Accepted manuscript for Coats, Steven, Articulation rate in American English in a corpus of YouTube videos, *Language and Speech* 63(4), 799–831. Copyright © 2019, Steven Coats. https://doi.org/10.1177/0023830919894720

Articulation rate in American English in a corpus of YouTube videos

Steven Coats

University of Oulu, Finland

Abstract

Previous studies of the temporal organization of speech in American English have found differences in speaking or articulation rate according to speaker dialect or location, but small sample sizes and incomplete geographic coverage have limited the generalizability of the findings. In this study, articulation rates in American English are calculated from the automatic speech-to-text transcripts of more than 29,000 hours of video from local government and civic organization channels on YouTube from the 48 contiguous U.S. states, containing more than 230 million individual word timings. Two questions are considered: Are there regional differences in articulation rate? And do urban speakers articulate faster than rural speakers? The study presents several methodological innovations: First, it identifies a genre of regional speech suitable for interregional comparisons (meetings of local governments or civic organizations). Second, it introduces a new method for the calculation of articulation rate using cue and word timestamps from captions files. Third, it leverages US Census data in order to correlate articulation rate with population for a large number of localities. The study shows that, in line with previous studies, Southerners articulate slower, and Americans from the Upper Midwest more quickly. In addition, there is a small but positive correlation between population size and articulation rate. Articulation rates are mapped using a measure of local autocorrelation.

1. Introduction

Do Americans from Southern states speak with a characteristic slow drawl that reflects their unhurried, down-to-earth attitudes? Do New Yorkers or inhabitants of other large American cities speak with a rapidity corresponding to the hectic pace of life in bustling metropolises? In recent years, research in sociophonetics and sociolinguistics has begun to systematically investigate the correlations between prosodic features and

traits associated with demographic or social identity, including speaker dialect as it is manifest in geographical location. Among the prosodic features that have been studied, variation in temporal organization has figured prominently—perhaps because speaking rate and articulation rate are thought to be psychologically salient features, apparent to interlocutors in conversation, but perhaps also because variation in temporal organization of speech can serve as the basis for linguistic stereotypes, such as that of the slow-speaking rural person versus the fast-talking city dweller, or the slow-speaking inhabitant of region A versus the fast-speaking inhabitant of region B. Roach (1998) noted that rural speakers of English, whether in the United States or the United Kingdom, are believed to speak slowly, whereas "urban accents such as those of London or New York are more often thought of as fast-speaking" (p. 150). In a study of dialect perceptions, Preston found that the distinction slow/fast was among the most frequent labels applied by respondents when they were asked to label regions of the United States based on the speech characteristics of the region's inhabitants (1999, p. 363).

In the United States, although speaker location has been considered as a variable in several recent studies of speaking or articulation rate variation, methodological considerations and limited sample sizes have made it difficult to infer patterns for the country as a whole. In addition, prosodic features pertaining to temporal organization such as stress timing, pause location and duration, or articulation rate are affected by a large number of interlocutor and situational factors, complicating efforts to disentangle the relationship between temporal variation and parameters of speaker identity and dialect. Rural-urban differences in articulation rate in American English have not been directly investigated in previous research.

In this study, a new methodological approach has been developed for the investigation of articulation rate, based on a corpus compiled from automatically-generated captions files of YouTube videos from the United States. The videos whose captions files are included in the corpus are from a range of spoken genres, communicative situations, and speaker configurations, but mostly consist of public meetings of local government or civic organizations. Articulation rates are calculated in aggregate for single videos, not for

individual speakers. The large number of recording transcripts in the corpus (48,945) permits a geographical analysis at a relatively fine level of granularity.

The text is organized as follows: In Section 2, a review is provided of some previous research on speech/articulation rate, automatic speech-to-text transcription, YouTube captions files, and the use of spatial statistics from geography for the study of language variation. In Section 3, the methods used to collect the data and assign each video to a location are described and the decision to focus on YouTube channels of local and state governments or civic organizations is discussed. In addition, a short description of one of the videos collected in the corpus–a school board candidate's forum in Tennessee–is provided. The section closes with a description of the method used to calculate aggregate articulation rates per video from captions files. In Section 4, the results of the study are presented: The geographical distribution of articulation rate differences in the U.S. is considered using two autocorrelation measures from spatial geography (Moran's I and the Getis-Ord G_i^*statistic); the values of the latter are shown on maps. Then, articulation rate is correlated with population size. Section 5 discusses caveats and possible interpretations of the results, and Section 6 summarizes the study and presents the outlook for future work with the corpus.

2. Previous work

Studies of the temporal organization of speech have utilized different measurement constructs, conducted measurements on different types of speech, and considered associations between temporal organization and various other factors, such as language used, demographic or social identity, or regional location.

In an early study, Goldman-Eisler (1961) found that the length and distribution of pauses within an utterance are variable, and therefore distinguished between the *speaking rate* and the *articulation rate*: The former is defined as units of speech such as phones, words or syllables divided by total utterance time, including pauses, while the latter omits pauses. Slow rates of articulation, when measured, will correspond to slow rates of speaking, but the inverse is not necessarily true: A slow speaking rate could result from rapidly articulated utterances, combined with frequent, lengthy pauses. Because pause duration itself is

known to be affected by individual and contextual factors (Kendall, 2013), the speaking rate often shows more variability than the articulation rate. In much of the recent research literature on the temporal organization of speech, in order not to conflate distinct components of the speech signal, the tendency has been to report articulation rate, rather than speaking rate (as well as, in some studies, pause duration). In line with this, in this study, articulation rate in syllables per second (σ/sec.) is reported, a rate measure which has been shown to be more psychologically salient, compared to some others (Plug & Smith, 2018).

Some research has analyzed speaking or articulation rates calculated from informants' reading of prepared texts, while other studies have been based on spontaneous or conversational speech. Ray and Zahn (1990) analyzed speaking rate in recordings of public and conversational speech by university students in seven American states, but did not find significant differences based on region. Byrd (1992, 1994) analyzed speaking rate, sex, and regional affiliation in eight groups of American English speakers using a subset of the TIMIT corpus, a database of speakers reading prepared sentences (Garofolo et al., 1993). Based on the analysis of two sentences, she found that women speak more slowly than men and Southerners speak more slowly than Northerners in terms of sentence duration, but that the difference may be due to more frequent pauses by southern speakers, not a difference in articulation rate (1994, p. 44). She further noted that "the geographical definitions used in TIMIT appear to be too broad for many linguistic purposes" (p. 52), and that regional differences in speaking rate may be artefacts of differences according to speaker sex. For the United Kingdom, few studies have considered speaking or articulation rate as a function of geographical location. Hewlett and Rendall (1998) reported conversational speech to be faster than reading, but found no significant differences in articulation rate between 12 urban informants from Edinburgh, Scotland and 12 rural informants from the Orkney Islands.

In a corpus of read speech from 60 undergraduate university students from different regions of the United States (Clopper & Pisoni, 2006), Clopper and Smiljanic (2011) reported higher pause frequency and longer pause duration for American speakers from the South compared to speakers from other regions. In a more detailed examination of the components of temporal variability in speech using the same corpus,

Clopper and Smiljanic (2015) measured articulation rate, pause frequency, pause duration, and the relative duration of consonant and vowel intervals (and derived measures) in two read passages from the 60 speakers, comprising in total one hour of speech. They reported articulation rates ranging from 5.36–5.73 σ/sec., with New England speakers articulating the fastest and Southern and Midland speakers the slowest. In a study of vowel duration in North American English dialects based on a sample of data from the *Atlas of North American English* (Labov, Ash & Boberg, 2006), Tauberer and Evanini (2009) found longer vowel durations for Southerners compared to speakers from other regions of the country. Although the data used in the study consisted solely of extracted vowels and hence was not suitable for the calculation of speaking or articulation rates, the authors also reported that they calculated speaking rates from the Fisher Corpus (Cieri, Miller, & Walker, 2004) and found no difference between Northern and Southern speakers. Jacewicz, Fox, O'Neill, and Salmons (2009) recorded 94 male and female speakers from two age groups (young adults aged 20–34 and older adults aged 51–65) from Wisconsin and North Carolina reading a set of sentences and speaking spontaneously for ten to fifteen minutes. The authors found that Wisconsin speakers had an articulation rate 12.5% higher than did North Carolina speakers, and males spoke faster than females, although the effect size was quite small (p. 244). They additionally noted the characteristic reticence of Americans from the Upper Midwest, writing that the "North Carolina… speakers clearly enjoyed sharing stories from their lives and mostly did not require prompting… Wisconsin speakers ran out of topics more often and needed a leading question when they stopped talking" (p. 241). Jacewicz, Fox, and Wei (2010) continued to analyze Wisconsin and North Carolina speakers by using a mixed-effects model on a slightly larger number of speakers, with similar findings.

Kendall (2013) analyzed articulation rate and pauses in two sets of data: one consisting of read passages, and the other of conversations conducted in the context of sociolinguistic interviews. In the first part of the analysis, the articulation rate of 42 young adults from Memphis, Tennessee, Oswego, New York, and Reno, Nevada was analyzed on the basis of recordings of a 266-word text read aloud. Articulation rate was found to vary from 4.28–4.97 σ/sec., with a mean of 4.44 σ/sec. (pp. 65, 63), and the Nevada speakers had the highest rate, followed by the New York State speakers and then the Tennessee speakers. Agreeing

with Jacewicz et al. (2010), Kendall noted that reading passage data may not be ideal for the investigation of speech timing, among other reasons because informants can differ in their reading ability (p. 81). In the second part of the study, a mixed-effects model was used to analyze articulation rate as a function of sex, age, region, and ethnicity in timed transcripts of sociolinguistic interviews with 159 informants from North Carolina, Ohio, Texas, and Washington, D.C. For the approximately 30,000 utterances in the data (i.e. segments of unbroken speech by a single speaker, excluding pauses of longer than 200 ms), the mean articulation rate in conversational speech was found to be 4.6 σ/sec. (p. 92). It was found that persons of European background spoke slightly faster than African Americans, Latinos, or Lumbee Native Americans, males spoke slightly faster than females, and persons aged 19–66 spoke slightly faster than persons younger or older. In terms of regional differences, no clear regional pattern was found: The speakers from Texas, Ohio, Southern North Carolina and Eastern North Carolina spoke slightly faster than speakers from Western North Carolina, Washington, D.C., or Central North Carolina (p. 91), and Kendall noted that "speech rates appear to vary as much within a single region… as they do across regions" (p. 210). Table 1 summarizes recent work on regional speaking rate and/or articulation rate variation in the United States.

Table 1: Summary of some recent work on regional speaking/articulation rate variation in the U.S.

	Size of sample	Locations	Type of speech	Measurement	Observed trends	Additional notes
Ray & Zahn (1990)	93 speakers, two-minute samples (3.1 hours of speech)	Single locations in Washington, Oregon, Texas, Louisiana, Wisconsin, Ohio, Utah	Public speaking and conversation from undergraduate university classes	Speaking rate	No significant differences	Conversational speech is faster than public speaking
Byrd (1992, 1994)	630 speakers, two sentences per speaker (~1 hour of speech)	New England, North, North Midland, South Midland, South, New York City, West, "Army Brat (moved around)". Exact locations not provided	Read sentences (laboratory environment)	Speaking rate	"Army Brat" > Northeast > North Midland > West > north, NY City > South Midland > South	Males > females
Jacewicz et al. (2009); Jacewicz, Fox & Wei (2010)	94 speakers, 120 read sentences per speaker + 4,930 phrases of five or more syllables	Single locations in Wisconsin and North Carolina	Read sentences and 10-15 minutes of conversation per speaker (laboratory environment)	Articulation rate	Wisconsin > N. Carolina	Mean articulation rate 5.12 σ/sec.

	without a pause					
Clopper & Smiljanic (2011, 2015)	60 speakers, two reading passages	Ten students enrolled at Indiana University from New England, the Mid-Atlantic, the North, the Midland, the South, and the West	Read passages (laboratory environment)	Articulation rate (pause frequency and duration also analyzed)	New England > Mid-Atlantic > North > West > South > Midland	Mean articulation rate 5.53 σ/sec. Southern speakers also have longer pause durations
Kendall (2013): Reading passages	42 speakers, 266-word text	14 speakers from each of the following locations: Memphis, TN, Oswego, NY, Reno, NV	Read passages	Articulation rate (pause frequency and duration also analyzed)	Nevada > New York > Tennessee	mean articulation rate 4.44 σ/sec.
Kendall (2013): Sociolinguistic interviews	159 speakers, ~40 hours of speech	Single locations in Ohio and Texas, Washington DC, four locations in North Carolina	Excerpts from sociolinguistic interviews	Articulation rate (pause frequency and duration also measured)	Ohio > Southern N.C. > Eastern N.C. > Texas > Western N.C. > Washington, D.C. > Central N.C.	mean articulation rate 4.6 σ/sec.

Speech and articulation rate variation according to dialect and/or location has been investigated in other languages. French speakers from France articulate more quickly than do Belgian or Swiss French speakers (Avanzi, Obin, Bardiaux, & Bordal, 2012; Avanzi, Dubosson, & Schwab, 2012). Verhoeven, De Pauw, and Kloots (2004) compared Dutch speakers from the Netherlands and Belgium, finding higher articulation rates for Netherlands speakers. Hahn and Siebenhaar (2016) analyzed articulation rate in German in recordings of reading passages by speakers in 67 localities in Germany, Austria, and Switzerland, and found that values generally increased from Northern Germany towards Southern Germany, Austria, and Switzerland. Leemann (2017) used a mobile telephone app to record 3,000 Swiss German speakers from 452 Swiss localities speaking 16 words, then analyzed the variation in the length of time between vowel onsets in six disyllabic words (*Abend, Augen, fragen, Donnerstag*, *heben*, and *trinken*). He found a regional pattern that corresponds to some previous results from studies of speaking or articulation rate in Swiss German (Bern speakers articulate more slowly), but acknowledged that inferring conversational articulation rates from vowel onset times for isolated words spoken in a non-naturalistic context may not be reliable. For the closely-

related Scandinavian languages of Norwegian, Danish, and Swedish, Hilton, Gooskens, and Schüppert (2011) reported higher articulation rates for Danish compared to Norwegian and Swedish, based on an analysis of recordings of news broadcasts.

Interlocutor and contextual factors have been shown to affect speaking or articulation rate in conversation. Yuan, Liberman, and Cieri (2006) analyzed speaking rate in recorded telephone conversations in English and Mandarin Chinese, and found that people who know each other tend to speak slightly faster than strangers. The topic under discussion can also affect rate: "important and unpredictable portions [of a conversation] are spoken at a relatively slower rate" (2006, p. 3). In addition, longer utterances and segments in the middle of an utterance are spoken more quickly, compared to shorter utterances and utterance-final segments (Oller, 1973; Yuan et al., 2006). As well as being affected by factors such as demographic, social, or regional identity, psychological state, interlocutor familiarity, topic under discussion, and utterance-internal considerations, a speaker's articulation may vary according to anatomical, physiological, or neurological parameters. Experimental studies have shown that the ratio between fastest possible articulation rate and normal articulation rate is relatively stable across speakers (Tsao & Weismer, 1997; Tsao, Weismer & Iqbal, 2006), suggesting that neuromuscular constraints on speech timing processes, and thus ultimately biological or genetic factors, may also contribute to articulation rate.

