 transcripts from 2,189 channel locations, comprising 1,149,031,002 words and corresponding to over 141,455 hours of video from locations in all 50 U.S. states and the District of Columbia (Tab. 1).

3.2.2 Canada

For Canada, a list of Canadian municipalities or other local administrative entities and their official or semi-official government websites was created by scraping public web resources such as web pages, PDF files, and databases of the 13 Canadian provincial and territorial governments, as well as Wikipedia lists of municipalities.[4] In total, the list comprised 3,401 localities or local government agencies (mostly cities, counties, towns, villages, rural municipalities,

altered, the link may direct to the social media presence of the service provider that created the template, rather than the account of the local government entity.

3 If only manually-uploaded transcripts are available for a YouTube video, YouTube-DL will download these transcripts, even if scripts are configured to download only automatic subtitles. Some of these manually-uploaded transcripts are identifiable on the basis of their textual features, such as all-capital-letter orthography.

4 Alberta: http://municipalaffairs.gov.ab.ca/cfml/officials/Official.xls; British Columbia: https://www.ubcm.ca/EN/main/about/ubcm-members/municipalities.html; Manitoba: https://www.gov.mb.ca/mr/contactus/pubs/mod.pdf; New Brunswick: https://www2.gnb.ca/content/gnb/en/departments/elg/local_government/content/community_profiles.html, Newfoundland and Labrador: https://en.wikipedia.org/wiki/List_of_municipalities_in_Newfoundland_and_Labrador; Northwest Territories: https://www.maca.gov.nt.ca/en/community-contact-listing; Nova Scotia: https://beta.novascotia.ca/sites/default/files/documents/1-1759/municipal-statistics-annual-report-2018-en.pdf; Nunavut: https://en.wikipedia.org/wiki/List_of_municipalities_in_Nunavut; Ontario: https://www.amo.on.ca/AMO-Content/Municipal-101/Ontario-Municipalities.aspx; Prince Edward Island: https://www.princeedwardisland.ca/sites/default/files/publications/municipal_directory.pdf; Quebec: https://www.donneesquebec.ca/recherche/fr/dataset/repertoire-des-municipalites-du-

districts, or settlements, but also other entities) with websites in all 13 Canadian territories or provinces, representing 65% of census subdivisions of the 2011 Canadian Census (Statistics Canada lists 5,253 census subdivisions for the 2011 Canada Census (Statistics Canada 2011).

Two approaches were used to find YouTube channels associated with the Canadian administrative bodies aggregated in this list. First, each website was scraped directly for links to YouTube channels present on the homepage, in the same manner as employed for the US Census list. For Canada, 205 of the homepages had links to YouTube channels, of which 112 were unique.[5] In a second approach, a script iteratively sent the name of each of the 3,401 locations and its province/territory name (e.g., "City of Calgary, Alberta") to YouTube's web search interface and the first two channel results were harvested. This method resulted in 679 channels, some of which were the YouTube channels of commercial entities or channels with no connection to a Canadian place.[6]

After manual filtering to remove commercial channels, non-Canadian channels, channels automatically generated by YouTube algorithms,[7] channels with no obvious locality, and channels for which transcripts were automatic translations of French videos,[8] the lists of YouTube channels identified using the two methods were merged. All available automatic speech-to-text transcripts were downloaded from the 407 channels identified in this manner, resulting in a corpus of 30,916 video transcripts and 103,035,369 words, corresponding to over 12,586 hours of video, from all 13 of Canada's provinces and territories. Summary statistics are presented in Tab. 2.

quebec/resource/19385b4e-5503-4330-9e59-f998f5918363; Saskatchewan: http://www.mds.gov.sk.ca; Yukon: http://www.gov.yk.ca/aboutyukon/communities.html.

5 Many municipal websites link to the same YouTube channel: For example, most of the homepages for Nunavut municipalities link to the YouTube channel of the Government of Nunavut.

6 YouTube's search function for channels returns hits if any video in a channel contains the search term in its title or the description on the "About" page.

7 Channels with the string "- Topic" in the title are automatically generated by YouTube; they contain videos that have been aggregated based on individual video metadata. In many cases "Topic" channels will contain content about a particular place, but as such content is not necessarily representative of speech in that place (for example, in the case of tourism videos profiling a particular location), they were removed from the download list.

8 This is the result of an issue with the YouTube-DL code: https://github.com/ytdl-org/youtube-dl/issues/13646.

Tab. 1: US Subcorpus Summary Statistics.

State	Channels	Videos	Words	Length (h)	State	Channels	Videos	Words	Length (h)
Alabama	27	2827	10,581,345	1,315.67	Montana	3	145	926,229	143.20
Alaska	6	451	1,854,654	248.37	Nebraska	16	677	2,487,171	312.51
Arizona	35	6356	26,393,272	3,063.73	Nevada	5	2,759	6,110,915	638.06
Arkansas	14	986	6,748,658	882.77	N.H.	11	1,305	10,913,552	1,469.04
California	211	18278	83,915,246	10,146.57	New Jersey	88	6,982	29,523,334	3,977.57
Colorado	56	8802	36,551,218	4,299.68	New Mexico	14	1,895	6,750,477	883.10
Connecticut	25	3731	24,549,746	3,010.04	New York	97	8,037	37,560,959	4,856.87
Delaware	3	148	242,073	25.45	N. Carolina	97	11,357	46,231,979	5781.40
District of Columbia	3	242	261,209	32.90	N. Dakota	10	768	3,616,363	442.05
Florida	89	17625	64,647,923	7,468.48	Ohio	97	7,647	33,695,476	4,268.46
Georgia	49	5487	18,565,796	2,421.53	Oklahoma	19	1,977	5,271,339	643.35
Hawaii	1	152	123,617	15.42	Oregon	38	2,769	15,675,898	1,992.84
Idaho	11	1547	8,747,885	1,012.14	Pennsylvania	74	6,984	32,571,217	3,970.32
Illinois	151	14243	54,613,612	6,725.31	Rhode Island	7	822	3,195,777	530.94
Indiana	46	4017	12,958,084	1,643.88	S. Carolina	24	3,894	8,716,589	1115.20
Iowa	43	7516	24,286,940	3,072.57	S. Dakota	12	1,819	18,619,258	2,172.97
Kansas	35	4444	19,862,293	2,504.08	Tennessee	33	7,194	43,286,858	5,127.52
Kentucky	26	4965	17,834,978	2,092.75	Texas	155	21,330	44,736,009	5,789.44
Louisiana	16	2018	10,500,407	1,221.96	Utah	21	2,561	7,766,782	940.21
Maine	12	819	5,879,165	797.01	Vermont	3	94	131,558	16.62
Maryland	32	7373	34,009,832	4,100.84	Virginia	42	9,209	34,806,149	4,059.67
Massachusetts	44	17596	11,517,230	14,682.19	Washington	51	6,178	28,949,403	3,387.77
Michigan	90	9832	51,293,982	6,079.47	W. Virginia	6	101	196,479	25.86
Minnesota	80	8666	31,366,468	3,661.89	Wisconsin	83	9,514	45,983,568	5,744.59
Mississippi	18	1448	2,613,901	346.07	Wyoming	7	251	2,638,963	348.39
Missouri	53	5093	15,094,086	1,946.43					

