A Corpus of Regional American Language from YouTube

Steven Coats[1][0000-0002-7295-3893]

[1] English Philology, University of Oulu, 90014 Oulu, Finland
steven.coats (at) oulu.fi

Abstract. Recent years have seen an increase in the number of corpora of regional language variation for English, allowing new types of aggregate analysis to be conducted. While the creation of a corpus from written language material is relatively straightforward, transcribing speech is time-consuming, and thus there are no large corpora of transcribed American speech with broad geographic coverage. This paper describes the creation of a new corpus of regional American English from the automatically generated captions of videos from YouTube channels with a local American focus – mainly channels of regional and local government entities or civic organizations. The corpus, which consists of transcripts of over 29,267 hours of spoken language, will enable the analysis of regional patterns of lexical, morphosyntactic, and other types of variation in spoken American English. Exploratory analysis and mapping of the corpus data indicates regional variation in spoken language is evident.

Keywords: Corpus linguistics, American dialects, YouTube, Social media

1 Introduction

1.1 Background

Differences between written and spoken language have been acknowledged in the linguistics literature. Compared to written texts, spoken language exhibits, for example, higher frequencies of grammatical and lexical features associated with interactional involvement, such as second person pronouns, present tense verb forms, or verbs of cognition (e.g. *believe, see, think, understand*), and lower frequencies of features associated with the presentation of information and organization of discourse, such as nouns, adjectives, past tense verbs, or longer words [1, 2].

Corpus-based analysis of written American English has shed light on regional differences in lexis and morphosyntax [3], but no large comparable corpora of spoken American English exist, and thus regional variation in the lexis and morphosyntax of speech have been documented mainly in regional linguistic atlases or in studies utilizing data from linguistic atlas projects; this data may not adequately represent the range of variation present in a speech community at a particular place.

A corpus-based approach to the study of regional spoken language variation offers distinct advantages over the traditional approach in which data is aggregated from linguistic atlases. In this paper, the methods used to create a large corpus of transcripts of spoken English at 539 locations within the United States from automatically generated captions of YouTube videos are described.

1.2 Organization of the Text

In the following section, a brief overview of some of the previous literature pertaining to American dialectology, speech-to-text technology, and YouTube video captions is presented. In Section 3, the methods used to collect and analyze the data are described. Section 4 presents an exploratory analysis and visualization using an autocorrelation statistic from spatial geography, and Section 5 discusses the outlook for future work with the corpus.

2 Previous work

2.1 Linguistic atlas projects

In the 20th century, regional language variation in the United States was investigated in the context of regional linguistic atlas projects. Armed with a questionnaire, fieldworkers conducted interviews with pre-selected informants and recorded their responses to several hundred questions designed to elicit lexical, phonetic, and morphosyntactic variation. The first projects collected data from New England and the Middle Atlantic States in the 1930s; later projects, not all of which were completed, collected material from the Gulf States, the North Central States, the Upper Midwest, Oklahoma, the Pacific Northwest, the West Coast, and the Western States, resulting in a large number of publications on regional and supra-regional patterns of phonetic, lexical, and lexico-grammatical variation (see e.g. [4, 5, 6, 7, 8]).[1] In earlier dialectological work, dialect areas were typically determined on the basis of the co-occurrence in geographical space of lines that separate linguistic features (isoglosses); from these "isogloss bundles", a researcher could divide an area into dialect regions.

While the publications resulting from the American linguistic atlas projects attest remarkable variability in American English, in some ways atlases present a simplified picture of language variation [9]. The use of the isogloss as a conceptual device for the identification of dialect areas suggests categoricity, although few language features occur categorically in a particular area and not at all in another area. Dialect atlases typically attest a single variant for a feature at the place where the corresponding linguistic interview occurred, rather than relative frequencies of several variants of a particular feature. In addition, the situation in which an informant produces an item during a linguistic interview by a fieldworker is not naturalistic, and although some linguistic atlas projects interviewed multiple informants in specific localities, most relied on the responses of a single informant for each locality – the item was then held to be representative for language use in that place in a resulting atlas.