2.1 YouTube captions files

Captions are text, representing the spoken language in a video, that appears at the bottom of a screen, synchronized to the audio signal. In 2009 YouTube began to provide captions generated automatically by Google's speech-to-text module for some videos (Google, 2009), and in recent years, the accuracy of neural-network based speech-to-text transcription models has increased significantly (Chiu et al., 2017; Liao, McDermott, & Senior, 2013; Sainath, Vinyals, Senior, & Sak 2015). Word error rates for some speech-to-text architectures are now in the 5–6% range for certain evaluation tasks, comparable with error rates of human transcribers (Xiong et al., 2017). YouTube's automatically generated speech-to-text transcripts are force-

aligned to the audio track; although the technical details of the procedures used for alignment have not been made public, the system's components are summarized in a patent filing (Harrenstien, Toliver, Alberti, & Black-Bilodeau, 2009). The accuracy of Google's automatic speech-to-text service has been evaluated in a few studies (Tatman, 2017; Ziman, Heusser, Fitzpatrick, Field & Manning, 2018), but as far as is known, caption file word timings have not yet been used for the study of speech timing phenomena.

2.2 Spatial analysis of language features

While several analyses have considered regional differences in articulation or speaking rate in American English, most of the studies have not been undertaken on data with geospatial granularity suitable for an analysis using the techniques of geographical statistics. Rather, articulation or speaking rates of speakers from different regions have been compared without a formal treatment of their geographical location. For example, in Byrd's (1994) analysis of speaking rate differences for speakers from seven different American regions, specific locations were not provided, and there was "no statement on the part of the database designers as to the motivation for establishing these particular dialect regions" (p. 43). Jacewicz et al. (2009) and Jacewicz et al. (2010) posed the question "is there a systematic dialectal difference in speech tempo between northern and southern regions?" (2010, p. 840), but their analyses were conducted on a small number of speakers from only two US locations (Madison, Wisconsin and three adjacent counties in North Carolina). Similarly, Kendall (2013) analyzed conversational data drawn from three US states and the District of Columbia.

Because the corpus data used in this study is drawn from 506 locations within all of the contiguous 48 U.S. states, the spatial analysis presented in Section 5 employs statistical methods of spatial autocorrelation. Spatial autocorrelation has been employed for the study of the spatial distributions of phonetic, lexical, and grammatical features, most notably by Jack Grieve and colleagues (Grieve, 2011, 2012, 2014, 2016; Grieve, Speelman & Geeraerts, 2011), but as far as is known, not yet in analyses of temporal

variation in speech. In the following, a brief summary of the treatment of geographical variation in dialectological data is presented.

Spatial analysis in dialectology has its roots in the linguistic atlases of the 19th century. Regional patterns in language were often marked on maps by drawing isoglosses, or lines that separate variants of a linguistic feature; dialect regions could then be identified on the basis of co-occurrence of isoglosses or isogloss bundles (Kretzschmar, McDavid, Lerud & Johnson 1993; Kurath, Hansen, Bloch & Bloch, 1972; McDavid & Cain, 1980; Pederson, McDaniel & Adams, 1986–93; Wenker, 1878). While the isogloss method often resulted in visually compelling and easily interpretable representations of regional language variation, the isogloss as a conceptual tool suffers from the deficiency of suggesting categoricity where it may not be the case: An isogloss implies that a language feature is used categorically in one place, and not at all in some other place. In addition, because the identification of dialect regions based on isogloss bundles relied at least in part on analyst intuition, the method was not necessarily replicable.

In recent decades, more objective statistical techniques from geography have been introduced in the study of dialect data in order to identify spatial patterns of variation and conduct analyses on spatially distributed language data. Spatial autocorrelation is a statistical technique for the identification of patterns in spatial data. Lee and Kretzschmar (1993) analyzed the spatial distribution of selected lexical items from the *Linguistic Atlas of the Middle and South Atlantic States* by using a joint count statistic of shared edges between polygons for categorical occurrence/non-occurrence of an item at a particular location. They demonstrated that a number of lexical items – those associated with regional dialects – exhibited clustering. A joint count statistic, however, is only suitable for analysis of categorical data, not for continuous variables such as articulation rate. Grieve, Speelman and Geeraerts (2011) described the compilation of a corpus of letters to the editor from American daily newspapers and analyzed the spatial distribution of some lexico-grammatical features using the spatial autocorrelation statistics Moran's global I (Moran, 1950) and Getis-Ord local G_i and G_i^* (Getis & Ord, 1992; Ord & Getis, 1995). Factor analysis conducted on spatial autocorrelation statistics showed that regional patterns of variation in the lexico-grammar of this genre of

writing corresponded, for the most part, to patterns of regional variation proposed for American English by earlier researchers. Similar techniques were employed to analyze the spatial distribution of contraction and of adverbial position in written American English (Grieve, 2011, 2012); regional variation in vowel quality in data from the *Atlas of North American English* (Labov et al., 2006) was subject to spatial autocorrelation analysis in Grieve (2014). Grieve (2016), a more detailed treatment of an expanded version of the corpus from Grieve et al. (2011), documented regional patterns of lexico-grammatical variation in detail and suggested some motivations for the differences found. In accord with these highly fruitful recent approaches, in the present study, two spatial autocorrelation statistics are employed in order to assess regional variation in articulation rate in American English: Moran's global I and the Getis-Ord G_i^* statistic. The statistics are introduced in Section 3.7.

3. Data and Methods

3.1 Decision to focus on YouTube channels of state and local government or civic organizations

Recent work in dialectology has emphasized the importance of corpus approaches to the study of regional language variation, based on the fact that with more data, one is more likely to detect legitimate regional patterns that are manifest in the relative frequencies of competing linguistic forms (Grieve, 2016; Nerbonne, 2009; Szmrecsanyi, 2011, 2013). For this study, the decision to focus on YouTube channels of state or local governments and civic organizations was motivated in part by this consideration: A relatively large number of local governments in the United States maintain a YouTube channel in order to provide citizens with access to information, services, and records of decision-making processes, and many channels feature a large number of videos. In addition to the reliability advantages implicit in large corpus size, there are benefits to working with captions files of video recordings of meetings of local governments or civic organizations in terms of the representativeness of the sample. First, it can be safely assumed that the majority of the persons recorded in videos of local government meetings are residents of the communities those channels were created to represent, such as local government council members or employees, members of local civic

organizations, or citizens bringing requests or posing questions to local government bodies. Many holders of public office in American municipalities are legally required to be residents of the area they represent (Mazo, 2016). Videos returned by YouTube's API (Application Programming Interface) on the basis of searches for place names alone, in contrast, often consist of content about a particular place, which may or may not include speech of local residents.[1] The locations of residence of the speakers in the corpus can be inferred from their participation in the affairs of local government in a particular place, and it seems probable that most elected officials of local governments in the United States are long-term residents, rather than new arrivals, simply because it takes time to establish the social contacts necessary for election to public office time (Buren & McHugh, 1992). Nevertheless, it cannot be guaranteed that all speakers in the videos are lifelong permanent residents of the places associated with the YouTube channels in the corpus. Geographical mobility is a fact of American life, and it may be the case that some speakers in the corpus have moved, either from other locations in the United States or from other countries, and that the prosodic qualities of their speech, including articulation rate, bear traces of their former places of residence. A detailed residence history of the speakers in the corpus may be a desideratum when investigating the territorial extent of traditional dialects or looking at changes in the areal distribution of language features over time, but as noted by Grieve (2016, p. 23), a corpus that excludes speakers based on prior residence history will not give an accurate snapshot of contemporary English use in the United States. Newly-arrived persons in a particular place contribute to the speech fabric of that community, and there is no reason to exclude such persons from the corpus if the goal is a record of contemporary language use in different locations. By focusing on videos of local government, the corpus used in this study helps to ensure a representative sample in terms of the speech patterns, including articulation/speaking rate, for locations within the United States.

Second, while the content of local government and civic organization YouTube channels is diverse, and typically includes videos from various speech genres and with a range of communicative configurations, a substantial proportion of the videos for which captions files were downloaded are directly comparable in terms of their communicative parameters: They are recordings of local government meetings. The meetings typically consist of structured group discussion, in the form of sequential individual utterances by a relatively

small number of people (circa 5 to 10), and are comparable in terms of register and formality. In a typical video of such a meeting, the mayor or chairperson of the body constituting the meeting speaks somewhat more than other speakers, but not always (see the description of a typical video in Section 3.4, below). Because they regularly attend local government sessions, council members or other representatives know each other, and because for the most part, the topical concerns of local government in the U.S. are predictable, the extent to which discussion topic may affect articulation rate is limited. Zoning regulations are discussed in town meetings in Massachusetts, in county council sessions in Iowa, or in city planning meetings in California. For these videos, the representativeness in terms of the regional affiliation of the speakers, the parallels in the communicative configurations and other contextual parameters of the recorded interactions, and the similarity in topics under discussion all help to ensure the comparability of the data in the captions files for the analysis language features such as articulation rate.

3.2 Data Collection

Scripts were written to access YouTube's API and download automatically-generated captions files from channels of local government or civic organizations in the United States. Channels of interest were identified by recursively passing searches to the API with regular expressions combining the substrings "county of", "city of", "municipal", "town meeting", "city council", "county supervisors", "board of supervisors", and "government" with the names and abbreviations for each of the 50 U.S. states (e.g. "municipal Arkansas", "town meeting New Mexico", "city council CA", etc.). In addition, the names of the 312 largest municipalities and the 100 largest counties by population in the United States were combined with the name or abbreviation of their states and the substring "official government" (e.g. "Los Angeles, CA official government", "Cook County, Illinois official government", etc.).[2] The procedure returned 1,680 channels, many of which were duplicates or false positives: For example, the search "city council CA", in addition to channels for cities in California, returns channels of Canadian municipalities. Because YouTube's search algorithm matches not only text in a channel name, but also text that appears in the titles of individual videos

in a channel or on a channel's "About" page, many channels were returned that had nothing to do with the specific place name, its local government, or civic organizations. These false positives were removed manually from the list of returned channels after checking the channel content, as were channels that could not be unambiguously assigned to a single U.S. state. In total, 579 channels were retained for corpus creation. Areas of the United States that are densely populated, such as the Eastern Seaboard and California, are well represented in the corpus. Few channels were found from less-densely populated areas, such as Montana or Wyoming.

In the next step, all available English-language automatically-generated captions files were downloaded from the 579 channels in .vtt format, using YouTube's API and youtube-dl (Hsuan, Amine & M., 2018). Some channels contained just a single video with automatically-generated captions, while others had many. In total, 53,743 unique captions files were downloaded. The text of the captions files (i.e. the speech transcript) and the individual word timings were then extracted using an additional script (see Section 3.5 below); non-speech content within word timing tags in the files (i.e. the automatically-generated content "[Applause]", "[Laughter]", and "[Music]") was removed.

The extracted texts of some captions files were extremely short. Many of these short texts were in a language other than English, despite captions file metadata labels, presumably due to misidentification of the language of the video by Google's speech-to-text algorithm, for unknown reasons. In many of these short captions files, the phonetic shape of a small number of English-language word segments from the video's audio approximately corresponded to words in the incorrectly identified language, which, upon text extraction, resulted in very short, incoherent texts in Spanish, Dutch, Swedish, or some other language. To remove these files, all extracted texts with 20 words or less (155 texts) were removed. The texts of the 53,588 retained files varied in length from 21 words to 50,349 words. The total size of this preliminary corpus was 252,277,053 words; the corpus was further processed to delimit its geographical scope to the 48 contiguous US states and to remove pauses (Sections 3.3 and 3.5, below). The corpus is described in more detail in Coats (2019).

3.3 Geolocation

The latitude and longitude coordinates for each channel were determined by passing the channel name appended to the name of the state for that channel to a geocoder API, using *geopy* (Esmukov et al., 2018). Captions file texts were then aggregated by channel; channels that could be assigned to a specific place with latitude and longitude coordinates within one of the 48 contiguous U.S. states were retained in the corpus. In order to calculate articulation rate from the corpus, it was necessary to remove words in the captions files that immediately preceded or followed longer pauses (see below). After this step, the channels whose aggregated captions files were at least 1,000 words were retained. In total, the corpus comprises 48,945 transcripts from 506 channels, totaling 233,127,501 words with individual word timings (Figure 1). The smallest channel subcorpus is that of the government of Peoria County, Illinois, with 1,031 words. The largest is the channel of Rutherford County, Tennessee, with 8,516,795 words.

State-level aggregation of captions results in subcorpus sizes ranging from 341,050 words (for Montana) to 19,558,326 words (for California). The state-aggregated subcorpora are at least 1 million words in size for 41 of the 48 contiguous U.S. states. A list of the sampled channels, with channel name, channel location, channel id, latitude and longitude coordinates for channel location, number of video transcripts downloaded, total word count of transcripts downloaded, population of channel location, total video duration, mean articulation rate, and standard deviation of articulation rate, is available at https://github.com/stcoats/YouTube_Corpus.