Tab. 2: Canada Subcorpus Summary Statistics.

Province/territory	Channels	Videos	Words	length (h)
Alberta	95	6,623	21,239,251	2,497.45
British Columbia	102	10,002	26,853,481	3,246.83
Manitoba	20	3,286	2,771,200	318.21
New Brunswick	8	382	2,347,141	278.05
Newfoundland and Labrador	2	108	186,070	29.99
Northwest Territories	3	32	21,404	3.27
Nova Scotia	11	332	1,229,149	148.38
Nunavut	1	6	1,230	0.23
Ontario	112	8,404	45,970,092	5,774.59
Prince Edward Island	6	753	777,772	95.87
Quebec	6	166	486,265	60.29
Saskatchewan	10	663	895,143	103.12
Yukon	7	159	257,171	30.48

3.2.3 CoNASE

The U.S. and Canadian resources were combined with the corpus described in Coats (2019) to create the Corpus of North American Spoken English (CoNASE) of more than 1.25 billion words (CoNASE; Coats 2021). Fig. 1 shows the locations of the channels from which transcripts were downloaded in this combined corpus. Circle sizes are proportional to the number of videos sampled from the channel(s) at that location.

3.2.4 British Isles

For the British Isles, a method similar to that employed for North America was employed: A list of the names of local government authorities in England, Scotland, Wales, Northern Ireland, and the Republic of Ireland was created in November 2019 from information available on Wikipedia,[9] then searches for the name of the

9 https://en.wikipedia.org/wiki/List_of_county_councils_in_England, https://en.wikipedia.org/wiki/Unitary_authorities_of_England, https://en.wikipedia.org/wiki/Metropolitan_borough, https://en.wikipedia.org/wiki/London_boroughs, https://en.wikipedia.org/wiki/Non-metropolitan_district, https://en.wikipedia.org/wiki/Subdivisions_of_Scotland, https://en.wikipedia.org/wiki/List_of_Welsh_principal_areas_by_area, https://en.wikipedia.org/wiki/Local_government_in_Northern_Ireland, and https://en.wikipedia.org/wiki/Local_government_in_the_Republic_of_Ireland. The council for the Isles of Scilly was added manually.

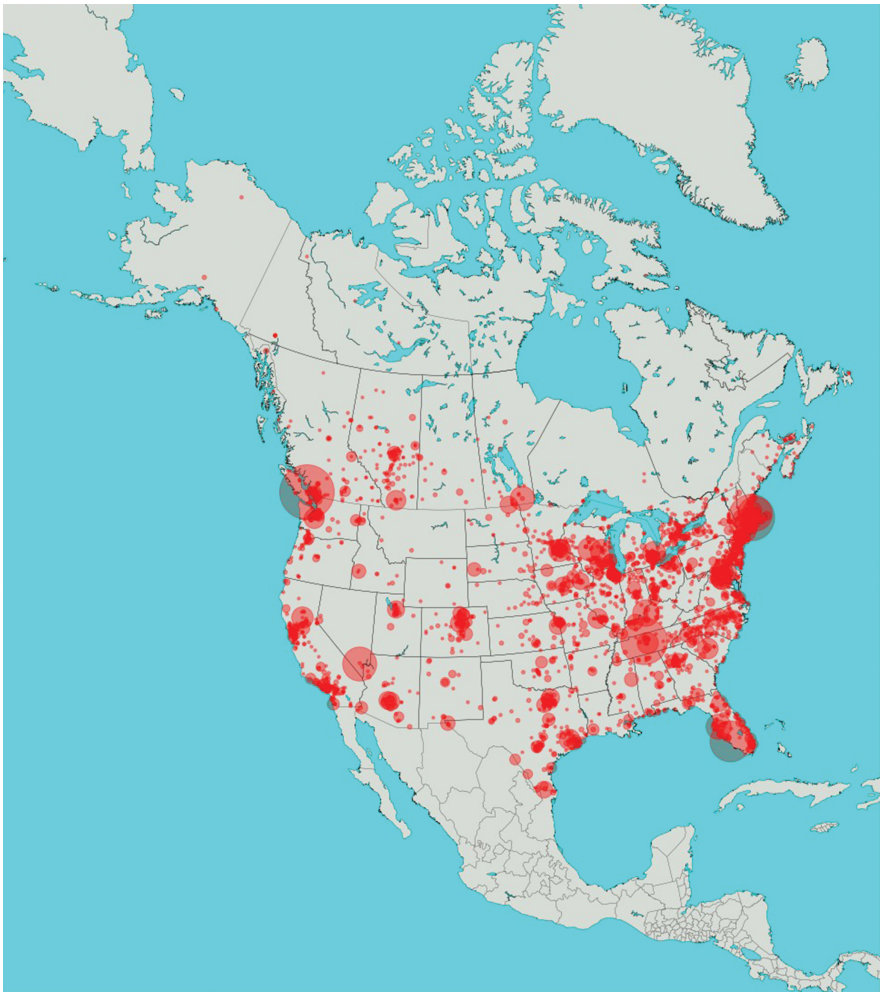


Fig. 1: Locations and sizes of sampled channels in CoNASE.

authority plus the string "Council" were sent to the search function on YouTube's web page for each of the 413 local government entities (e.g., "Dorset Council", "East Ayrshire Council", "Mayo County Council", etc.). The first three "channel" results ranked in order of relevance were retrieved. Results were then filtered to retain channels that included the strings "council" or "cc" in the channel name. Almost all of these were the official YouTube channels of the regional authorities targeted by the search procedure, although in a few cases, both an official and an unofficial channel existed for a given local authority with the same name or very

similar names.[10] In addition to these 'unofficial' channels, likely created automatically by scripts, channel duplicates, channels automatically generated by YouTube, and channel false positives (e.g., the channel "Boston City Council" from the United States, rather than Lincolnshire, or "Ipswich City Council TV" from New South Wales, Australia) were removed after a content check.