---

[1] Records and materials from most of the American regional linguistic atlas projects are maintained at the University of Georgia by William A. Kretzschmar; some of the audio data has been digitized and made available at http://www.lap.uga.edu/.

Dialect corpora, in contrast, if of sufficient size, will contain frequency information about the use of different variants of a particular feature in a place, allowing regional variation to be assessed on the basis of relative frequencies of competing forms – a model that may better line up with data from perceptual salience studies and thus better represent the actual language situation [10, 11].

For these reasons, a corpus-based approach to the analysis of regional language variation has been advocated, and several recent studies have utilized corpora in order to analyze aggregate regional language variation [11, 12, 13, 3].

A corpus of transcribed American speech with regional coverage represents a desideratum, but significant time and resources are necessary for the manual transcription of large amounts of audio or audiovisual data. For example, the Sociolinguistic Archive and Analysis Project [14], a digitized archive of more than 4,400 sociolinguistic interviews, many of which are associated with specific places, is only approximately 5% transcribed. Likewise, digitized recordings of fieldworker interviews of informants for the *Linguistic Atlas of the Gulf States* and other regional linguistic atlas projects have not yet been extensively transcribed.

2.2 Automatic Speech Recognition and studies of regional language on YouTube and YouTube captions

Significant advances have been made in recent years in the field of automatic speech recognition using neural network-based approaches [15, 16], with some recent system architectures reporting accuracy comparable to that of human transcribers in word error rate (i.e. the proportion of words incorrectly transcribed in a given audio file) [17, 18]. Word error rates of speech-to-text architectures from Google, the owner of YouTube, are reported to be in the range of 5–6% for certain types of evaluation tasks [19]. In the context of investigating the effect of speaker gender and regional accent on the accuracy of YouTube's automatically generated captions, Tatman [20] uploaded videos of speakers of American English to YouTube, and found that very low error rates are possible (p. 56). Ziman et al. [21] utilized Google's Speech-to-text service to automatically transcribe lists of words spoken by experimental subjects taking part in a word recall experiment. They found that the service offers high accuracy in terms of word identification and word onset times.

Linguistic analyses of YouTube videos with regional language content have mainly been qualitative in nature and have been conducted on small numbers of videos. For example, Androutsopoulos [22] analyzed selected excerpts from videos with German dialects, and then compared two videos returned from a search for *Berlinerisch* ('Berlin dialect'). The study offered perspectives on the nature of the relationship between performance of dialect and global mediality, concluding that "representations of dialect are embedded in heteroglossic contrasts within a spectacle" (p. 67).

Some work has used automatically generated YouTube captions. Marrese-Taylor et al. [23] used a neural network to conduct fine-grained aspect extraction and sentiment analysis on captions downloaded from seven YouTube videos (product reviews of a

mobile phone). They found that, compared to written texts, it is more difficult to accurately extract sentiment and aspect features from captions of spoken language, presumably due to the significant differences between speech and writing in terms of relative frequencies of language features.

Tatman [20] analyzed the effect of gender and geographical location on automatic speech-to-text word accuracy. Comparison of automatically-generated captions to manual transcriptions for 62 words showed that the captions were less accurate for Scottish speakers compared to American speakers and for females compared to males.

3 Methods

Using YouTube's API, a script was written to conduct searches for the strings "county of", "city of", "municipal", "town meeting", "city council", "county supervisors", "board of supervisors", and "government" in combination with the names and abbreviations of each of the 50 U.S. states, as well as the string "official government" in combination with the names of the 312 largest municipalities by population and the 100 largest counties by population in the United States. Each search returned a maximum of 50 matches to YouTube channel names. The 1,680 channel matches were manually checked to remove duplicates and false positives, such as channels containing content not associated with state or local organizations in the U.S., channels containing exclusively non-English-language content, or channels which could not be unambiguously assigned to one of the U.S. states or territories.[2]

For each of the 579 channels associated with a state or local civic or governmental organization, the transcript files of all videos in the channel containing automated speech-to-text captions were downloaded in the .vtt file format. Some channels contained just a single video with automatic captions, while others had many -- the channel with the most captions was "City of Murfreesboro, TN - Government", with 1,153 video caption files. In total, 53,743 captions files were downloaded. Speech content was extracted from the transcript files using a script.