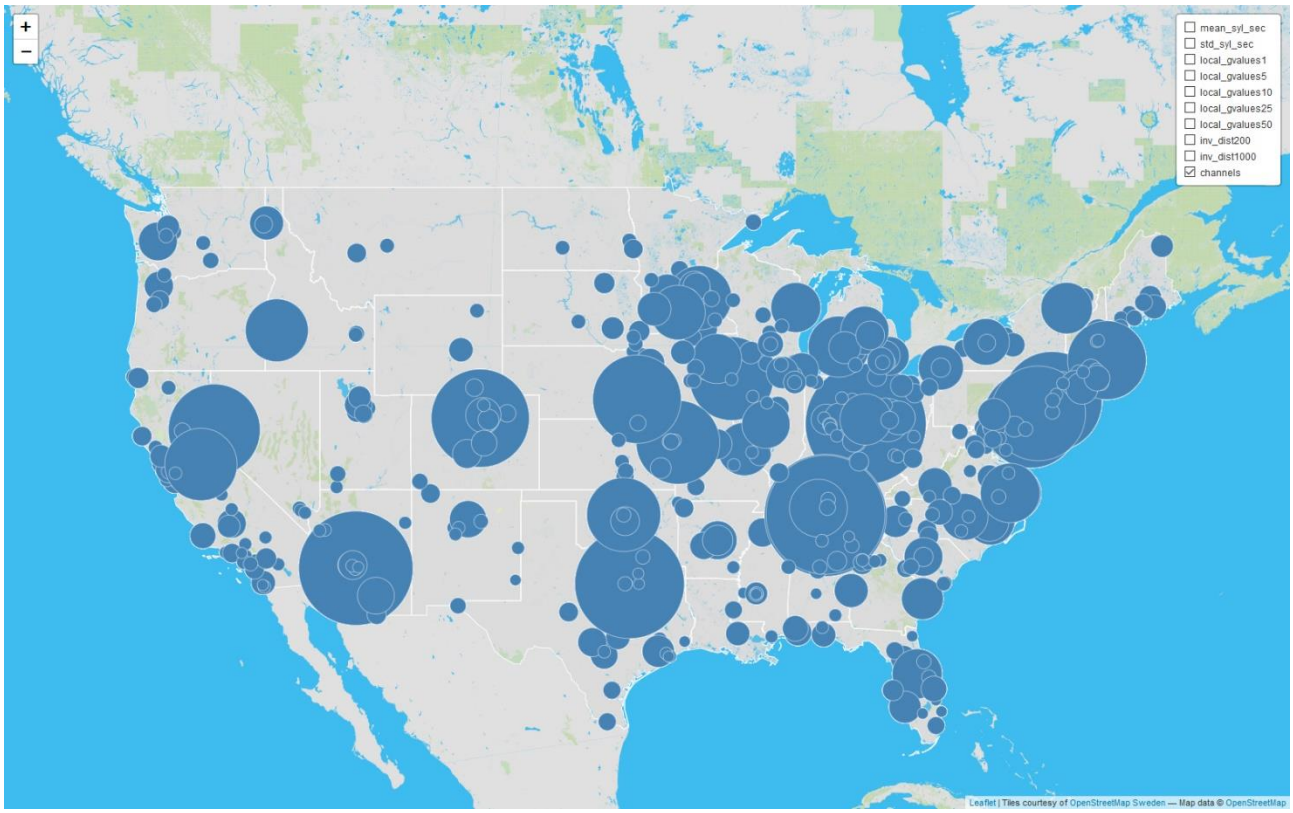

Figure 1: Locations of the YouTube channels sampled in the corpus. Circle size corresponds to number

3.4 Description of videos

Most of the channels in the corpus have videos from different genres. For example, the channel with the second largest number of downloaded captions files in the corpus, that of Murfreesboro, Tennessee (1,153 downloaded files), contains, in addition to videos of meetings of the city council, planning commission, local school board, parks and recreation commission, and zoning appeals board, many short local news reports, typically consisting of multiple segments of video with a voice-over announcer. For the most part, however, the corpus consists of transcripts of public meetings. An example is the video "City School Board Candidate Forum (July 16, 2018)".[3] In this 49-minute video, candidates for a local election to the city school board are introduced. The video begins with a volunteer from a local civic organization speaking directly into a camera which has been set up in the city council meeting room. The volunteer introduces herself and explains that the purpose of the meeting is for "voters to become more familiar with the candidates and their positions". After a cut, the video shows the candidate forum. The moderator for the forum introduces herself to the

local participants and video viewers, then explains the format for the forum: Candidates will make opening statements, then each candidate will be asked short questions by the moderator and will be permitted a short response. The moderator then introduces the six candidates.

The first candidate states that she was born in born in Murfreesboro and has worked in the school system for most of her life. The second candidate introduces himself as a member of various local organizations and thanks the organizers of the forum. The other candidates introduce themselves in a similar manner. The candidates are then asked the question "how would you improve city schools?" The responses include improving school security, pushing for more funding, increasing parent involvement in the schools, or using common texts in classroom teaching. Several other questions are posed on matters such as addressing the needs of non-first-language English speaking children in the schools, school bullying, recruiting teachers, and other topics. At the end of the video, the moderator thanks the candidates, the viewers, the city government, and the people who provided technical support for the video production, and reminds viewers that the school board election will take place on August 2nd. The transcript is 8,256 words long.

3.5 Description of YouTube automatic captions and calculation of articulation rate

YouTube captions are automatically time-aligned to the audio track of the video, with timing tags for each word. Audio segments that are not deciphered by the speech-to-text algorithm leave no trace in the transcript. For overlapping speech segments, transcripts will sometimes contain a correctly-transcribed word, but more often will have no text for the overlapping segments. For non-overlapping speech, word timings correspond to the audio signal for individual words, but not exactly: In order that the captions be legible, utterance-initial words are shown on the screen slightly before they are spoken in the video. Words fill up a caption line, which is bumped up on the screen when a second caption line appears. Figure 2 shows a screenshot with two caption lines from the video "City School Board Candidate Forum (July 16, 2018)" at video time 00:26:23.

Figure 2: Screenshot from the video https://www.youtube.com/watch?v=ON8rdTMh9q8 at video time

At this point in the video, two cues (lines) in the automatically generated speech-to-text transcript are shown on the screen – an excerpt from a question posed by the moderator to one of the school board candidates. Caption cues and timestamps from the captions file for the video are shown in Figure 3. The line 00:26:19.440 --> 00:26:21.620 indicates that during this time span, the text "dedication thank you miss X" was visible as the top captions line.[4] During this interval, the text "what can the City School Board do to" appeared in the bottom captions line, with each word appearing at the time indicated in the immediately preceding tag (the first word in a line appears at the cue start timestamp). The end timestamp for the cue, 00:26:21.620, is the time at which the top line disappeared, and the text "what can the City School Board do to" was "bumped" to the top line. The next timestamp, 00:26:21.630 --> 00:26:23.680, indicates the span of time during which the following cue ("help recruit and retain good teachers") appeared in the same fashion. At the end timestamp for this entire cue, 00:26:23.680, the cue is "bumped up": The line "what can the City School Board do to" disappears, "help recruit and retain good teachers" moves to the top line, and a new line appears at the bottom.

```
00:26:19.440 --> 00:26:21.620
dedication thank you miss X
what<00:26:20.010>  can<00:26:20.220>  the<00:26:20.340>  City<00:26:20.550>
School<00:26:20.850> Board<00:26:20.880> do<00:26:21.270> to

00:26:21.630 --> 00:26:23.680
what can the City School Board do to
help<00:26:21.780> recruit<00:26:22.170> and<00:26:22.200> retain<00:26:22.560>
good<00:26:22.980> teachers
```

Figure 3: Excerpt from .vtt file for https://www.youtube.com/watch?v=ON8rdTMh9q8, (additional class and color tags have been removed for illustrative purposes).

Because words are arranged sequentially in a captions cue and timings within a cue do not overlap, if speech is continuous, word durations can be calculated by subtracting a word's start time from that of the following word. The simplest method to derive a rate value for a video would be to divide the number of syllables in the orthographic transcript of the captions file by the sum of the word timing tag durations. For example, the total duration of the two cues shown in Figures 2 and 3 is 4.24 seconds. The number of syllables in the text (18) divided by the sum of word timings in the excerpt gives an articulation rate of 4.25 σ/sec. for this very short excerpt. While the passage excerpted is within an utterance of fluent continuous speech by one speaker that contains no pauses, using this method to calculate articulation rate from caption cues becomes more complicated when there are intra- or inter-utterance pauses.

Some channels include videos where a council session or other meeting is preceded or followed by a long test screen showing the text "city council session will begin soon" or similar, in some cases because the videos are streamed live and the council session is not yet ready to begin. Other videos have long interruptions between utterances by different speakers or within utterances by a single speaker, for various reasons, such as a speaker needing time to approach a microphone, the chairperson of a meeting searching for a document or adjusting a setting on the computer used to track the meeting's agenda, the council waiting until members have submitted their votes on their computers, or other reasons.

An example occurs in the video for the meeting of the Bellevue City Council, Nebraska, on March 12, 2018, at 00:17:30.00 (https://youtu.be/cK3CXpoH0qg?t=1050, Figures 4 and 5): The city mayor, commenting on a vote that has just been conducted on the council members' computers, says "it's a 3:3 vote", followed by a pause of 16.340 seconds, before continuing "because it's a personnel issue, we're trying to do the right thing". The words "it's a 3:3 vote" remain on the screen during the pause. Calculating the articulation rate by dividing the total duration of the timestamps in a single cue (the cue "it's a 3:3 vote because it's a personnel") by the number of words and syllables would result in the misleading value of 0.610 σ/sec.

Figure 4: Screenshot from the video https://www.youtube.com/watch?v=cK3CXpoH0qg&feature=youtu be&t=1050 at video time 00:17:30.00.

```
00:17:30.820 --> 00:17:50.680
it's<00:17:31.820>  a<00:17:31.940>  3:3<00:17:32.660>  vote<00:17:49.000>
because<00:17:50.000> it's<00:17:50.150> a<00:17:50.240> personnel
```

Figure 5: Excerpt from .vtt file for https://www.youtube.com/watch?v=cK3CXpoH0qg.

In addition, words spoken after pauses (whether between speakers or within the turn of a single speaker) will be shown for exactly one second on the screen, always appearing slightly before the audio, in

order to enhance the legibility of the captions. For these reasons, articulation rates for the videos were calculated after filtering out word tokens with long durations (utterance-initial words and words spoken immediately before or following longer pauses; boldface in Table 2). This intra-utterance continuous articulation rate is defined to be the articulation rate, in syllables per second, for all word tokens in a captions file whose sequential duration is less than 1 second.

Table 2: Calculation of articulation rate from individual word timings. Words with a duration of 1 second or longer are in bold.

word	start time	end time	duration
it's	00:17:30.820	00:17:31.820	**00:00:01.000**
a	00:17:31.820	00:17:31.940	00:00:00.120
3:3	00:17:31.940	00:17:32.660	00:00:00.720
vote	00:17:32.660	00:17:49.000	**00:00:16.340**
because	00:17:49.000	00:17:50.000	**00:00:01.000**
it's	00:17:50.000	00:17:50.150	00:00:00.150
a	00:17:50.150	00:17:50.240	00:00:00.090
personnel	00:17:50.240	00:17:50.680	00:00:00.440

Applying this procedure to the text from Table 2 results in an articulation rate of 5.26 σ/sec.–a value comparable to those reported in recent studies of articulation rate in acoustic phonetics.

Because individual speakers are not tagged with metadata in captions files downloaded from YouTube, the articulation rates in this study are calculated for entire videos/captions files, not for individual speakers. Thus, for a typical local government meeting, the rate is based on all the utterances of all speakers in the video (typically between 5 and 10 persons). In other captions files in the corpus, such as those extracted from interviews, news reports, or vlog-style videos, the calculated value may represent the rate of just a few or a single speaker. Because the method used to calculate articulation rate is based on within-utterance timings and omits longer pauses, whether between utterances of different speakers or within the utterances of a single speaker, differences in articulation rate that may arise due to differences in communicative configuration parameters for the different types of videos in the corpus are minimized. Syllabification of the

text was undertaken using a script in *R* prepared by Tyler Kendall (Kendall, 2013) after converting numerals in the text to their word forms.

This method for the calculation of articulation rate omits longer pauses but could also potentially omit words with articulation durations of longer than one second. While such words are certainly possible in spoken English, previous studies suggest that most words have a much shorter duration. For example, Baker and Bradlow reported word durations ranging from .142 to .740 s, with a mean value of .362 s, in an experiment in which American English speakers read different texts (2009), and Yuan et al. reported mean word durations of .2 to .45 s in shorter utterances in a corpus of American English telephone conversations (2006). In addition, because this study focuses on comparison of articulation rates at different locations and the method used for the calculation of articulation rate is the same for all locations, unless the distribution of word durations is substantially different for different channel locations, it is unlikely that filtering out words with longer calculated durations will bias the results. The similarity of communicative contexts and the relatively large number of speakers for most channel subcorpora also help to ensure that the word duration distributions at channel level are comparable.

3.6 Validation of method

In order to test whether the automatic calculation of the intra-utterance continuous articulation rate produces values that are comparable to those derived from other methods, articulation rates were calculated for the first 10 minutes of 20 randomly selected transcripts from the corpus with the method described above and with a semi-manual method using the corresponding audio tracks, which were scraped from YouTube using youtube-dl and converted to .wav format using FFmpeg (Hsuan, Amine & M., 2018; FFmpeg Developers, 2019). Articulation rates were then calculated directly from the audio using the *speechrate* script (De Jong & Wempe 2009) in Praat (Boersma & Weenink, 2010). The script detects syllable nuclei from an audio signal on the basis of sound intensity in decibels.

Figure 6 shows a Praat screenshot of the signal from 9.25 seconds of audio from the beginning of the meeting of the Bellevue City Council, Nebraska, on March 12, 2018. The annotation tiers 1 and 2 are the output of the *speechrate* script, using a silence threshold of -20 decibels and a minimum pause duration of 100ms. In tier 1, syllable nuclei are indicated with individually-numbered boundary markers, and in tier 2, sounding segments are distinguished from pauses. Tier 3 shows the automatically-generated transcript, with word timings, for the excerpt.[5] The *speechrate* script correctly counts 29 syllables and calculates an articulation rate based on the duration of the sounding segments of 4.66 σ/sec. In the speech-to-text transcript, there are four words whose duration is one second or longer ("good", "again", "welcome", and "2018"). Using the method described above, the articulation rate for the entire 9.25-second segment is calculated as 5.14 σ/sec.

There are two reasons why the articulation rate calculated from the speech-to-text transcript is slightly higher: First, while the method effectively excludes pauses from the articulation rate calculation, it also excludes most utterance-initial and utterance-final words, which tend to be articulated somewhat slower than the utterance-internal words due to phrase boundary constraints (Oller, 1973; Byrd & Saltzman, 1998). In addition, the transcript generated by the speech-to-text service consists of standard orthographic word forms, some of which are normally subject to elision or other phonological reduction processes in spoken language. In the audio signal shown in Figure 6, for example, "evening" is realized not with three syllables, but as [iv.nɪŋ]. The orthographic transcript thus can contain more syllables than are articulated in the audio file, and hence the articulation rates determined by the procedure used in this study are slightly higher than articulation rates as calculated semi-manually. Nevertheless, because the same procedure has been used to analyze all of the transcripts in the corpus, it is unlikely that the slightly higher articulation rates will introduce bias in terms of the regional or the urban-rural analyses.

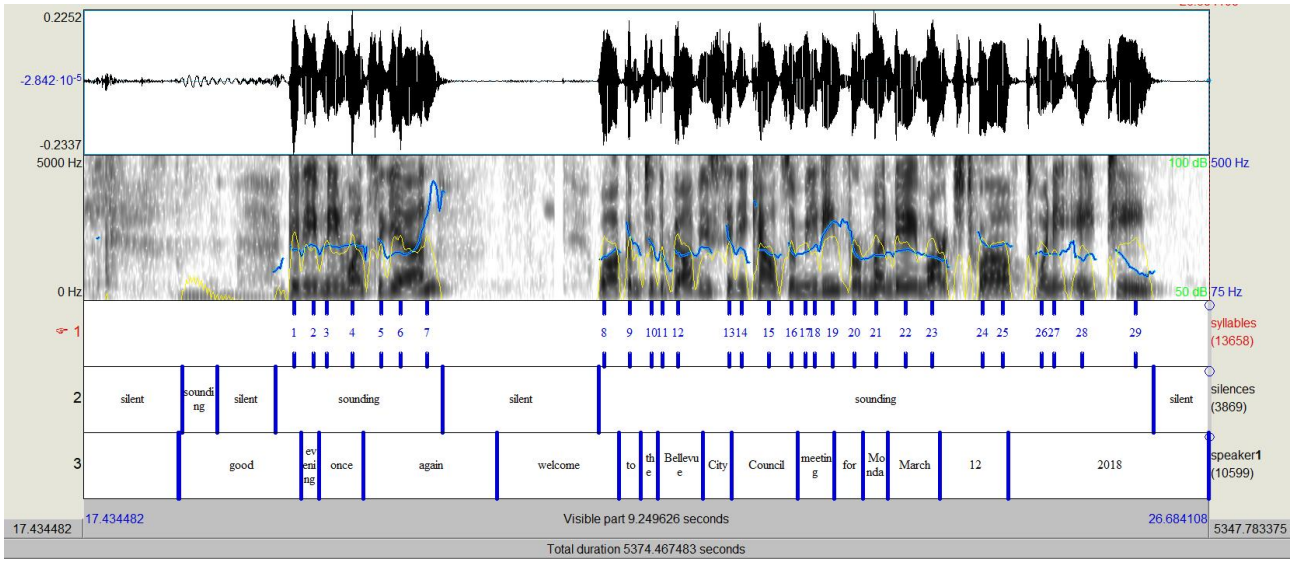

Figure 6: Screenshot of Praat excerpt for Bellevue City Council, Nebraska, 12 March 2018.