In 2021, websites of local governments in England, Scotland, and the Republic of Ireland were scraped to retrieve several additional channels.[11] In total, the British Isles corpus contains transcripts from 453 geolocated channels, comprising 38,680 transcript files and 111,563,614 tokens, and corresponding to more than 12,801 hours of video. A summary of the results by country is presented in Tab. 3.

Tab. 3: UK and Ireland Corpus Summary Statistics.

Country	Channels	Videos	Words	Length (h)
England	324	23,657	72,879,173	8,521.71
Northern Ireland	10	1,898	6,508,505	770.84
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

The map in Fig. 2 depicts the locations assigned to the channels by the geocoding procedure with circle sizes proportionate to the number of videos in each location. As can be seen, channel density is high in relatively densely-populated parts of the British Isles such as London, the Midlands, and the 'Central Belt' of Scotland, but lower in the North of England, Wales, most of Scotland, and Ireland.

10 For example, the channel "Stoke-on-Trent City Council" (https://www.youtube.com/channel/UCTrvOc-4pd_ME-RyuN5ZBMQ) contains a large number of videos and is the official channel of the authority. "Stoke City Council" (https://www.youtube.com/channel/UCngBVsm9z3OAR3j7vV2AF8Q) contains only four videos.

11 The channels listed at https://www.local.gov.uk/our-support/guidance-and-resources/communications-support/digital-councils/social-media/go-further/a-z-councils-online plus channels scraped from sites listed at https://www.mygov.scot/organisations#scottish-local-authority and https://www.gov.ie/en/organisation-information/fd139-local-government-councils-and-councillors.

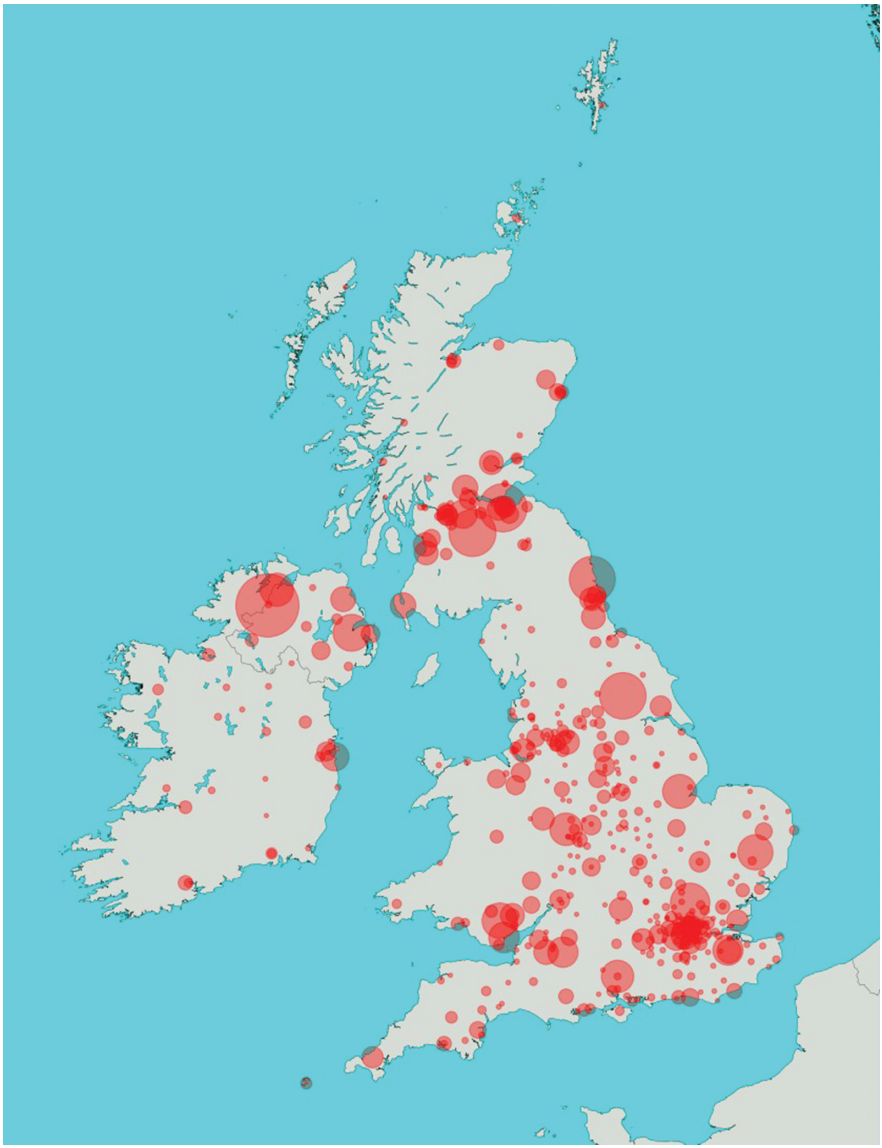


Fig. 2: Locations and sizes of sampled channels in the United Kingdom and Ireland.