Manual examination showed that some transcripts were incoherent – in many cases due to the language of the video (English) being incorrectly identified by the automatic speech-to-text system, for example as Spanish or Dutch. For these videos, the transcripts were typically very short, consisting of only those words extracted from the

[2] YouTube's search API also returns objects in which search terms match the text in the "About" page of individual channels, but the "About" page cannot be searched directly. Thus, a search of YouTube channels for "city of New York" returns (among other channels) "にゅうよくさん" ('Mr. Nyu yoku', https://www.youtube.com/channel/UCGfSGSR1qaYq7G3BkcnU7Hw), a channel devoted to online gaming whose "About" page contains the string "city_of_NewYork". Similarly, a channel search for the string "city government Montana" will return (among other results), a channel for a Dutch field hockey club ("HC Gooische YT", https://www.youtube.com/channel/UCDcQHBcTsDW4UGxiH6HgWA).

audio stream whose phonetic values approximately correspond to segments in the misidentified language. To reduce the signal of these faulty transcriptions in the corpus, 155 transcripts with 20 or fewer words were removed. The resulting transcripts vary in length from 21 words to 50,349 words, for the transcript of a city council meeting for Santa Rosa, California lasting 5 hours and 49 minutes.

The latitude and longitude coordinates for each channel were determined passing the name of channel appended to the name of the state for that channel to a place name API, using geopy [24]. Channels that could be assigned to a specific place with latitude and longitude coordinates were retained.

Transcripts were then aggregated by channel; the 539 channels with at least 1,000 words were retained in the corpus. In total, the corpus comprises 53,675 transcripts from locations in all of the 50 U.S. states, totaling 252,259,141 words. The smallest channel subcorpus is that of the Peoria County, Illinois government, with 1,031 words. The largest is the channel of Rutherford County, Tennessee, with 8,516,795 words. The mean channel word count is 468,013. The corpus was tagged for part of speech using the nltk tagger [25].

State-level aggregation of captions results in subcorpus sizes ranging from 52,911 words (for Hawaii) to 21,897,145 words (for California). The state-aggregated subcorpora are at least 1m words in size for 41 of the 50 U.S. states.[3]

3.1 Description of videos

Many of the videos from which captions were collected are recordings of local government meetings. For example, the video "Bellevue Planning Commission Feb. 22, 2018" is a 1 hour and 17-minute-long video in which the planning commission of Bellevue, Nebraska holds a hearing on a proposed zoning change for land within the city.[4] First, a decision is made by commission members to make use of an electronic voting system within meetings. Then, the commissioners address a request by a city property owner for a zoning change for a 14 acre (5.67 ha) plot of land which would permit commercial development of the plot. A lawyer for the property owner speaks in favor of the zoning change, as does a representative of a real estate development company with plans to develop part of the plot. A member of the local chamber of commerce speaks in favor of the proposed change. Next, several of the commissioners speak about the proposed zoning change, both for and against, and pose questions to the lawyer and the real estate company representative. The video concludes with a vote in which the nine commissioners approve the proposed zoning change by a 7 to 2 margin. The chief commissioner remarks that the approved zoning change "will go to the city council for a hearing".

[3] A tabular presentation of the data, with channel name, channel id, channel state location, latitude and longitude coordinates for channel location, number of video transcripts downloaded, total word count of transcripts, and total speech duration is available at https://github.com/stcoats/YouTube_Corpus/blob/master/YouTube_Channels.csv.

[4] https://www.youtube.com/watch?v=WY9RPeXA3pw

The chief commissioner then closes the hearing, and the video ends. The transcript of the video contains 12,542 words.