Figure 7 shows the best-fit regression line for the articulation rate values calculated from the automatic speech-to-text transcripts versus the values as semi-manually calculated from the audio signal using *speechrate*. As can be seen in the figure, while the intra-utterance continuous articulation rates are slightly higher, a strong correlation is found between the values calculated according to the two methods.

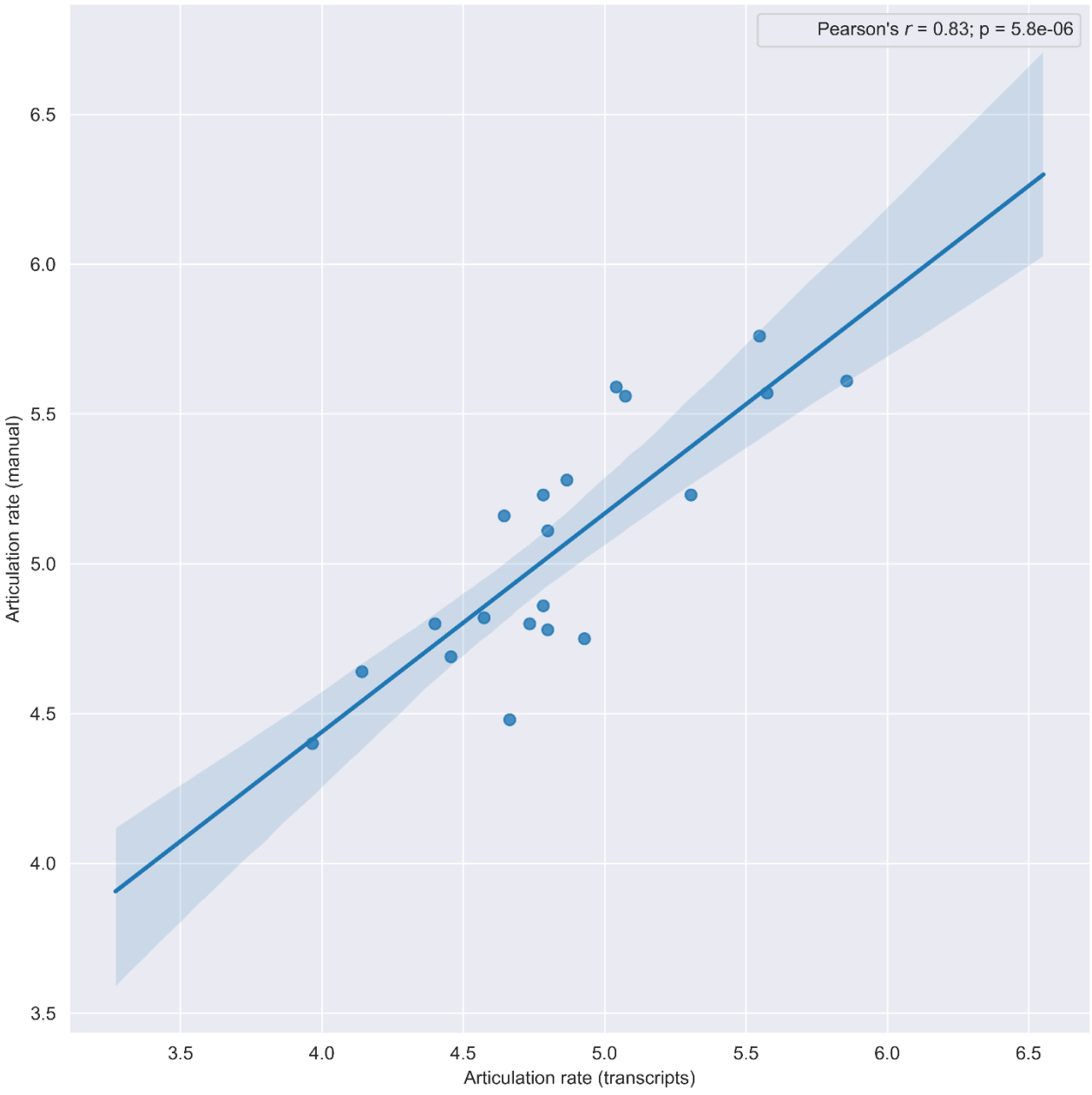

Figure 7: Articulation rate (speech-to-text transcripts) versus articulation rate (semi-manually calculated), 20 videos.

For tracks with poor audio quality, high levels of background noise, long passages of incoherent or overlapping speech, loud background music, or other fidelity issues, the automatic speech-to-text transcripts are often incorrect, resulting in inaccurate intra-utterance continuous articulation rate values. Filtering out transcripts with fewer than 20 words of text during corpus creation eliminated some of the videos with low quality audio. In general, the automatic calculation of articulation rate from transcript word timings is

accurate for recordings with high audio fidelity, whose audio tracks have been correctly transcribed. Future work with this corpus, and with other materials using this or similar methods, will need to focus on the identification of those transcripts with the highest levels of accuracy.

3.7 Spatial statistics

In recent years, alongside the development of quantitative and statistical techniques in corpus linguistics in general, several open-source libraries in programming languages such as *R* or *Python* have been developed that simplify the calculation of various spatial autocorrelation statistics (Bivand, Pebesma & Gomez-Rubio, 2013, Rey et al. 2015). Autocorrelation statistics can be used to assess the degree to which all points associated with a variable exhibit a spatial pattern (a global statistic), or can assess the degree to which individual points exhibit high or low values, compared to neighboring points (a local statistic). In this study, the global spatial statistic Moran's I and the local statistic Getis-Ord G_i^* were calculated, with the latter used to assess the regional patterning of the calculated articulation rate values.

Moran's global I (Moran 1950), a commonly used measure of global autocorrelation, quantifies the extent to which the values associated with a set of spatial points are similar to the values of neighboring points. The statistic, essentially Pearson's product-moment correlation between the values of a variable and their "spatial lag" as defined by a weighting function, is calculated by

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

where n is the total number of points in the data, $\bar{x}$ is the mean of the variable and i and j index locations. A spatial weights matrix $\boldsymbol{W}$ summarizes the connections between the points in the data set with a weighting

value w_{ij} for all location pairs: the value can be binary (i.e. 1 for points considered to be neighbors and 0 for non-neighboring points) or continuous (e.g. based on the inverse distance between points or a function thereof). Neighbors can be determined by polygon or choropleth continuity (areas with shared borders or vertices on a map), based on a distance threshold (all points within a certain distance of one another are considered neighbors), or defined as a set of k-nearest neighbors (e.g. the 5 nearest points to any given point are its neighbors). Finally, the decision has to be made whether the matrix values are to be normalized, and if so, what form the standardization should take (e.g. according to row sums, global sums, or some other value). Moran's I is a global value for an entire data set ranging between a theoretical -1 (for data that is perfectly dispersed spatially) to 1 (for data that is maximally clustered spatially). A value of 0 indicates random spatial dispersion of values.

To analyze local differences, the Getis-Ord G_i^* statistic can be utilized (Getis & Ord 1992). For an area divided into n regions indexed by $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, n$, the value of the statistic for region i is calculated by

$$G_i^*(d) = \frac{\sum_j w_{ij}(d) x_j}{\sum_j x_j} \tag{2}$$

where $w_{ij}(d)$ is the value drawn from the spatial weights matrix $\boldsymbol{W}$ for all points within distance d of the centroid for i. As is the case with Moran's I, the spatial weights matrix can be binary or continuous and based on choropleth contiguity, a cutoff distance, or a distance function.[6] Ord and Getis (1995) transformed equation (2) into a standard variate by subtracting expected values and dividing by the square root of the variance:

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j - \bar{X}\sum_{j=1}^{n} w_{ij}}{\sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{n-1}}} \quad (3)$$

Positive values of G_i^* at a given point mean that point is located in a cluster of high values for a variable, whereas negative values mean the point is in a cluster of low values. G_i^* is a standard variate; values of ± 1.645 are significant at p = 0.05.

4. Results

The calculated mean articulation rate for the 48,945 videos in the corpus is 5.12 σ/sec., with a standard deviation of 0.37 and a range of 4.39 to 5.88 (Figure 8). The calculated mean articulation rate is comparable to rates reported for American English in previous literature such as Kendall (2013) or Jacewicz et al. (2011), or for British English as reported by Goldman-Eisler (1961), as well as for articulation rates reported for Dutch or German (Verhoeven et al. 2004; Jessen 2007).[7] When the videos are aggregated by channel, the highest rate was found for the channel "Ownby VA/US Government" from Virginia (6.59 σ/sec.); the channel consists of four videos with vlog-style commentary by a single speaker. Because the articulation rate value for this channel was much higher than for the other channels in the data, and because the value represents a single speaker, the data point was discarded for the calculation of spatial autocorrelation statistics and mapping presented in the next sections. The channel with the lowest articulation value was found to be "Mississippi Department of Human Services" (4.39 σ/sec.).

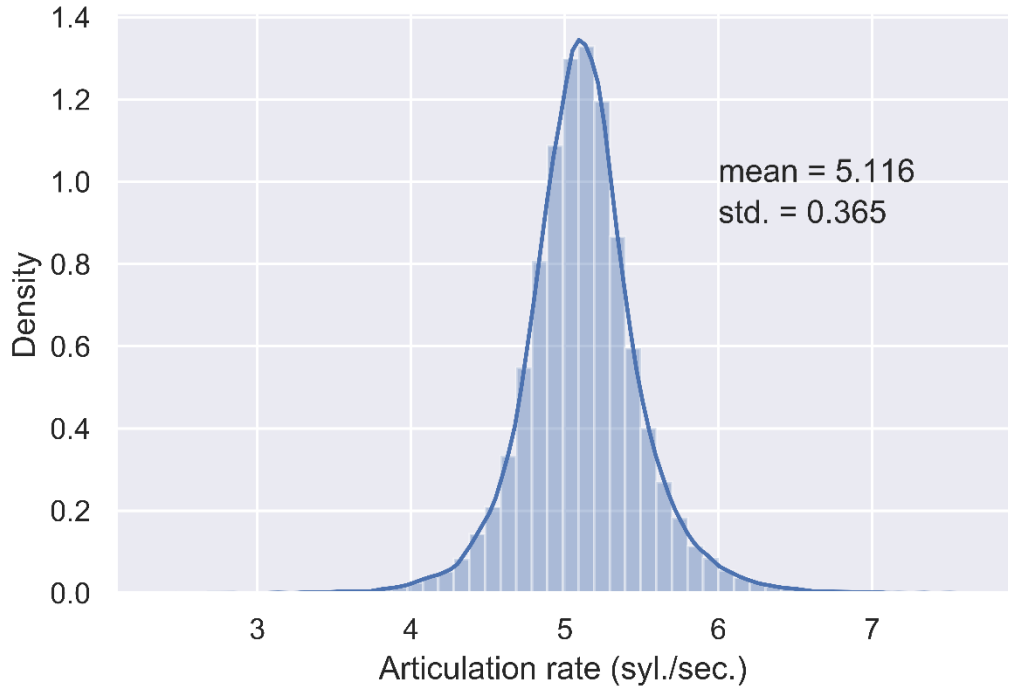

Figure 8: Density plot of articulation rates for the 48,945 videos.

4.1 Regional distribution of articulation rate

To assess the degree of spatial clustering of articulation rate values, following recent approaches in dialectology (Grieve et al., 2011; Grieve, 2013, 2016), measures of global and local spatial autocorrelation were calculated with the R package *spdep* (Bivand et al., 2013). Map visualization was achieved by creating a Voronoi tessellation (Voronoi, 1907) with the *deldir* package (Turner, 2019) for the 506 channels located within the contiguous 48 US states. Interactive maps were created using *Leaflet* in *R* (Agafonkin et al., 2018, Cheng, Karambelkar, & Xie, 2018). A Voronoi tessellation was chosen for visualizations due to the unbiased manner in which the tessellation partitions space into polygons: All points in a polygon are closer to that polygon's centroid than to any other centroid. The spatial contiguity of the tessellation makes color-shaded heat maps easily interpretable. It should be noted, however, that polygon size has no relation to the size of the sample (in words) associated with that polygon's centroid: Polygons in regions with many channels are quite small, and those in regions with few channels large. The size of the polygons depends only on the locations of the nearest sampling points, not on the size of the subcorpus for that channel.

Figure 9 shows the raw values for mean articulation rate for the channels in the corpus: The channel locations are the polygon centroids and mean articulation rate corresponds to color intensity. Values range from 4.39–5.88 σ/sec., a range significantly larger than the approximately 5% threshold for hearers to perceive tempo differences (Quené, 2007). There is no immediately apparent geographical pattern in the raw articulation rate values, although some regional trends are discernible. It should be noted, however, that for areas with higher densities of sampled YouTube channels, such as the Eastern Seaboard or the metropolitan areas of larger cities, the some of the Voronoi cells are quite small, making it more difficult to perceive areas with more uniform values.