3.3 Use of the Tor Network

The procedures described above require large numbers of requests to be sent iteratively by scripts to YouTube servers, which can result in the sender's IP address being blocked for 24 hours, 48 hours, or longer. To surmount this problem, scripts can be designed to send requests from multiple IP addresses, automatically switching addresses after a certain number of requests. Most researches do not have access to multiple IP addresses, and the cost of acquiring multiple IPs via a virtual private network may be prohibitive. For this reason, the Tor network was used to send requests to YouTube servers. Tor, an open-source software protocol for anonymous internet use, sends encrypted HTTP requests to a target via a randomized network of node servers (Loesing/Murdoch/Dingeldine 2010). Periodically generating a new Tor connection changes the Tor 'exit node' and thus the IP address of the server from which the request is passed YouTube. For the collection of transcripts described in this paper, the Tor exit node was changed every 1,000 calls to YouTube made by the YouTube-DL library. While using Tor can circumvent IP blocking, it reduces the download speed of the script pipeline. To generate the corpora described in this paper, it was necessary to run the download scripts for several weeks.

Although the methods described above focus on the creation of corpora of ASR transcripts from specific locations, they could also be used for the creation of other types of specialized corpora, for example pertaining to specified content, communicative situations, or speaker demographic attributes. In addition, because the functionality of YouTube-DL allows users to download the original video file as well as captions or other metadata, the basic procedure described above can be employed for the creation of specialized corpora of video or audio files from YouTube or other websites; these could then be subjected to acoustic or audio-visual analysis.

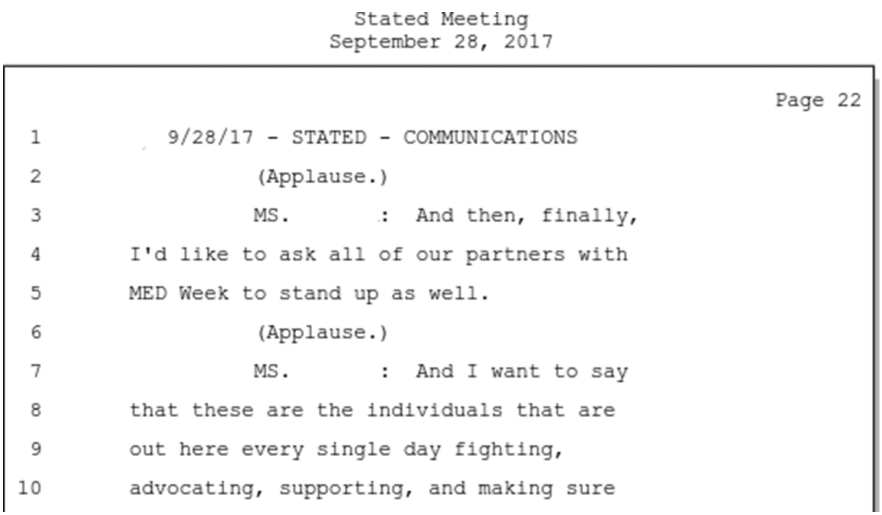
4 Test Cases

YouTube ASR transcripts can be considered a type of 'noisy' data: they contain errors, which can be due to low acoustic fidelity in the audio source, inaccurate identification of the language being spoken by the ASR algorithm, overlapping speech, music in the background, or other causes. In the following two subsections, the accuracy of the ASR transcripts is measured and an example of transcript classification using noisy corpus data is described.

4.1 WER of ASR Transcripts

The WER of ASR transcripts was calculated by comparing them with publicly-available manual transcripts of council sessions of the American city of Philadelphia, Pennsylvania. The city of Philadelphia, like many larger American cities, hires stenographers to produce official transcripts of meetings of local government bodies. In Philadelphia, the service is provided by a stenography firm that specializes in the transcription of courtroom proceedings (which for most types of trials are required by law to be transcribed).

In order to retrieve the official transcripts of the 40 Philadelphia City Council meetings whose ASR transcripts were present in the North American corpus described above, a script was written to scrape the website of the city of Philadelphia for links to the corresponding transcript files, which were then downloaded.



Stated Meeting
September 28, 2017

Page 22

9/28/17 - STATED - COMMUNICATIONS

(Applause.)

MS. : And then, finally, I'd like to ask all of our partners with MED Week to stand up as well.

(Applause.)

MS. : And I want to say that these are the individuals that are out here every single day fighting, advocating, supporting, and making sure

Fig. 3: Excerpt of official transcript of the Philadelphia City Council meeting of 28 September of 2017.

The files, in PDF format (an example excerpt is provided in Fig. 3), were converted to text using Apache Tika (2021), then processed to remove all text that did not correspond to speech, such as the title of the transcript, the time and location of the transcribed meeting, the list of participants, page headers and page numbers, the name and telephone number of the company that prepared the transcript, the certification of the stenographer that the transcript is accurate, the index at the end of the transcript, and all indications of speaker diarization

(names of speakers followed by colons).[12] Parenthetical annotations that did not correspond to speech were also removed, such as "(Councilmember and guests approached podium.)", "(No response.)", "(Applause.)", or "(The council is at ease.)". After the cleaned texts were stripped of remaining punctuation and excess whitespace and converted to lower case, they were used to calculate the WERs of the corresponding ASR transcripts.

Word error rate is calculated with

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D the number of deletions, and I the number of insertions necessary to transform the 'hypothesis' text (i.e., the text whose accuracy is to be tested, in this case the ASR transcript) to the 'ground truth' text (i.e., the manual transcript); N is the number of words in the 'ground truth' text. WER ranges from 0 (texts are identical) to 1 (texts have zero overlap). For example, the WER of the strings "welcome to our council meeting" and "welcome to the city council meeting", where the first string is the hypothesis and the second string the ground truth, would be 2/6 or 0.333. Word error rate (WER) was calculated using the *jiwer* library in Python (Vaessen 2020). For the 40 transcript pairs, the mean WER was 0.22, with a standard deviation of 0.03 and a range from 0.15 to 0.29.