Although many of the transcripts record hearings, council sessions, or meetings of local government, other transcripts are from a wide range of genres, such as ceremonies for the presentation of awards to government or civic organization employees, ceremonies for openings of buildings or other infrastructure, interviews with local government representatives (such as mayors), vlog-style videos made by local government representatives or civic organization employees, fire safety tips from county fire departments, videos made to promote tourism, videos profiling pets for adoption at the local animal control center, and many other types of videos. Some videos have many speakers (for example council meetings), and others a single speaker (for example vlog-style videos or videos with a single voice-over speaker). Although the video genres for which transcripts were collected are diverse in terms of their content, it is reasonable to assume that the overwhelming majority of the videos has been locally produced for community purposes, and thus their transcripts can be considered representative of language use in that locality.

3.2 Accuracy of captions

The accuracy of the automatically created transcripts varies, which is unsurprising considering the range of communicative situations in which the videos were made, the large number of individual speakers, and the varying quality of the audio signal, as well as presumable changes in recent years of the implementation of the speech-to-text system used by YouTube. Transcript accuracy was provisionally tested by calculating the word error rate for the first minute of one randomly selected video from 20 channels selected at random. For this sample, error rates ranged from 1.2% to 53.4%, with a mean error rate of 18.6%. While the rate is higher than typical word error rates for orthographic transcriptions, and the material would therefore be unsuitable for certain types of investigations (e.g. regional comparisons of the relative frequencies of rare lexical items), the size of the corpus ensures that it will be useful for many types of analysis.

For the investigation of grammatical variables that are manifest in several forms with competing variants, for example, the corpus is expected to produce useful results. First, due to the large number of transcripts in the corpus and the size of the aggregated channel subcorpora, it is expected that for most lexico-grammatical features, enough instances will be recognized for features to be compared geographically, even if not all instances of a particular variant are identified due to errors in the transcript. This is because the frequency of a feature within a corpus is a function of corpus size. A minimum channel subcorpus size of 1,000 words and a mean subcorpus size of 435,711 words should ensure a sufficient signal for many kinds of lexical and lexico-grammatical features (see the methodological considerations offered by Grieve [3]). Second, because the subcorpora for most of the channels in the corpus consist of a relatively large number of videos drawn from different genres and recorded under different conditions, it is reasonable to assume that the channel-level word error rate does not differ drastically between channel subcorpora, given the number of transcripts per channel

(the mean number of transcripts per channel is 99.6) and the corresponding diversity of communicative situations, speakers, and audio track qualities for that channel's videos. Nevertheless, more rigorous comparison of the word error rate of the automatically generated captions with manually created transcriptions is planned before the corpus is used for an analysis of lexico-grammatical variation in spoken language.

3.3 Map of channel locations

Figure 1 shows a screenshot of the interactive map with the sizes of the channel sub-corpora. The geographical coverage of the corpus is good for the densely populated eastern seaboard from the Washington, D.C. area to Boston, for most of the Southeast, for the Great Lakes and the Upper Midwest, for Colorado, and for California, but not for a large swath of eastern Montana and the western Dakotas. These are, however, some of the least-densely populated areas of the country.

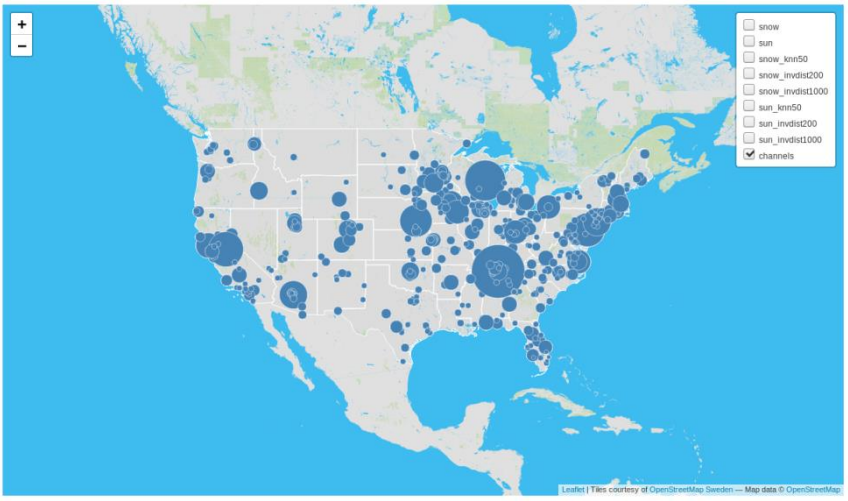

Fig. 1. Screenshot of interactive map showing locations of downloaded channels within the contiguous 48 US. Each circle represents a YouTube channel; circle sizes are proportional to the word count of the aggregated transcripts for that channel.