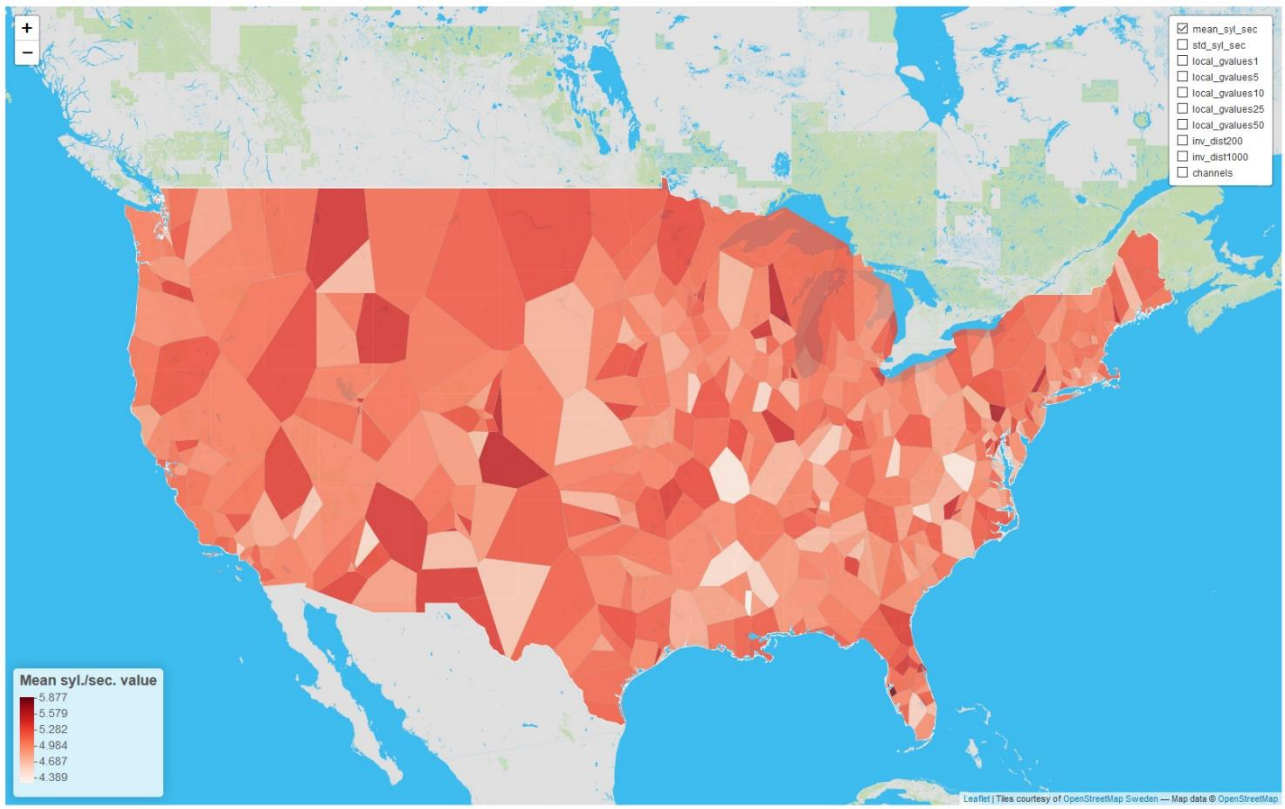

Figure 9: Mean articulation rate (σ/sec).

To better understand the data, Moran's I and the Getis-Ord G_i^* statistic were calculated using several different spatial weights matrices based on different combinations of parameters: first, with a binary contiguous spatial weights matrix (a "queen" continuity matrix, i.e. one in which region i's neighbors are all polygons with a shared edge or vertex in the Voronoi tessellation), second, with binary weights matrices with

various distance cutoff thresholds (50, 100, 200, 500, 1000, 2000, and 5000 km), third, with binary weights matrices based on different numbers of nearest neighbors (5, 10, 25, and 50), and finally with a continuous spatial weights matrix for all points in the data based on inverse distance, with the same seven distance thresholds as above. All matrices were row normalized. As noted by Grieve (2016, p. 115), the choice of the most appropriate spatial weights matrix depends on the nature of the data being analyzed and the hypothesis being explored. Exploratory analysis showed that for this data, the calculated Moran's I values varied between -0.03 (for a binary spatial weights matrix based on a threshold distance of 200km) and 0.16 (for a binary spatial weights matrix based on 5 nearest neighbors), suggesting that from a global perspective, articulation rate measures are not strongly autocorrelated in the data, but rather dispersed somewhat randomly. The lack of global autocorrelation, however, does not imply that regional patterns are absent in the data – on the contrary, as noted by Ord and Getis, "when global autocorrelation exists, local pockets are harder to detect. Conversely, when no global pattern exists, G_i^* helps to monitor local behavior" (1995, p. 299).

The maps in Figures 10–16 show the spatial distribution of G_i^* scores as calculated from seven different spatial weights matrices: the polygon continuity matrix, k nearest-neighbor matrices with 5, 10, 25, and 50 nearest neighbors, and the continuous inverse distances matrices with 200 km and 1000 km cutoffs.[8] Figure 10 shows G_i^* values calculated from a binary spatial weights matrix based on polygon continuity (i.e., each polygon's neighbors are those that share an edge or a vertex). Although the values are somewhat randomly distributed, a cluster of low scores is evident in the states of Mississippi, Arkansas, and Louisiana. Figure 11 shows the calculation based on 5 nearest neighbors: Here again, the pattern is mostly random, but a pocket of low values is discernible in the South. Increasing the number of nearest neighbors in the spatial weighting function to 10, 25, and 50 (Figures 12, 13, and 14) reveals a clear regional pattern for articulation rates within the US: The Western South (Louisiana, Mississippi, Arkansas, Tennessee, and Alabama) shows the lowest articulation rates, and the Upper Midwest (Minnesota, Wisconsin, and the Upper Peninsula of Michigan) the highest rates. The Mountain West and the Pacific Northwest show higher values, as do, to a lesser degree, Florida and the Mid-Atlantic. Figure 15 shows the calculated G_i^* values for a continuous spatial

weights matrix with a cutoff of 200km. The pocket of low values in the Western South is perceptible, but much of the territory of the US shows uniform values, because the centroids of the polygons are not within 200km of any neighbors. When the cutoff distance is increased to 1000km, so that each point has a relatively large number of neighbors (Figure 16), the regional patterning seen in Figure 14 is again evident: Lower values are concentrated in the South, and higher values in the Upper Midwest.

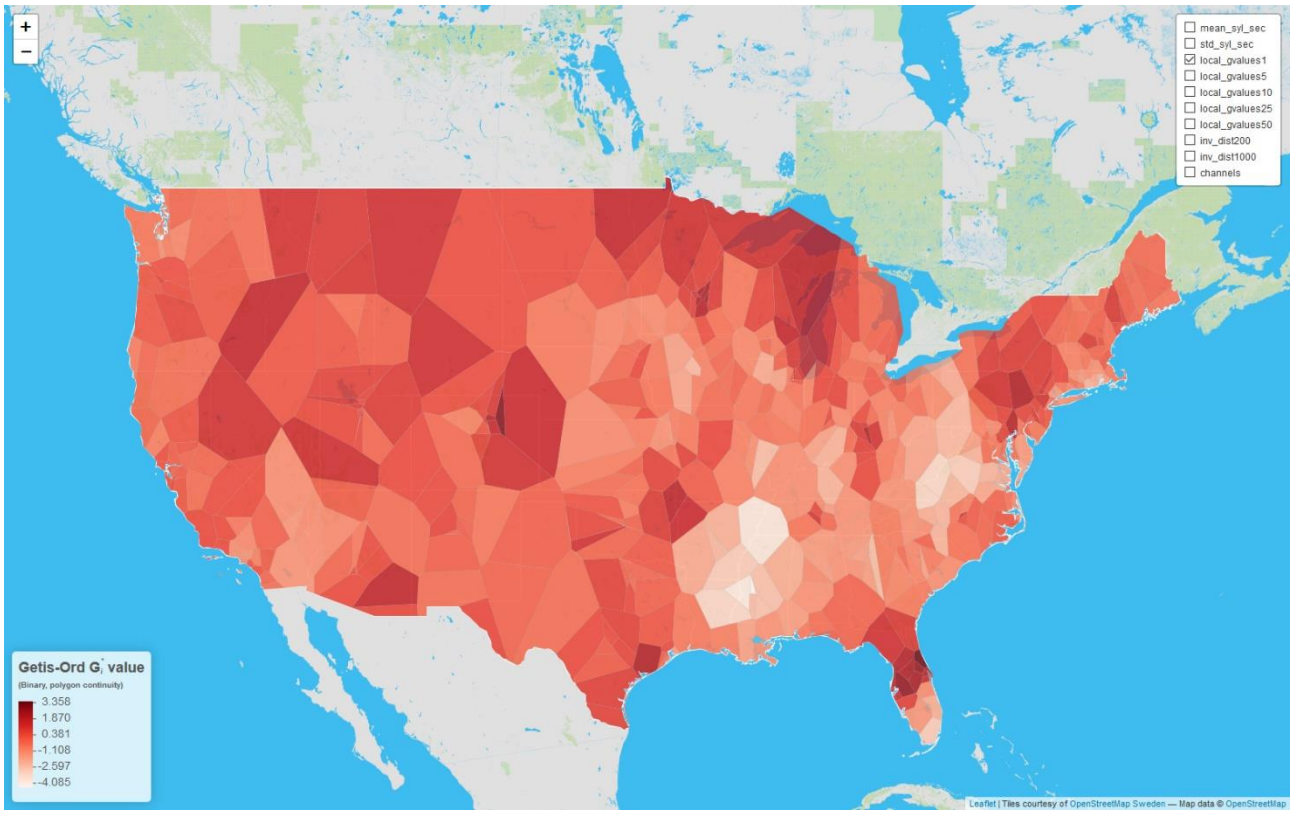

Figure 10: Getis-Ord Gi values for a binary spatial weights matrix based on polygon contiguity.*

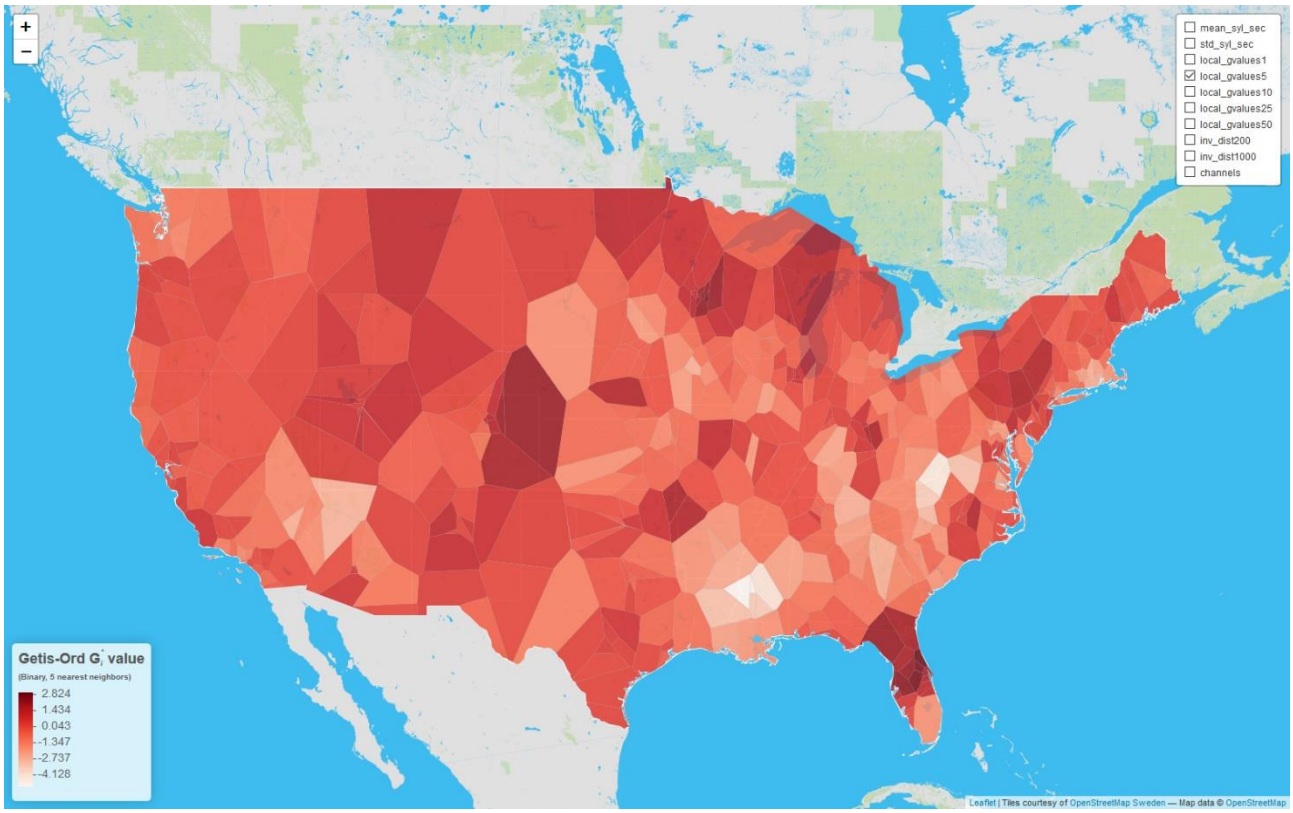

Figure 11: Getis-Ord Gi values for a binary spatial weights matrix based on five nearest neighbors.*

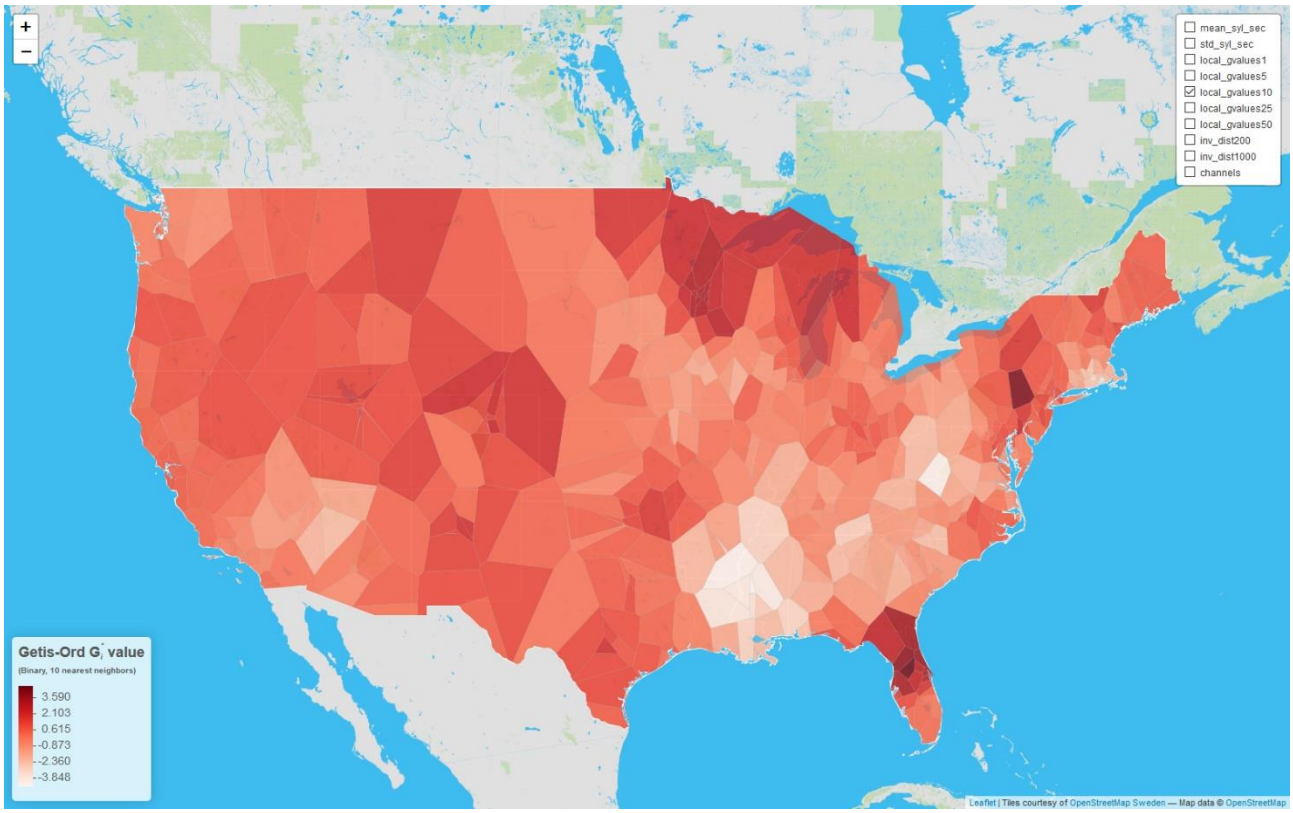

Figure 12: Getis-Ord Gi values for a binary spatial weights matrix based on ten nearest neighbors.*