This WER is comparable to some values reported in the literature, but does not give a good indication of how useful the ASR transcripts may be for linguistic analysis. In order to gauge the comparability of the ASR and manual transcripts, word frequencies in aggregated transcripts were compared. ASR transcripts were aggregated into one text, and manual transcripts into another. The relative frequencies of all word types were then calculated in both aggregated texts. The log-likelihood score (Dunning 1993; Rayson/Garside 2000) and corresponding p-value were used to compare the frequencies of the 14,433 word types with at least one occurrence in each of the aggregated texts. For 13,929 types (96.5% of the shared word types), no significant difference in usage was found at an alpha level of $p = 0.05$. For 504 word types (3.5% of the shared word types), a significant difference in frequency was found at $p = 0.05$. The types that exhibit significant differences in use between the manual and automatic transcripts are various: Many are personal names ("Clarke", "Belen", "Bill") or other proper nouns such as place names ("Roxborough", a suburb of Philadelphia, "Leverington", a street name). Legal terminology ("writ", "mandamus") and words common in the

12 YouTube ASR transcripts do not contain diarization metadata as of 2021.

specific context of a council meeting but otherwise relatively rare in spoken language ("rezoning", "councilperson") show significant frequency differences, as do some digits and numerals ("12", "706"), possibly in part due to the various ways in which numbers can be phonetically realized in spoken English.[13] In addition, some words that are homonyms show significant frequency differences between the ASR and manual transcripts, such as "gym" ("Jim") and "I" ("aye"). Among the types that show significant differences in frequency but are otherwise relatively common English words are "teen", "emotion", and "meaning", among others. Further investigation is necessary to determine why such types may inaccurately transcribed in this data.

In Fig. 4, the logarithm of the frequency for each of the 14,433 word types is plotted in the ASR transcripts (x-axis) and the manual transcripts (y-axis). If the two aggregate transcripts were exactly equivalent, all words would have the same frequency in both texts and scatterplot points would fall on a straight line. As can be seen, for low-frequency items there is considerable variation in word frequencies (i.e., many errors), but more frequent words tend to show comparable frequencies.

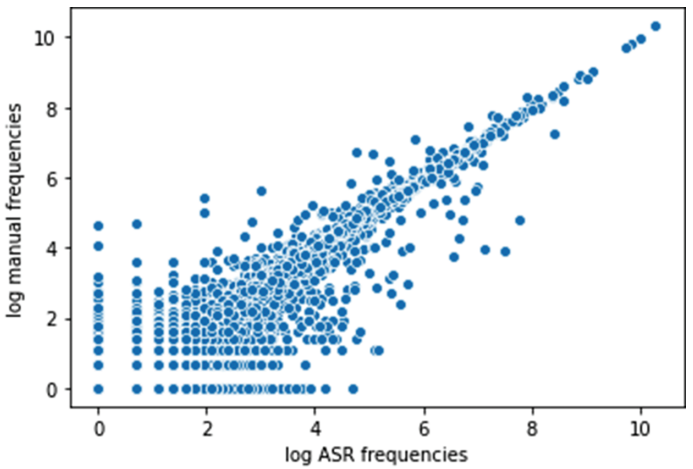


Fig. 4: Log-log plot of frequencies of shared types.

13 For example, 344 can be "three hundred and forty four", "three forty four", or "three four four", depending on if it is spoken as part of a residential address, a telephone number, or some other numerical quantity.

Vector representations of documents (individual video transcripts) and vocabulary items were created from a subset of the US corpus comprising 78,238 transcripts whose video titles included the words "council", "session", or "meeting", totalling 691,442,599 words. SpaCy (Honnibal 2019) was used for tokenization, part-of-speech tagging, removal of named entities such as organizations and place names, and restriction of the vocabulary to nouns, verbs, adjectives, and interjections. Doc2Vec (Le/Mikolov 2014), a variant of the popular Word2Vec neural network model (Mikolov/Yih/Zweig 2013) which also allows tagged documents (in this case, individual transcripts) to be embedded in the same multidimensional space as individual words, was employed to generate a model in which each of the 78,238 transcripts was tagged with one of 51 labels (for the 50 US states and the District of Columbia). The Gensim implementation of Doc2Vec was used, with distributed bag-of-words training, a window size of 15 words, 300-dimensional vectors, a minimum frequency of 50 occurrences per word type, and 20 training epochs (Rehurek/Sojka 2011).

This model, which embeds vectors for individual words and vectors for document tags (state names) in the same multidimensional space, makes it possible to see which words are closest to each state. In addition to words denoting activities, geographical features or crops important in some states (for example, the closest words for Alaska included "fisheries" and "harbor", while the closest words for some Midwestern states included "corn" and "vetch"), the model managed to capture some features of American lexis that may be regionally distributed: For example, the vocabulary items "folks", "alrighty", and "sir" were found to be among the vectors nearest to the Southern states of North Carolina, South Carolina, and Georgia.

To test the ability of the model to accurately predict the provenance of regional language, a simple logistic regression classifier was trained for the transcripts from California and New York, using 90% of the transcripts from those two state locations as training material and 10% as test material. Classifier accuracy was 96.7% for the test transcripts: Of the 634 test transcripts from California, 618 were accurately classified; of the 251 New York test transcripts, 238 were accurately classified.

Next, t-SNE (van der Maaten/Hinton 2008) was used to project the 300-dimensional vectors into 2-dimensional space. Fig. 5. visualizes vector similarity for the state-level labels based on the aggregate documents and vocabulary from that state. As can be seen, vector representations derived from ASR transcripts recapitulate to some extent geographical proximity: A Southern cluster, comprising Tennessee, Florida, Louisiana, North Carolina, Virginia, South Carolina, Alabama, Mississippi, and Georgia is evident at the top of the figure. A New England cluster of Maine, Massachusetts, New Hampshire, Rhode Island, and Connecticut is

apparent to the left, and the Midwestern states of Illinois, Wisconsin and Minnesota form a cluster to the right of Fig. 5 in close proximity to the neighbouring states of North and South Dakota, Montana, Iowa, and Nebraska. At the bottom of the figure the Western states of Utah, Wyoming, Washington, Oregon, Colorado, and Idaho are clustered together.