4 Preliminary analyses

When comparing the relative frequencies of lexical items, a common approach in aggregate analyses has been to calculate the relative frequencies of words with the same

referential meaning [29, 30]. In a similar manner, the frequencies of lexico-grammatical features that have two or more variant realizations (e.g. “do not” and “don’t” for negatives) can be calculated. In order to identify such features and calculate their relative frequencies, it will be necessary to use part-of-speech tags in combination with regular expressions. Of particular interest will be to consider features that have been shown to vary regionally in written American English and investigate the extent to which they may vary in spoken language (cf. [3]).

As an exploratory step, a correlation of geographic distance versus linguistic distance based on pairwise comparison of 504 channels (i.e. 126,756 location pairs) was undertaken. Linguistic distance was calculated using a Euclidean distance metric for the relative frequency of the 100 most frequent lexical items in the corpus, after removal of stopwords such as articles, prepositions, and pronouns. The correlation is positive but weak (Figure 2).

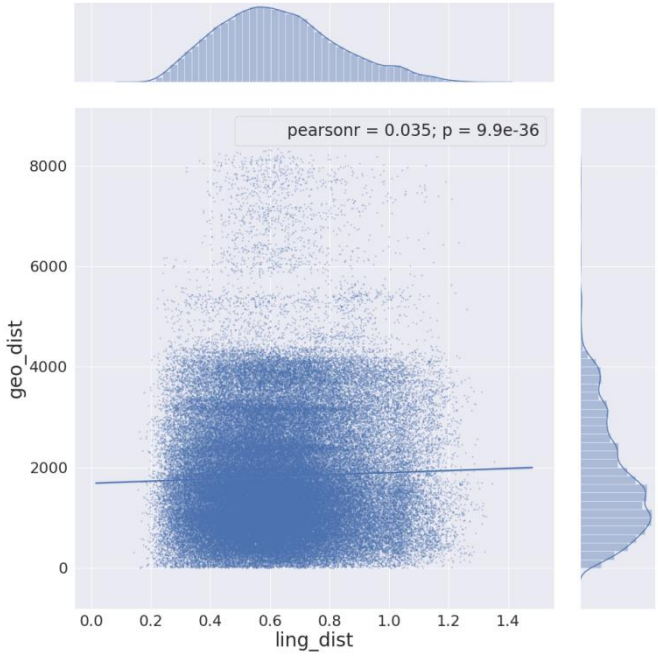

Fig. 2. Linguistic distance versus geographic distance (in km) for the 126,756 location pairs

It is planned that grammatical and lexical variation within the corpus will be analyzed using techniques developed in dialectometry. In dialectometry, linguistic atlases have often been used as data sources in order to analyze aggregate regional patterns in language variation (e.g. [31, 26, 27]; [29] for a dialectometric analysis of American linguistic atlas material). As noted above, however, the frequency-based information from a corpus of regional language may better represent the range of language variation in different parts of the country. Dimensionality reduction techniques such as principal

components analysis, factor analysis, or multidimensional scaling are commonly used in dialectometry. For geographical analysis, statistical techniques such as spatial autocorrelation may be used to assess patterning of the linguistic feature matrix [3, 32, 33].

Continuing with an exploratory analysis of lexical variation, in order to confirm that the frequency information from the corpus can demonstrate regional variation, the relative frequencies of lexical items associated with common weather conditions were subject to local spatial autocorrelation using the Getis-Ord G_i^* statistic [34, 35]. The statistic is a standard deviate indicating whether a given point in a set of spatially distributed values is located in a cluster of high values (positive G_i^* score) or in a cluster of low values (negative G_i^* score). The calculation of G_i^* requires a spatial weights matrix to be defined; in line with recent approaches [3], a binary weights matrix based on the 50 nearest neighbors to each channel location was used.