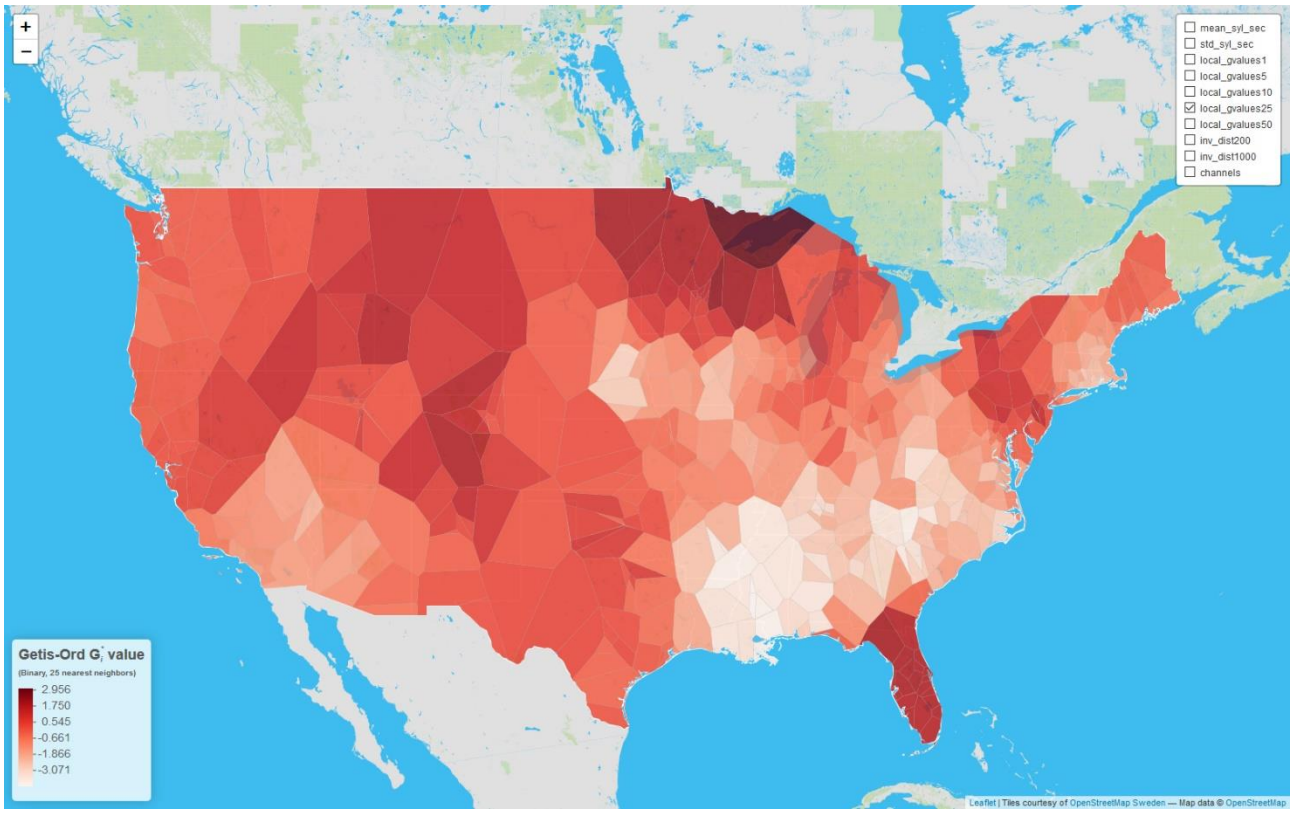

Figure 13: Getis-Ord Gi values for a binary spatial weights matrix based on 25 nearest neighbors.*

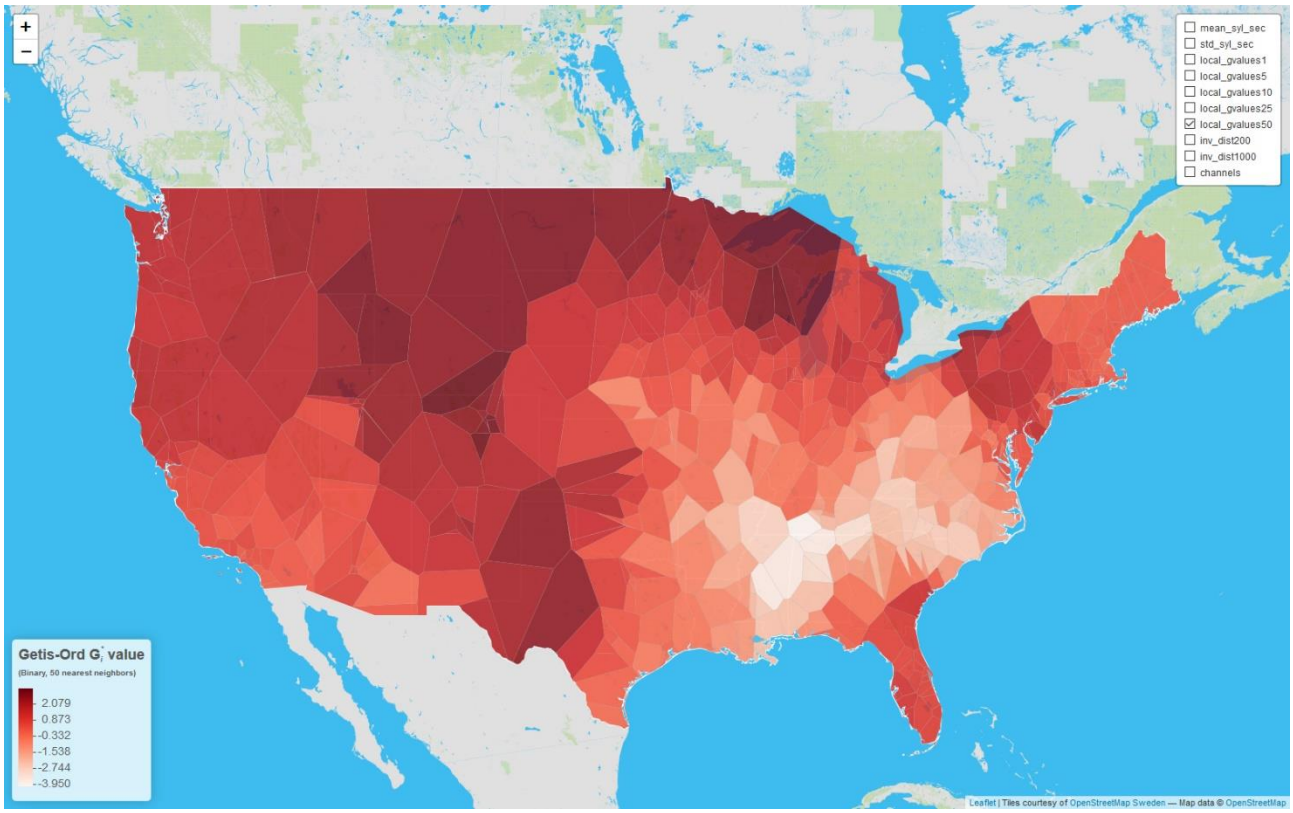

Figure 14: Getis-Ord Gi values for a binary spatial weights matrix based on 50 nearest neighbors.*

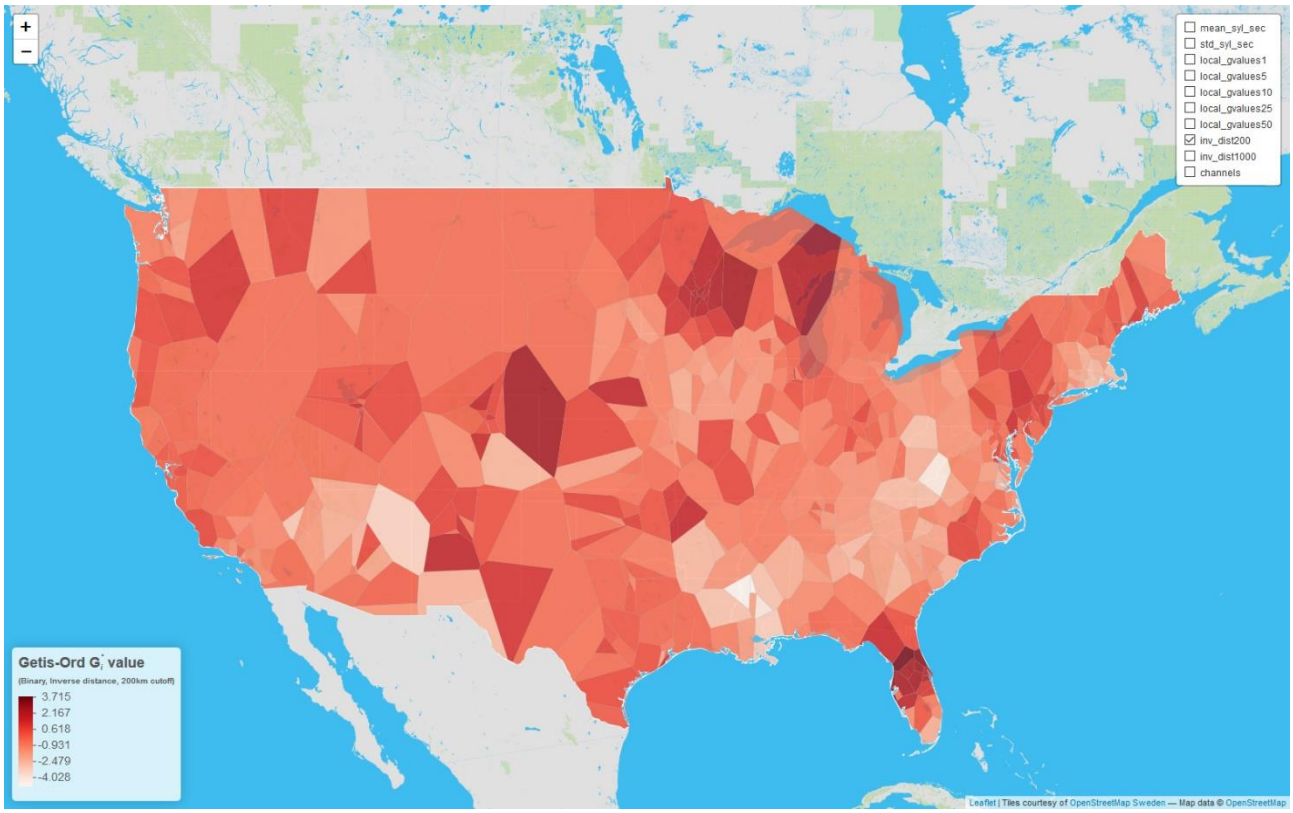

Figure 15: Getis-Ord Gi values for a continuous inverse distance spatial weights matrix with a 200 km cutoff.*

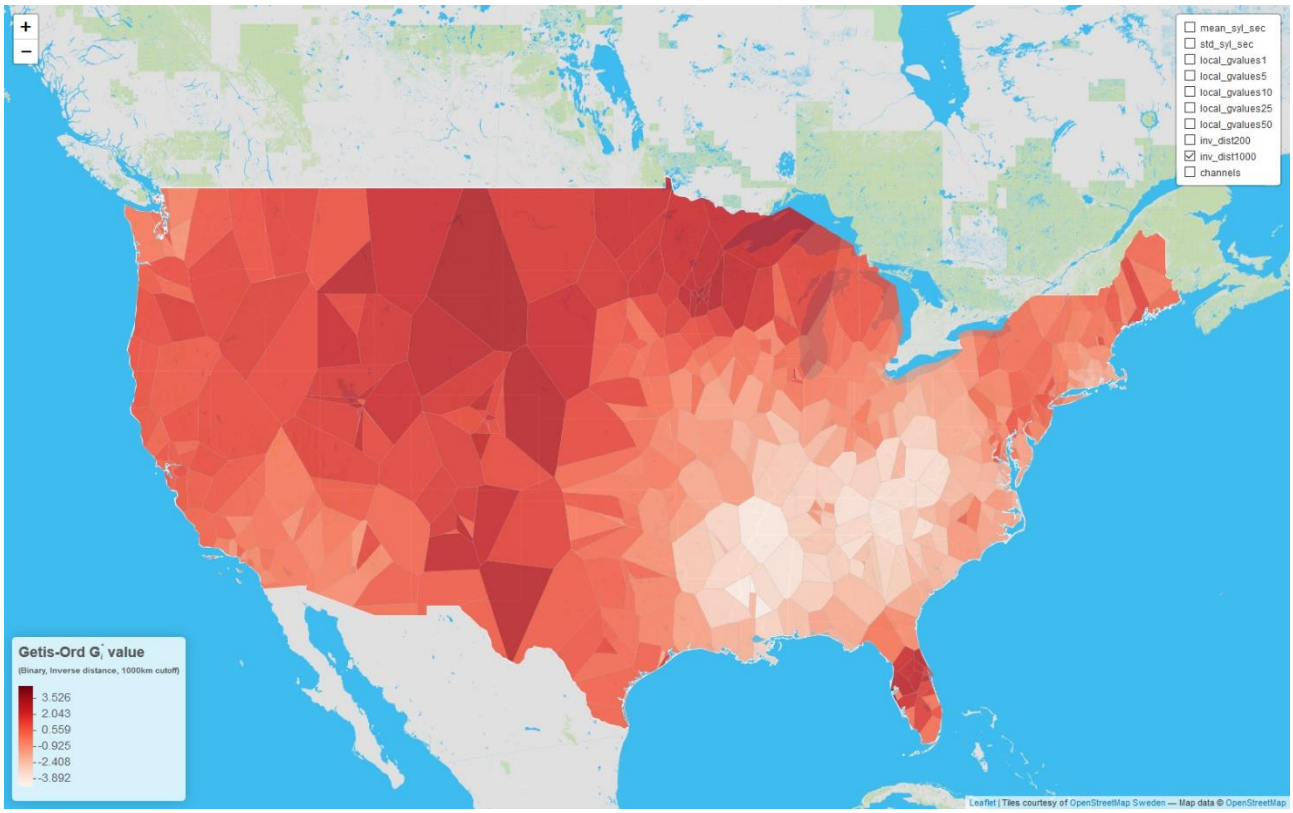

Figure 16: Getis-Ord Gi values for a continuous inverse distance spatial weights matrix with a 1000 km cutoff.*

In terms of statistical significance, the proportion of channels with a G_i^* value significant at p = .05 ranges from .119 (for the 5 nearest-neighbor spatial weights matrix) to 0.234 (for the continuous inverse distance spatial weights matrix with a 1000 km cutoff), with the significant clusters indicated on the maps in Figures 10–16 by the darkest and palest shades. Although changing the parameters for the calculation of the spatial weights matrix slightly changes the pattern of significant clusters in the resulting maps, in all of the maps a significant cluster of lower articulation rates is centered in Mississippi and a significant cluster of higher rates in the Minneapolis-St. Paul area of Minnesota. High articulation rate clusters in central Florida and in the Mountain West are evident for some of the spatial weights matrices. In addition, in most of the maps a lower articulation rate cluster is evident in New England, centered on the towns of Groton, Connecticut and Coventry, Rhode Island.