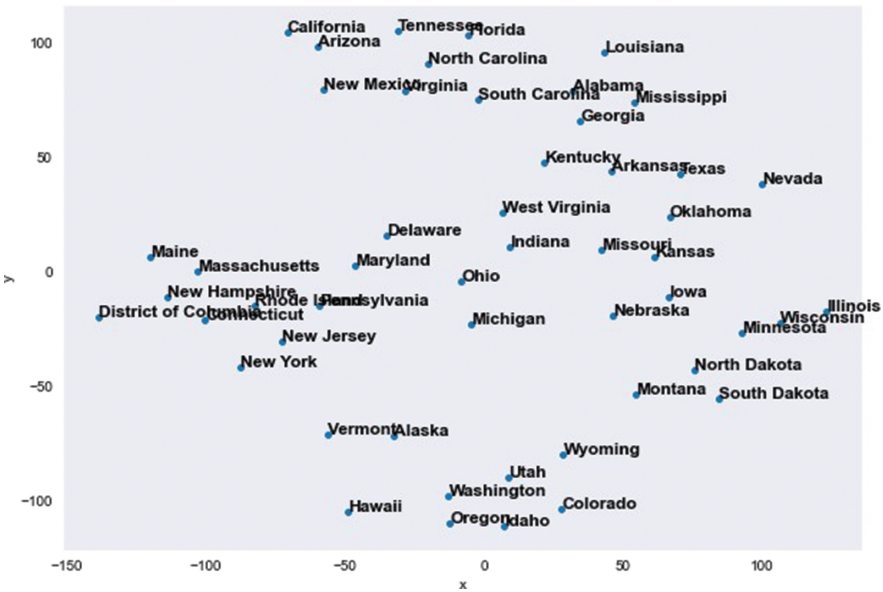


Fig. 5: t-SNE map of vector similarity for US states and Washington, DC.

Document classification or the calculation of cosine similarity for vector representations may not be tasks that directly correspond to analysis of linguistic variation in terms of lexis or morpho-syntax, but they are ultimately also based on frequency information. The high level of accuracy achieved by the classification task and the geographical patterns of similarity generated from vectors suggest that relative frequencies in a constrained vocabulary model can be used to identify basic patterns of regional variation in spoken American English. More sophisticated feature representations, for example in which morpho-syntactic variability is identified using regular expressions, may further increase the accuracy of NLP tasks, as well as provide more direct insight into linguistic variation.

5 Discussion

5.1 Methodological Caveats

Inherent features of YouTube ASR transcripts as well as methodological procedures pertaining to sampling and filtering techniques and the assignment of transcripts to geographical locations need to be kept in mind when considering the types of analysis that can be undertaken using these and similar corpora.

ASR transcripts contain errors, and rare lexical items are often incorrectly transcribed. In addition, some potential phonological or morpho-syntactic linguistic features are subject to normalization by the ASR algorithm and therefore may be inaccurately recorded in the transcripts. These include non-standard stem vowels in past tense forms of strong verbs ("I sot" for *to sit*) or non-standard weak past tense forms for verbs that are typically conjugated according to the strong paradigm (e.g., *blowed, dealed, drinked*), which are attested as features in some varieties of English dialectal speech, but are have not been found in the ASR transcripts, likely due to the ASR model having been trained mainly on transcripts of standard speech. Similarly, non-standard verbal agreement (e.g., "I likes", "they was") in speech may be rendered according to the standard paradigm in ASR transcripts due to the preponderance of standard forms in the training data for YouTube's ASR algorithms.

The transcripts used for the creation of these corpora do not contain speaker metadata or any indication of speaker diarization. However, the structure of the corpora facilitates manual annotation of this and other metadata: Because the word tokens in the corpora contain timing information, the corresponding videos can be checked at the time of utterance for a given phenomenon of interest, and relevant metadata recorded.

The WER analysis presented in Section 4 shows that ASR and manual transcripts are not equivalent, but the manual transcripts from Philadelphia may also be inaccurate: Taylor et al. (2019) tested a sample of Philadelphia courtroom stenographers and found that their transcripts of recordings of speech of African-Americans who had a history in the criminal justice system did not necessarily correspond to the researchers' own transcripts, either for verbatim transcripts or for a "paraphrase task" in which the speech was translated into Standard American English, particularly for the representation of aspectual properties of the verbal phrase.[14] The assessment of transcript accuracy in Section 4, however, is

14 The authors found that the accuracy of transcripts prepared by experienced court stenographers varied from 8.8% to 41.6%, with black court reporters showing higher WERs.

based on language delivered in the relatively formal situational context of a city council meeting and hence more likely to correspond to the norms of standard American English than to African-American Vernacular English.

5.2 Potential Features for Analysis

Due to the normalization of the ASR transcripts, variation that occurs within the constraints of standard orthographical forms is better suited for the exploration of regional variation in the YouTube corpora. A large number of potential morphological and syntactic features have been identified in previous studies, including lexical and word order variation features that could be examined in orthographic transcripts. In a study of patterns of negation in spoken British English, Anderwald (2002) made use of orthographic transcripts from the BNC as well as smaller corpora. Kortmann/Szmrecsanyi (2004) summarized morpho-syntactic variation in global English varieties on the basis of 76 grammatical features grouped into 11 categories. Szmrecsanyi (2011), in a discussion of the outlook for corpus-based dialectological studies, used the frequencies of 57 morphosyntactic features in the *Freiburg Corpus of English Dialects* (Szmrecsanyi/Hernández 2007) to explore patterns of regional variation in spoken British English. Grieve (2016) showed that lexical and morpho-syntax features in written American letters to the editor of newspapers exhibit regional variation.

Additional features that could be examined in this framework include, for example, politeness words (Culpepper/Gillings 2018), intensifiers (Aijmer 2018), variation manifest in multi-word sequences such as dative alternation (Jenset/McGillivray/Rundell 2018) or non-standard reflexive pronoun deixis (Paterson 2018). The corpora may also be suitable for studies of conversational phenomena such as word repetition or repair sequences.

6 Summary and Future Outlook

Automated methods were used to create large corpora of ASR speech transcripts from YouTube channels of geographically localized local government entities in the United States, Canada, and the British Isles. Web-scraping scripts, the Tor network, and the open-source YouTube-DL library, when used in concert, allow the researcher to create large corpora of ASR transcripts that may be suitable for linguistic analysis of regional variation in English. In addition, with minor script modifications, such a corpus-creation pipeline allows the collection of transcript

material according to pre-defined register, genre, or other parameters, as well as the download of video and audio data for acoustic analysis.