Following a common mapping procedure in dialectometry [26, 27], a map was created for the channel locations that fall within the boundaries of the contiguous 48 US states using a Voronoi tessellation based on the locations of the individual channels [28]. Mapping the G_i^* scores for the types *snow* (Figure 3) and for *sun* (Figure 4) shows that the use of these lexical items is spatially distributed in the corpus in a manner that is easily interpretable: snow, ice, and other winter-related phenomena are more often mentioned in the Great Lakes region, the Rockies, and New England, where provisions for snow removal need to be made by local governing bodies, and less often discussed in the southern and southeastern parts of the country, where a milder climate means snow and ice rarely occur. Similarly, sun is mentioned in climates where it is more intense: in Florida, the Southwest, and much of the Southeast, sun may be discussed in the context of its relevance for tourism or in public service videos warning of its dangers. The sun is less often discussed in the relatively cloudy and cool Upper Midwest.

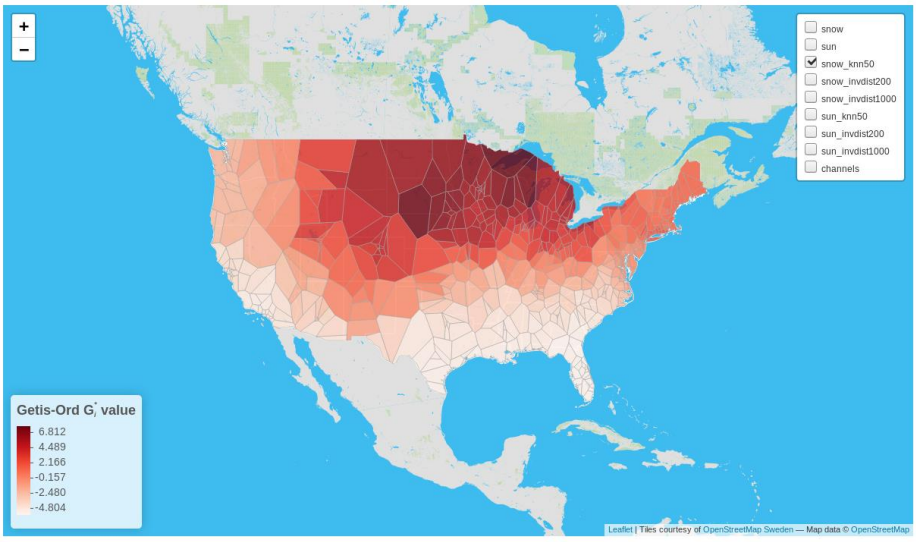

Fig. 3. Getis-Ord G_i^* scores for *snow*, 50-nearest-neighbor binary spatial weighs matrix

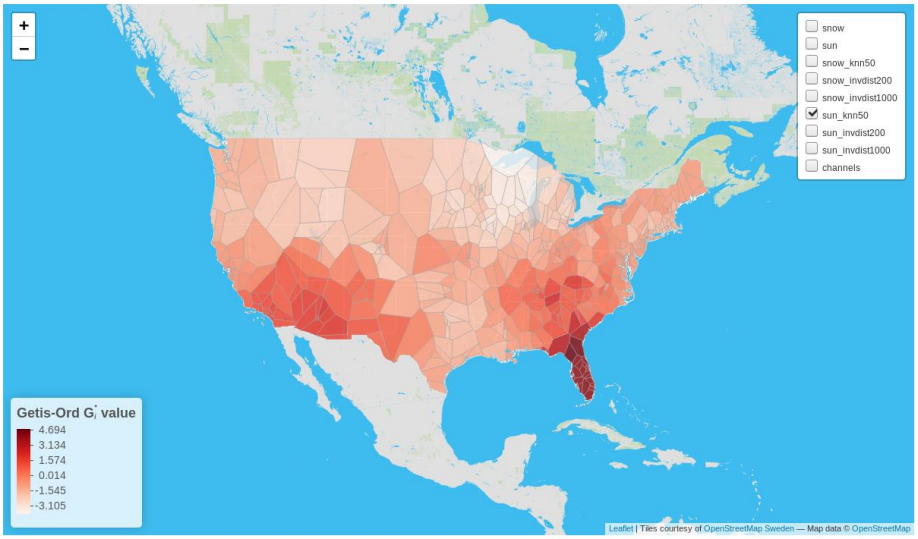

Fig. 4. Getis-Ord G_i^* scores for *sun*, 50-nearest-neighbor binary spatial weighs matrix