Overall, the spatial weights matrix parameter dependency of the G_i^* values, as well as the low values calculated for the global Moran's I statistic, suggest that variation in the articulation rate in the U.S. does not

show a clear monotonic association with location. Two regions show a consistent pattern of higher and lower articulation rates when changing the spatial weighting model: the Upper Midwest and the South. For the most part, these articulation rate findings are in accord with regional patterns reported in previous studies based on smaller datasets with coarser geographic granularity, although there are some differences. Possible interpretations of the regional patterning of articulation rates are provided in Section 5.

4.2 Articulation rate and population size

To assess the hypothesis that residents of towns or rural areas with relatively smaller populations exhibit a slower articulation rate than do residents of larger cities or urban areas, population data from the US Census Bureau was correlated with articulation rate. To retrieve population statistics, a script was written to extract the place names associated with the exact latitude-longitude coordinates determined for each channel (described in Section 3.3, above) and return the population estimate for the year 2017 in data from the U.S. Census Bureau (U.S. Census Bureau, 2017). Populations were normalized by taking their natural logarithm, after which a linear regression with articulation rate was conducted. Figure 17 shows the relationship between mean articulation rate and the natural logarithm of the population size. As can be seen, a weak but significant positive correlation exists between log population and articulation rate, suggesting that at least for this sample of spoken data, city dwellers may indeed speak slightly more quickly than persons who reside in smaller communities or in rural areas.

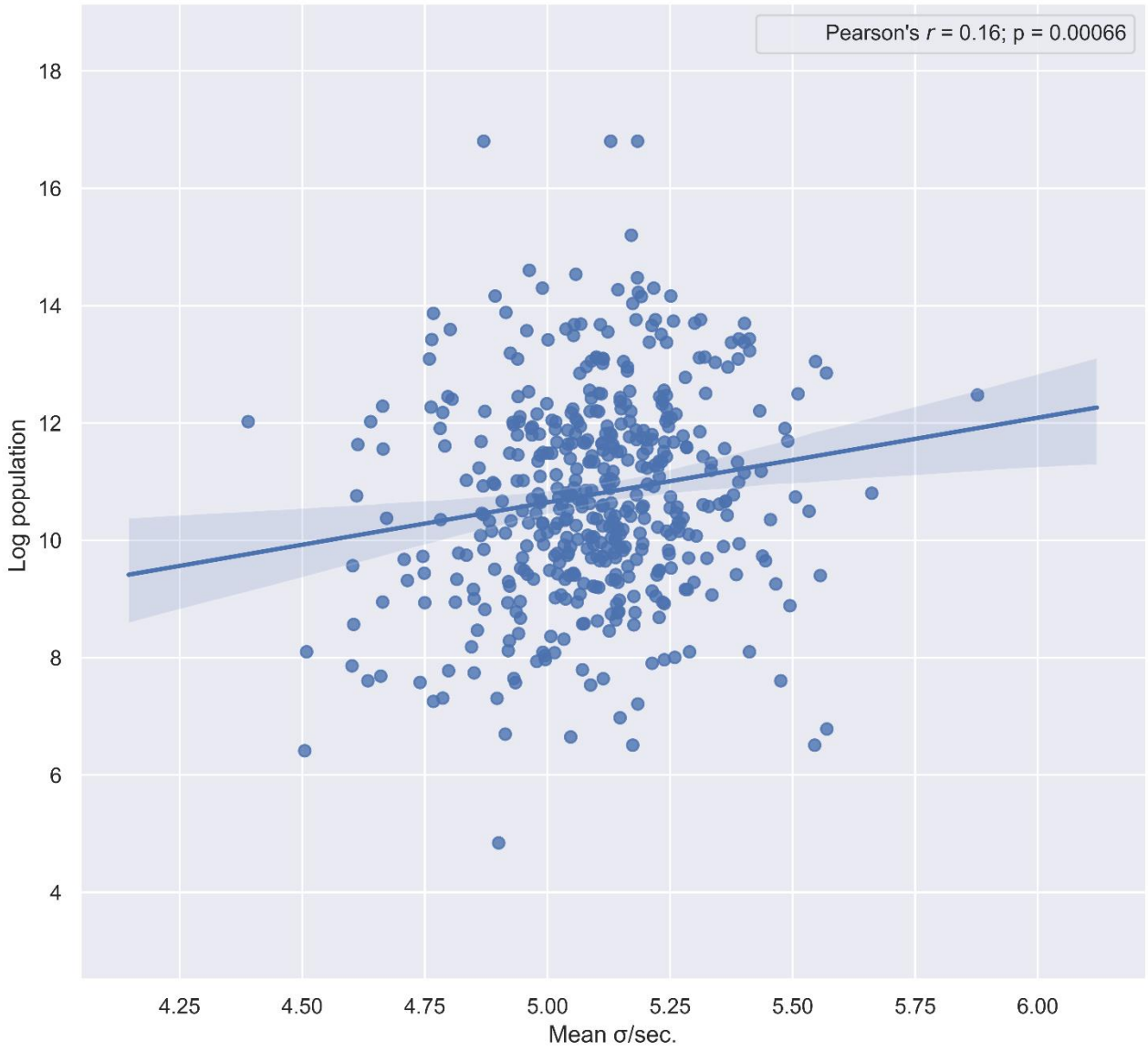

Figure 17: Mean articulation rate (σ/sec) versus log population.

5. Discussion

The findings of this study offer partial confirmation of earlier research based on smaller samples of mostly reading passage data in which Southerners were found to have lower speaking or articulation rates than Americans from other parts of the country (Byrd, 1992, 1994; Jacewicz et al., 2009, Jacewicz et al., 2010; Clopper & Smiljanic, 2015). This confirmatory finding, as a proof-of-concept, suggests that the method

devised to utilize speech timing information from automatic speech-to-text transcripts can accurately capture variation in the temporal organization of speech.

Although no obvious regional patterns were evident when viewing the distribution of raw articulation rate values on a map, the use of the Getis-Ord G_i^* statistic showed a consistent regional pattern, especially when using a spatial weights matrix with a bandwidth that encompasses a relatively large area. The local spatial autocorrelation shows that speakers in the American South, especially in Louisiana, Mississippi, Arkansas, Alabama, and Tennessee, articulate slower than speakers elsewhere in the contiguous 48 states. Speakers in the Upper Midwest, especially in Minnesota and Wisconsin, articulate faster, as do speakers in the Mountain West from Idaho to West Texas, and speakers in central Florida.

The regional pattern of articulation rate differences is apparent when Getis-Ord G_i^* values are mapped using a contiguous binary spatial weights matrix (Figure 10), but becomes more pronounced when the spatial weighting function includes increasing numbers of nearest neighbors or when a continuous spatial weights matrix based on inverse distances with an increasing cutoff threshold is used (Figures 11–16). These facts suggest that while articulation rates can be highly variable within smaller geographical regions (as found also by Kendall, 2013, for North Carolina), differences are evident when the focus is shifted to larger numbers of speakers and broader geographical expanses of the North American continent.

Getis-Ord G_i^* values calculated on the basis of a binary spatial weights matrix taking into account the 50 nearest neighbors (Figure 14) show that areas of intermediate articulation rates include a swath of territory from Nebraska to Western Pennsylvania, California and the West Coast, and Arizona. In line with popular conceptions and some previous dialectological research identifying the American Midwest as the home of the least marked and most intelligible American English variety in terms of phonological features (Clopper & Bradlow, 2008), the two channels with the raw articulation rate values closest to the mean value for the entire data, 5.12 σ/sec., are in this region: the channel for the city of Wayne, Michigan, and the channel of New Carlisle, Ohio.

Some differences are found in the regional pattern of articulation rates compared to findings reported in previous studies. For example, Clopper and Smiljanic (2015) reported higher articulation rates for New England speakers than for speakers from the Upper Midwest/Inland North. In their sample, however, the Inland North was represented by eight speakers from Indiana and Illinois, one from Wisconsin, and one from New York State, with none from Minnesota, North Dakota, South Dakota, Iowa, or Michigan (Clopper & Pisoni 2006, p. 640). The higher articulation rated found for Upper Midwestern speakers in this study are in accord with results reported by Byrd (1992) and are indirectly supported by the findings of Jacewicz et al. (2009) and Jacewicz et al. (2010). The findings for Florida, where rates are higher than in the rest of the American South, are in accord with work in perceptual dialectology in which Floridians generally do not characterize the spoken English of their state to be "Southern" (Garzon, 2017, p. 29), and also in line with previous dialectological studies in which the phonology of Floridians is found to be different from that of residents of other Southern states (Labov et al., 2006), possibly due in part to the sustained high levels of intra-American immigration to Florida since the 1950s, especially from Northern states.

While this study tentatively confirms some previous findings for articulation rate in American English, namely that Southerners articulate slightly slower, the reasons for the geographical pattern suggested by the Getis-Ord G_i^* values are not clear, and different interpretations are conceivable. The possibility that prosodic features such as articulation rate bear indexical meaning and, together with the temporal organization of pauses in speech, can be used as stylistic devices that communicate aspects of social identity within the United States in the same way as do characteristic phonological or lexical features, is plausible, and recent work in sociophonetics has considered the indexicality of prosodic features (e.g. Kendall, 2013; Podesva, 2007; Thomas, 2011; Yuasa, 2010). While indexicality could explain the existence of regional variation in articulation rate in general, it does not explain why particular regions exhibit faster or slower articulation rates. In order to further investigate this relationship, evidence would be needed in the form of a model that incorporates not only articulation rate and location, but also social identity parameters, as well as the relative frequencies of phonological, grammatical, or lexical features that are known to be regionally distributed.

In terms of the cognitive processes that underlie the temporal organization of speech, it has been suggested that prosodic features such as tempo or pause duration are linked to cognitive activity during speech production (Tsao & Weismer 1997; Tsao et al. 2006). The envelope of variation for articulation rate for an individual speaker (i.e. the ratio between his or her normal and fastest possible rate) is relatively stable (Tsao & Weismer 1997; Tsao et al. 2006), presumably resulting from neurological factors that likely have a biological and genetic component. While it is beyond the scope of this study to consider physiological and neurological correlates of articulation rate, some previous research has found differences between ethnic groups for articulation rate or for other articulatory timing components (Kendall, 2013).

Language interference effects may also play a role. Amino and Osanai (2011) found that second-language articulation rate was positively correlated with first-language articulation rate. Yuan et al. (2006) found that for non-L1 fluent speakers of English, speaking rate in English was dependent on their L1. Similarly, for English speakers in London with immigrant backgrounds, Torgersen and Szakay (2012) found that not only was speech timing affected by the other languages they used regularly, but faster articulation rates were associated with changes in speech rhythm, as measured by the normalized Pairwise Variability Index (PVI), a ratio based on vowel duration in adjacent syllables (Ling, Grabe, & Nolan, 2000). Differences in syllable timing were found between Anglo and non-Anglo London speakers, with Anglo speakers tending towards stress timing and non-Anglo speakers towards syllable timing (Torgersen & Szakay, 2012, p. 829). For American English, Thomas and Carter (2006) analyzed speech rhythm in historical recordings of African-American ex-slaves and recent recordings of speakers of different ethnicities, finding that while speech rhythm does not differ significantly between European-Americans and African-Americans at present, it may have in the past.

In this study, the regional pattern of Getis-Ord G_i^* values evident in Figures 14 and 16 bears resemblance to the geographical distribution of African-Americans in the United States, raising the possibility that regional articulation rate differences may in part reflect historical legacies of cultural, ethnic, and linguistic contact: in the American South between English and the African languages spoken by slaves, and in the Upper Midwest between English and the languages most often spoken by European settlers (often

German or Scandinavian languages). At this point such an interpretation is highly speculative, but, along with other types of evidence, a corpus with word timing information may be the first step towards an analysis along these lines; one that has been annotated with speaker social and demographic identity attributes may allow such questions to be investigated in the future.

As for the relationship between town, city, or county population and speech rate, the data in this study suggest a weak but positive correlation (r = .16, p = .00066), lending some credence to popular perceptions that inhabitants of cities speak faster than rural people. It may be the case that residents of urban areas have more face-to-face encounters with potential interlocutors in their daily lives than do persons residing in more rural areas, and thus experience more demands on their (limited) time. By imparting certain types of information with a higher articulation rate, urban speakers may pursue a time optimization strategy. As is the case with the regional distribution of articulation rates, further information about individual speaker attributes would be necessary in order to propose such an interpretation – in this case, possibly complemented by measures of conversational density such as words spoken per day (cf. Mehl et al., 2001; Mehl, 2017).

Several caveats must be offered for the analysis: First, although the corpus was designed to provide a snapshot of language use in comparable contexts in many American locations, demographic factors which may also affect articulation rates, such as speaker age, gender, or length of residence in a particular location, have not been taken into account. Second, although most of the videos for which captions files were downloaded consist of meetings of local government or civic organizations, a variety of different genres are represented in the corpus, including non-conversational speech such as news reports or public service announcements. Because not every video in the corpus was manually inspected, the possibility that some channels consist primarily of non-naturalistic spoken language cannot be ruled out.

Third, recordings of public meetings, although they largely consist of authentic, naturalistic, conversational speech, cannot be considered representative of conversational speech in general. Public meetings represent a specific communicative genre in which interlocutors typically exhibit a somewhat

constrained repertoire of conversational styles: The formality of many types of governmental meetings and the corresponding dearth of those types of conversational interaction common in other domains of daily life, such as the telling of jokes, gossiping, or the use of emotional speech, swearing or profanity, places limits on the interpretations that can be drawn from this corpus material. Fourth, many of the recordings of public meetings include passages of non-conversational speech of a formulaic or reading nature, for example the roll call, the reading of the agenda, or the recitation of the Pledge of Allegiance.

A fifth point concerns the residence locality of the speakers in the videos. In many U.S. states, holders of public office in American municipalities are legally required to be residents of the area they represent (Mazo, 2016), and hence are by definition locals, and not persons who travel long distances to take part in local government meetings. It is also probable that public office holders are more likely to be longer-term residents (and thus have speech patterns similar to the population of that place), rather than recent arrivals: Establishing the social network necessary for election to public office typically takes some time (Buren & McHugh, 1992). Nevertheless, it is possible that some of the speakers in the corpus are new arrivals or temporary residents in the locations that the sampled YouTube channels represent. Because the analysis in this study intends to shed light on the geographical patterning of articulation rate differences in contemporary American English, and not (e.g.) to establish the territorial extent of traditional dialect regions, the inclusion of these speakers does not bias the findings.

Finally, because the calculation of the intra-utterance continuous articulation rate from individual word timestamps excludes all words with calculated durations of 1 second or longer, if some parts of the country exhibit significantly different distributions of word durations, the calculated articulation rates that serve as the basis for regional analysis would be inaccurate. However, due to communicative economy considerations, it is not expected that the proportion of longer words in discourse differs substantially according to geography (Piantadosi, Tily, & Gibson, 2011). A comparison of word duration distributions by channel prior to filtering of pauses/long words also suggested that this condition does not obtain in this data.