Word error rates for a subset of the ASR transcripts in the US corpus were found to be approximately 22%, making some types of analysis less feasible. However, in aggregate, only 3.5% of the word types attested in both ASR and manual transcripts showed a significant difference in frequency, according to a log-likelihood test.

The findings of Agarwal et al. (2007), Eder (2013), Franzini et al. (2018) and Pentland et al. (2019), as well as the simple classification presented in this study, suggest that some tasks may be relatively robust to high error rates in transcripts, presumably due to the fact that the even in transcripts with many errors, with sufficient sample sizes, distinct patterns emerge in the relative frequencies of accurately transcribed features (i.e., words). Vector representations of corpus vocabulary and corpus transcripts can be used to investigate patterns of geographical variability – a simple embeddings model using a restricted vocabulary was found to recapitulate some state-level geographical clusters, and some lexical items associated with particular regions in the US were found to be among the items closest to state labels.

Future work could be organized along the following lines: First, similar corpora are planned for other countries in which local government business is conducted in English, such as Australia, New Zealand, and South Africa. Continual refinements to YouTube's language models should allow ASR corpora in other languages to be compiled, as well. Second, the investigation of variation in spoken English in North America and the British Isles can proceed, for example, by using regular expressions to capture morpho-syntactic variants rendered in standard orthography or by using word-vector based methods (Hovy/Purschke 2018). Large corpora of geo-located speech obtained from ASR transcripts will open up new possibilities to explore the diversity and development of spoken English in terms of its geographical variability.

References

Agarwal, Sumeet/Godbole, Shantanu/Punjani, Diwakar/Roy, Shourya (2007): "How much noise is too much: A study in automatic text classification." In: Geetha Jagannathan/Rebecca N. Wright (Eds.): *Proceedings of the Seventh IEEE international Conference on Data Mining (IDCM 2007), Omaha, NE, 2007*. 3–12.

Aijmer, Karin (2018): "'That's well bad': Some new intensifiers in spoken British English." In Vaclav Brezina/Robbie Love/Karin Aijmer (Eds.): *Corpus Approaches to Contemporary*

British Speech. Sociolinguistic Studies of the Spoken BNC2014. New York: Routledge, 60–95.

Anderwald, Liselotte (2002): *Negation in Non–Standard British English. Gaps, Regularizations and Asymmetries*. London: Routledge.

Tika.apache.org. (2021): *Apache Tika – Apache Tika*. Online at: https://tika.apache.org <05.11.2021>.

Bokhove, Christian/Downey, Christopher (2018): "Automated generation of 'good enough' transcripts as a first step to transcription of audio–recorded data." In: *Methodological Innovations* 11, 1–14.

Busse, Beatrix (2018): "Current British English: The sociolinguistic perspective." In: Vaclav Brezina/Robbie Love/Karin Aijmer (Eds.): *Corpus Approaches to Contemporary British Speech. Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge, 16–26.

Chiu, Chung–Cheng/Sainath, Tara/Wu, Yonghui/Prabhavalkar, Rohit/Nguyen, Patrick/Chen, Zhifeng/Kannan, Anjuli/Weiss, Ron J./Rao, Kanishka/Gonina, Ekaterina/Jaitly, Navdeep/Li, Bo/Chorowski, Jan/Bacchiani, Michiel (2018): "State–of–the–art speech recognition with sequence–to–sequence models." arXiv:1712.01769v6 [cs.CL], 1–5.

Coats, Steven (2019): "A corpus of regional American language from YouTube." In: Constanza Navarretta/Manex Agirrezabal/Bente Maegaard (Eds.): *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference (DHN 2019)*. CEUR, 79–91.

Coats, Steven (2020): "Articulation rate in American English in a corpus of YouTube videos." In: *Language and Speech* 63, 799–831.

Coats, Steven (2021): *Corpus of North American Spoken English* (CoNASE). Online at: http://cc.oulu.fi/~scoats/CoNASE.html <12.05.2022>.

Culpeper, Jonathan/Gillings, Matthew (2018): "Politeness variation in England: A North–South divide?" In: Vaclav Brezina/Robbie Love/Karin Aijmer (Eds.): *Corpus Approaches to Contemporary British Speech. Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge, 33–59.

Dunning, Ted (1993): "Accurate methods for the statistics of surprise and coincidence." In: *Computational Linguistics* 19, 61–74.

Eder, Macei (2013): "Mind your corpus: Systematic errors in authorship attribution." In: *Literary and Linguistics* 28, 603–614.

Esmukov, Kostya (2018): *Geophy* [Python module]. Online at: https://github.com/geopy/geopy <14.04.2022>.

Franzini, Greta/Kestemont, Mike/Rotari, Gabriela/Jander, Melina/Ochab, Jeremi K./Franzini, Emily/Byszuk, Joanna/Rybicki, Jan (2018): "Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm." *Frontiers in Digital Humanities* 5, 1–15.

Google Developers (2021): *API Reference*. Online at: https://developers.google.com/youtube/v3/docs <14.04.2021>.

Grieve, Jack (2016): *Regional Variation in Written English*. Cambridge: Cambridge University Press.

Halpern, Yoni/Hall, Keith/Schogol, Vlad/Riley, Michael/Roark, Brian/Skobeltsyn, Gleb/Bäuml, Martin (2016): "Contextualizing prediction models for speech recognition." In: *Proceedings of Interspeech 2016*, 2338–2342.

Honnibal, Matthew (2019): *SpaCy* [Python Module]. Online at: https://github.com/explosion/spaCy <14.04.2022>.

Hovy, Dirk/Purshke, Christoph (2018): Captioning regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4383–4394.

Jenset, Gard B./McGillivray, Barbara/Rundell, Michael (2018): "The dative alternation revisited: Fresh insights from contemporary British spoken data." In: Vaclav Brezina/Robbie Love/Karin Aijmer (Eds.): *Corpus Approaches to Contemporary British Speech. Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge, 185–208.

Kawahara, Tatsuya (2012): "Transcription system using automatic speech recognition for the Japanese parliament (Diet)." In: *Proceedings for the Twenty–Fourth Innovative Applications of Artificial Intelligence Conference*, 2224–2228.