It should be noted, however, that Figures 3 and 4 do not present regional lexical variation, but rather demonstrate that lexical items may be distributed geographically in the corpus.

5 Summary and future outlook

This paper has described the methods used to create a large corpus of transcribed American English speech from the automatically generated captions of YouTube channels of local government and civic organizations in the United States. Such a corpus may prove to be a useful resource for the analysis of regional variation in spoken American English, especially compared to some data from traditional linguistic atlas projects, which for some features may present a simplification of the range of variation in a particular place. It is planned that the corpus will enable a geographical analysis of lexical and grammatical variation in speech, allowing patterns of regional variation to be identified and, if present, to be compared to patterns of regional lexical and grammatical variation that have been found in linguistic atlas data and in corpora of written language.

Although the corpus consists of orthographic transcripts and is thus only suitable for analyses of lexical and grammatical variation, the data collection pipeline could also be utilized for the automatic download of the videos for which the transcripts were created; the audio tracks of the video files could then be analyzed in terms of phonetic or prosodic variation such as formant frequencies, speech rate, or other features.

Future work with the corpus will involve more rigorous testing of the accuracy of the automatically created transcripts and the preparation of scripts for the extraction of lexico-grammatical variants in the corpus. It is planned that the geographical analysis

of variable relative frequencies will involve aggregation techniques common in dialectometry as well as statistical tools employed in geographical analysis such as spatial autocorrelation.

References

1. Biber, D.: Variation across speech and writing. Cambridge University Press, Cambridge, UK (1988).
2. Biber, D.: Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press, Cambridge, UK (1995).
3. Grieve, J.: Regional variation in written American English. Cambridge University Press, Cambridge, UK (2016).
4. Kurath, H., Hansen, L., Bloch, B., Bloch, J.: Linguistic atlas of New England (3 vols.). Brown University Press, Providence, RI (1939–1943; reprinted 1972).
5. McDavid, R., O'Cain, T.: Linguistic atlas of the Middle and South Atlantic States (2 fascicles published before discontinued). University of Chicago Press, Chicago, IL (1980).
6. Kretzschmar, W. A., McDavid, V., Lerud, T., Johnson, E.: Handbook of the linguistic atlas of the Middle and South Atlantic States. University of Chicago Press, Chicago, IL (1993).
7. Pederson, L., McDaniel, S. L., Adams, C. M.: Linguistic Atlas of the Gulf States (7 vols.). University of Georgia Press, Athens, GA (1986–1993).
8. Kretzschmar, W. A.: Mapping Southern English. American Speech 78, 130–149 (2003).
9. Kretzschmar, W. A.: The linguistics of speech. Cambridge University Press, Cambridge, UK (2009).
10. Nerbonne, J.: Data-driven dialectology. *Language and Linguistics Compass* 3(1), 175–198 (2009).
11. Szmrecsanyi, B.: Corpus-based dialectometry: a methodological sketch. Corpora 6(1), 45–76 (2011).
12. Szmrecsanyi, B.: Grammatical variation in British English dialects: A study in corpus-based dialectometry. Cambridge University Press, Cambridge, UK (2013).
13. Szmrecsanyi, B.: Forests, trees, corpora, and dialect grammars. In: Szmrecsanyi, B., Wälchli, B. (eds.), Aggregating Dialectology, Typology, and Register Analysis, pp. 89–112. De Gruyter, Berlin/Boston (2014).
14. Kendall, T.: Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. University of Pennsylvania Working Papers in Linguistics 13(2), 15–26 (2007).
15. Liao, H., McDermott, E., Senior, A.: Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 368–373 (2013).
16. Sainath, T., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4580–4584 (2015).
17. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner,

C., Gao, L., Gong, C., Hannun, A., Han, T., Lappi J., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Zhijian, Wang, Zhiqian, Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama,D., Yuan, B., Zhan, J., ZhuZ.: Deep Speech 2: End-to-end speech recognition in English and Mandarin. In: Balcan, M., Weinberger, K. (eds.) Proceedings of the 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research vol. 48, pp. 173–182 (2016).

18. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, F., Stocke, A., Yu, D., Zweig, G.: Toward human parity in conversational speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(12), 2410–2423 (2017).
19. Chiu, C.-C., Sainath, T., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., Bacchiani, M.: State-of-the-art speech recognition with sequence-to-sequence models. arXiv:1712.01769v6 [cs.CL] (2018).
20. Tatman, R.: Gender and dialect bias in YouTube's automatic captions. In: Proceedings of the First Workshop on Ethics in Natural Language Processing, pp. 53–59 (2017).
21. Ziman, K., Heusser, A., Fitzpatrick, P., Field, C., Manning, J.: Is automatic speech-to-text transcription ready for use in psychological experiments? In: Behavior Research Methods 50, 2597–2605 (2018).
22. Androutsopoulos, J.: Participatory culture and metalinguistic discourse: Performing and negotiating German dialects on YouTube. In: Tannen, D., Trester, A. M. (eds.), Discourse 2.0: Language and New Media, pp. 47–72. Georgetown University Press, Washington, DC (2013).
23. Marrese-Taylor, E., Balazs, J. A., Matsuo, Y.: Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Copenhagen, Denmark, September 7–11, 2017, pp. 102–111. Association for Computational Linguistics, Stroudsburg, Pennsylvania (2017).
24. Esmukov, K. et al.: Geopy. https://github.com/geopy/geopy (2018).
25. Bird, S., Loper, E., Klein, E.: Natural language processing with Python. O'Reilly, Newton, MA (2009).
26. Goebl, H.: Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF ("Dialectometric studies: On the basis of italo-romance, rhaeto-romance, and gallo-romance language material from the AIS and the ALF"). Tübingen: Niemeyer (1984).
27. Goebl, H.: Recent advances in Salzburg dialectometry. Literary and Linguistic Computing 21: 411–435 (2006).
28. Voronoi, G.: Nouvelles applications des paramètres continus à la théorie des formes quadratiques ("New applications for continuous parameters in the theory of quadratic forms"). Journal für die Reine und Angewandte Mathematik 133, 97–178. (1907).
29. Nerbonne, J., Kleiweg, P.: Lexical distance in LAMSAS. In: Nerbonne, J. and Kretzschmar, W. (eds.) Computational methods in dialectometry. Special issue of Computers and the Humanities, 37(3), 339–357 (2003).
30. Heeringa, W., Hinskens, F.: Convergence between dialect varieties and dialect groups in the Dutch language area. In: Szmrecsanyi, B., Wälchli, B. (eds.), Aggregating Dialectology, Typology, and Register Analysis, pp. 27–52. De Gruyter, Berlin/Boston (2014).

31. Séguy, J.: La relation entre la distance spatiale et la distance lexicale (“The relation between spatial and lexical distance”). Revue de Linguistique Romane 35, 335–57 (1971).
32. Grieve, J.: A statistical comparison of regional phonetic and lexical variation in American English. Literary and Linguistic Computing 28, 82–107.
33. Grieve, J.: A comparison of statistical methods for the aggregation of regional linguistic variation. In: Szmrecsanyi, B., Wälchli, B. (eds.), Aggregating Dialectology, Typology, and Register Analysis, pp. 53–88. De Gruyter, Berlin/Boston (2014).
34. Getis, A., Ord, J. K.: The analysis of spatial association by use of distance statistics. Geographical Analysis 24(7), 189–206 (1992).
35. Ord, J. K., Getis, A.: Local spatial autocorrelation statistics: Distributional issues and application. Geographical Analysis 27(4), 286–306 (1995).