Këpuska, Veton/Bohouta, Gamal (2017): "Comparing speech recognition systems." In: *International Journal of Engineering Research and Applications* 7, 20–24.

Kim, Joshua Y./Liu, Chunfeng/Calvo, Rafael A./McCabe, Kathryn/Taylor, Silas C. R./Schuller, Björn W./Wu, Kaihang (2019): "A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech." arXiv:1904.12403, 1–13.

Kortmann, Bernd/Szmrecsanyi, Benedikt (2004): "Global synopsis: Morphological and syntactic variation in English." In: Bernd Kortmann/Edgar W. Schneider/Kate Burridge/Rajend Mesthrie/Clive Upton (Eds.): *A Handbook of Varieties of English, Vol. 2. Morphology and Syntax*. Berlin/New York: Mouton de Gruyter, 1142–1202.

Le, Quoc V./Mikolov, Tomas (2014): "Distributed representations of sentences and documents." In: *Proceedings of the 31st International Conference on Machine Learning (ICML–14)*, 1188–1196.

Liao, Hank/McDermott, Erik/Senior, Andrew (2013): "Large scale deep neural network acoustic modelling with semi–supervised training data for YouTube video transcription." In: *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. New York: IEEE, 368–373.

Loesing, Karsten/Murdoch, Steven J./Dingledine, Roger (2010): "A case study on measuring statistical data in the Tor anonymity network." In: Radu Sion/Reaz Curtmola/Sven Dietrich/Aggelos Kiayias/Josep M. Miret/Kazue Sako/Francesc Sebé (Eds.): *Financial Cryptography and Data Security. FC 2010 Workshops, RLCPS, WECSR, and WLC 2010 Tenerife, Canary Islands, Spain, January 2010, Revised Selected Papers*. Springer, 203–215.

Maaten, Laurens van der/Hinton, Geoffrey E. (2008): "Visualizing high–dimensional data sing t–SNE." In: *Journal of Machine Learning Research* 9, 2579–2605.

McEnery, Tony (2018): "The spoken BNC2014: The corpus linguistic perspective." In: Vaclav Brezina/Robbie Love/Karin Aijmer (Eds.): *Corpus Approaches to Contemporary British Speech. Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge, 10–15.

Mikolov, Tomas/Yih, Wen–Tau/Zweig, Geoffrey (2013): "Linguistic regularities in continuous space word representations." In: *Proceedings of HTL–NAACL 13*, 746–751.

Muthukadan, Baiju (2018): *Selenium with Python*. Online at: https://selenium–python.readthedocs.io/ <14.04.2022>.

Peterson, Laura L. (2018): "'You can just give those documents to myself': Untriggered reflexive pronouns in 21st century spoken British English." In: Vaclav Brezina/Robbie Love/Karin Aijmer (Eds.): *Corpus Approaches to Contemporary British Speech. Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge, 235–255.

Pentland, Steven/Spitzley, Lee/Fuller, Christie/Twitchell, Doug (2019): "Data quality relevance in linguistic analysis: The impact of transcription errors on multiple methods of linguistic

analysis." In: *AMCIS 2019: Proceedings of the 25th Americas Conference on Information Systems*. Online at: https://aisel.aisnet.org/amcis2019/human_computer_interact/human_computer_interact/12 <05.09.2022>.

Ranchal, Rohit/Taber–Doughty, Teresa/Guo, Yiren/Bain, Keith/Martin, Heather/Robinson, J. Paul/Duerstock, Bradley S. (2013): "Using speech recognition for real–time captioning and lecture transcription in the classroom." In: *IEEE Transactions on Learning Technologies* 6, 299–311.

Rayson, Paul/Garside, Roger (2000): "Comparing corpora using frequency profiling." In: *WCC '00 Proceedings of the Workshop on Comparing Corpora*, 1–6. ACM: New York.

Rehurek, Radim/Sojka, Petr (2011): "Gensim–python framework for vector space modelling." In: *NLP Centre, Faculty of Informatics* 3. Masaryk University: Brno, Czech Republic.

Sainath, Tara N./Vinyals, Oriol/Senior, Andrew/Sak, Hasim (2015): "Convolutional, long short–term memory, fully connected deep neural networks." In: *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*. New York: IEEE, 4580–4584.

Statistics Canada (2011). Tab. 5 Census subdivision types by province and territory, 2011 Census.

Szmrecsanyi, Benedikt (2011): "Corpus based dialectometry: A methodological sketch." In: *Corpora* 6, 45–76.

Szmrecsanyi, Benedikt/Hernández, Nuria (2007): *Manual of Information to Accompany the Freiburg Corpus of English Dialects Sampler ("FRED–S")*. Freiburg: University of Freiburg. Online at: http://www.freidok.unifreiburg.de/volltexte/2859/ <14.04.2022>.

Tatman, Rachael (2017): "Gender and dialect bias in YouTube's automatic captions." In: *Proceedings of the First Workshop on Ethics in Natural Language Processing, April 4th, 2017, Valencia, Spain*. Stroudsburg, PA: Association for Computational Linguistics, 53–59.

U.S. Census Bureau (2017): *Public Use Files*. Online at: https://www.census.gov/data/datasets/2017/econ/gus/public–use–files.html <14.04.2022>.

Vaessen, Nik (2020): *JiWER: Similarity measures for automatic speech recognition evaluation* (version 2.1.0). Online at: https://pypi.org/project/jiwer <14.04.2022>.

Xiong, Wayne/Droppo, Jasha/Huang, Xuedong/Seide, Frank/Seltzer, Michael/Stolcke, Andreas/Yu, Dong/Zweig, Geoffrey (2017): "Toward human parity in conversational speech recognition." In: *IEEE/ACM Transactions on Audion, Speech and Language Proceedings* 25, 2410–2423.

Yen, C. H./Remite, A/Sergey, M. (2020): *YouTube–dl* [Software]. Online at: https://github.com/rg3/youtube–dl/blob/master/README.md <14.04.2022>.

Ziman, Kirsten/Heusser, Andrew/Fitzpatrick, Paxton/Field, Campbell/Manning, Jeremy (2018): "Is automatic speech–to–text transcription ready for use in psychological experiments?" In: *Behavior Research Methods* 50, 2597–2605.