6. Summary and conclusion

Previous research into variation in speech and articulation timing processes in American English has mainly been conducted on short samples of recordings of speakers from a small number of locations, limiting the generalizability of findings. In this study, a corpus of automatic speech-to-text transcripts with individual word timings from more than 29,000 hours of natural speech from the YouTube channels of 506 American local governmental and civic organizations has been utilized to explore regional differences in articulation rate in English. Mean articulation rates per channel were calculated from the word timestamps of force-aligned automatic speech-to-text transcripts, and spatial autocorrelation statistics (global Moran's I and the local Getis-Ord G_i^* score) were used to explore regional patterning of values. The results of the geographical analysis largely confirm findings of some previous studies–Southerners articulate somewhat more slowly than Americans from other regions, and Upper Midwesterners somewhat more quickly. The relationship between location population and articulation rate was explored by correlating the natural logarithm of channel location population at city or county level with articulation rate. A weak but significant positive relationship was found between population and articulation rate.

In terms of future work, it may be the case that regional differences in articulation rate are associated with demographic and social identity parameters in a manner that allows correlations to be drawn between articulation rate and frequencies of use of other regionally distributed language features. Because an orthographic transcript exists for the corpus, this type of analysis represents one possible line of future research. In addition, because the corpus is large, it may be possible to examine articulation rate (and, potentially, lexical and grammatical feature frequency) differences according to communicative situation–for example, by filtering the corpus additionally for recordings of council meetings, news-type broadcasts, face-to-face interviews, or other genres.

As for the intriguing possibility that regional articulation rate differences in the United States may in part reflect the legacy of earlier language contact situations which arose due to different regional patterns of settlement and immigration in the United States, much additional work would be necessary before the

hypothesis could be investigated. A first step in this direction would be to annotate the corpus with individual speaker demographic information at utterance level, which would also enable more comprehensive sociolinguistic analysis of the dataset.

While this study offers new methods for the large-scale study of articulation rate variation and contributes to our understanding of the regional patterning of articulation rate in American English, at least in one type of communicative situation (meetings of local governments), much work remains to be done in the investigation of the relationship between speech timing and other prosodic phenomena, lexico-grammatical language features, and parameters of individual identity. In coming years, as the availability of large data sets of audio and video recordings of natural language interaction continues to increase, language researchers will harness advances in automated methods of natural language processing such as automatic speech-to-text transcription and other neural network-based approaches in order to further investigate the history, current state, and future development of American English dialects and how their feature distributions reflect the identities of their past and present users. For now, the ongoing collection and analysis of large data sets of natural language from specific geographic locations allows us to continue to document how the use of English varies in different parts of the United States.

Acknowledgements

The author would like to thank the editors of *Language and Speech* and three anonymous reviewers for helpful suggestions that improved the manuscript, as well as Finland's Centre for Scientific Computing for access to computational resources and storage.

Declaration of Conflicting Interests

None.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. For example, the automatically-created "Topic" channels in YouTube for specific places often contain videos by non-residents such as tourists or other short-term visitors or non-local news and documentary content producers.
2. Scraped from https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population and https://en.wikipedia.org/wiki/List_of_United_States_counties_and_county_equivalents. Various Python script fragments used to create the corpus are available at https://github.com/stcoats/YouTube_Corpus.
3. https://www.youtube.com/watch?v=ON8rdTMh9q8.
4. The personal name has been anonymized in this example transcript excerpt.
5. One can note that the start and end timings (and durations) for individual words, as given in the speech-to-text transcript, do not necessarily correspond exactly to the audio signal – for this reason, the transcripts will not be useful for (e.g.) the study of variation in the duration of particular words or segments. Articulation rates are nevertheless mostly accurate when the text transcript correctly represents the number of syllables in the audio segment and the phonation duration is reliable.
6. For the calculation of G_i^* , the values for the reference point i are included in the weighting function (i.e. j can equal i, or points are considered to be neighbors of themselves). The G_i statistic is calculated under the condition $j \neq i$. Getis and Ord note that "G_i and G_i^* typically convey much the same information" (1992, p. 194).
7. Kendall (2013) reported a mean rate of 4.6 σ/sec. for conversational speech after omitting pauses of 200ms or longer (p. 92); Jacewicz et al. (2011) reported 5.21 σ/sec. and 4.80 σ/sec. for unconstrained speech from Wisconsin and North Carolina, respectively, and Goldman-Eisler (1961) reported a range of

4.4 to 5.9 σ/sec. Verhoeven et al. reported rates between 3.91 and 5.42 σ/sec. for Dutch varieties (2004, p. 303), while Jessen reported 5.19 σ/sec. for spontaneous German conversation (2007, p. 56).

8. The maps are screenshots of an interactive mapping tool publicly available at https://stcoats.github.io/artic_rate_new.html.

References

Agafonkin, V., et al. (2018). *Leaflet* 1.3.4 [Computer software]. https://leafletjs.com/.

Amino, K., Osanai, T. (2015). Cross-language differences of articulation rate and its transfer into Japanese as a second language. *Forensic Science International, 249*, 116–122.

Avanzi, M., Dubosson, P., Schwab, S. (2012). Effects of dialectal origin on articulation rate in French. In Sproat, R. (Ed.), *INTERSPEECH 2012: 13th Annual Conference of the International Speech Communication Association* (pp. 651–654). Portland, OR: ISCA.

Avanzi, M., Obin, N., Bardiaux, A., Bordal, G. (2012). Speech prosody of French regional varieties. In Ma, Q., Ding, H., Hirst, D. (Eds.), *Proceedings of the 6th International Conference on Speech Prosody, Shanghai* (pp. 603–606). Shanghai: Tongji University Press.

Baker, R. E., Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech 52*(4), 391–413.

Bivand, R. S., Pebesma, E., Gomez-Rubio, V. (2013). Applied spatial data analysis with R, Second edition. New York: Springer.

Boersma, P., Weenink, D. (2010). *Praat: Doing phonetics by computer* [Computer software]. http://www.praat.org/

Buren, B.A., McHugh, K.E. (1992). Residence histories and electoral success in the Arizona legislature. *Social Science Journal, 29*(1), 107–118.

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America, 92*, 593–596.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15*, 39–54.

Byrd, D., Saltzman, E. (1998). Intragestural dynamics of multiple phrasal boundaries. *Journal of Phonetics* 26, 173–199.

Cheng, J., Karambelkar, B., Xie, Y. (2018). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.2. https://CRAN.R-project.org/package=leaflet.

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, M., Li, B., Chorowski, J., Bacchiani, M. (2018). State-of-the-art speech recognition with sequence-to-sequence models. *arXiv*:1712.01769v6 [cs.CL].

Cieri, C., Miller, D., Walker, K. (2004). The Fisher corpus: A resource for the next generations of speech-to-text. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R., Silva, R. (Eds.), *Proceedings of the 4th LREC Conference* (pp. 69–71). Lisbon, Portugal: ELRA.

Clopper, C. G., Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech*, *51*(3), 175–198.

Clopper, C. G., Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, *48*, 633–644.

Clopper, C. G., Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, *39*, 237–245.

Clopper, C. G., Smiljanic, R. (2015). Regional variation in temporal organization in American English. *Journal of Phonetics, 49*, 1–15.

Coats, S. (2019). A corpus of regional American language from YouTube. In Navarretta, C. et al. (Eds.), *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference, Copenhagen, Denmark, March 6–8, 2019* (pp. 79–91). Aachen, Germany: CEUR.

De Jong, N.H., Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, *41*(2), 385–390.

Esmukov, K., et al. (2018). *Geopy*. [Python module]. https://github.com/geopy/geopy.

FFmpeg Developers. (2019). *ffmpeg tool* (Version 4.1.3) [Computer software]. http://ffmpeg.org/

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. (1993). *DARPA TIMIT: Acoustic-phonetic continuous speech corpus*. Gaithersburg, MD: National Institute of Standards and Technology.

Garzon, D. (2017). *Exploring Miamians' perceptions of linguistic variation in Miami-Dade County and the state of Florida* (Master's thesis). Retrieved from https://digitalcommons.fiu.edu/linguistics_ma/5/.

Getis, A., Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*(7), 189–206.

Goldman-Eisler, F. (1961). The significance of changes in the rate of articulation. *Language and Speech*, 4(4), 171–174.

Google. (2009). Automatic captions in YouTube. https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html.

Grieve, J. (2011). A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics, 16*, 514–46.

Grieve, J. (2012). A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguistics and Linguistic Theory, 8*, 39–72.

Grieve, J. (2014). 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). Berlin/Boston: De Gruyter.

Grieve, J. (2016). *Regional variation in written American English*. Cambridge, UK: Cambridge University Press.

Grieve, J., Speelman, D., Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change, 23*, 193–221.

Hahn, M., Siebenhaar, B. (2016). Sprechtempo und Reduktion im Deutschen (SpuRD) [Speaking tempo and reduction in German (SpuRD)]. In Jokisch, O. (Ed.), *27. Konferenz Elektronische* Sprachsignalverarbeitung/*27th Conference on Electronic Speech Signal Processing* (pp. 198–205). Dresden: TUDpress.

Harrenstien, K., Toliver, J., Alberti, C., Black-Bilodeau, N. D. (2009). *Generation of timed text using speech-to-text technology and applications thereof*. U.S. Patent No. 8645134B1. Washington, DC: U.S. Patent and Trademark Office.

Hewlett, N., Rendall, M. (1998). Rural versus urban accent as an influence on the rate of speech. *Journal of the International Phonetic Association 28*(1–2), 63–71.

Hilton, N.H., Gooskens, C., Schüppert, A. (2011). Syllable reduction and articulation rates in Danish, Norwegian and Swedish. *Nordic Journal of Linguistics, 34*, 215–237.

Hsuan, Y. C., Remite A., Sergey M. (2018). Youtube-dl [Computer software]. https://github.com/rg3/youtube-dl/blob/master/README.md.

Jacewicz, E., Fox, R. A., O'Neill, C., Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change, 21*, 233–256

Jacewicz, E., Fox, R. A., Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America, 128*(2), 839–50.

Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science and Justice, 47*, 50–67.

Kendall, T. (2013). *Speech rate, pause, and sociolinguistic variation: Studies in corpus sociophonetics*. London: Palgrave-Macmillan.

Kretzschmar, W. A. (2003). Mapping Southern English. *American Speech, 78*, 130–149.

Kretzschmar, W. A., McDavid, V., Lerud, T., Johnson, E. (1993). *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.

Kurath, H., Hansen, L., Bloch, B., Bloch, J. (1939–1943; reprinted 1972). *Linguistic atlas of New England* (3 vols.). Providence, RI: Brown University Press.

Labov, W., Ash, S., Boberg, C. (2006). *The Atlas of North American English*. Berlin/New York: Mouton de Gruyter.

Leemann, A. (2017). Analyzing geospatial variation in articulation rate using crowdsourced speech data. *Journal of Linguistic Geography, 4*, 76–96.

Liao, H., McDermott, E., Senior, A. (2013). Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 368–373). New York: IEEE.

Ling, L. E., Grabe, E., Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech, 43*, 377–401.

Mazo, E. (2016). Residency and democracy: Durational residency requirements from the framers to the present. *Florida State University Law Review, 43*(2), 611–678.

McDavid, R., O'Cain, T. (1980). *Linguistic atlas of the Middle and South Atlantic States* (2 fascicles published before discontinued). Chicago: University of Chicago Press.

Mehl, M. R. (2017). The Electronically Activated Recorder (EAR): A method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science 26*(2), 184–190.

Mehl, M. R., Pennebaker, J. W., Crow, M., Dabbs, J., Price, J. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers 33*(4), 517–523.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*, 17–23.

Moreno, P., Alberti, C. (2009). A factor automaton approach for the forced alignment of long speech recordings. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4869–4872). New York: IEEE.

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

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America, 54*, 1235–1247.

Ord, J. K., Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and application. *Geographical Analysis, 27*(4), 286–306.

Pederson, L., McDaniel, S. L., Adams, C. M. (1986–1993). *Linguistic Atlas of the Gulf States* (7 vols.). Athens, GA: University of Georgia Press.

Piantadosi, S. T., Tily, H., Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America 108*(9), 3526–3529.

Plug, L., Smith, R. (2018) Segments, syllables and speech tempo perception. In Klessa, K., Bachan, J., Wagner, A., Karpiński, M., Śledziński, D. (Eds.), *Proceedings of the 9th International Conference on Speech Prosody* (pp. 279–283). Poznań, Poland: Adam Mickiewicz University Press.

Podesva, R. (2007). Phonation type as a stylistic variable: the use of falsetto in constructing a persona. *Journal of Sociolinguistics 11*, 478–504.

Preston, D. (1999). A language attitude approach to the perception of regional variety. In Preston, D. (Ed.), *Handbook of Perceptual Dialectology vol. 1* (pp. 359–373). Amsterdam/Philadelphia: John Benjamins.

Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of the Acoustical Society of America 35*(3), 353–362.

Ray, G. B., Zahn, C. J. (1990). Regional speech rates in the United States: A preliminary analysis. *Communication Research Reports, 7*(1), 34–47.

Roach, P. (1998). Myth 18: Some languages are spoken more quickly than others. In Bauer, L., Trudgill, P. (Eds.), *Language myths* (pp. 150–158). London/New York: Penguin.

Sainath, T., Vinyals, O., Senior, A., Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4580–4584). New York: IEEE.

Szmrecsanyi, B. (2011). Corpus-based dialectometry: A methodological sketch. *Corpora* 6(1), 45–76.

Szmrecsanyi, B. (2013) *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge, UK: Cambridge University Press.

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

Tauberer, J., Evanini, K. (2009). Intrinsic vowel duration and the post-vocalic voicing effect: Some evidence from dialects of North American English. In Uther, M., Moore, R., Cox, S. (Eds.), *Proceedings of Interspeech 2009, 6–10 September 2009, Brighton, UK* (pp. 2211–2214). Brighton: ISCA.

Thomas, E. R. (2011). Sociophonetics: An Introduction. New York/Basingstoke, Hampshire: Palgrave Macmillan.

Thomas, E., Carter, P. M. (2006). Prosodic rhythm and African American English. *English World-Wide, 27*, 331–355.

Torgersen, E., Szakay, A. (2012). An investigation of speech rhythm in London English. *Lingua 122*, 822–840.

Tsao, Y.-C., Weismer, G. (1997). Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research, 40*, 858–866.

Tsao, Y.-C., Weismer, G., Iqbal, K. (2006). Interspeaker variation in habitual speaking rate: Additional evidence. *Journal of Speech, Language, and Hearing Research, 49*, 1156–1164.

Turner, R. (2019). deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation. R package version 0.1-16. https://CRAN.R-project.org/package=deldir.

United States Census Bureau. (2017). Subcounty Resident Population Estimates: April 1, 2010 to July 1, 2017. [Data set]. https://www2.census.gov/programs-surveys/popest/datasets/2010-2017/cities/totals/sub-est2017_all.csv (accessed 14 March 2019).

Verhoeven, J., De Pauw, G., Kloots, H. (2004). Speech rate in a pluricentric language: a comparison between Dutch in Belgium and the Netherlands. *Language and Speech 47*(3), 297–308.

Voronoi, G. (1907). 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.

Wenker, G. (1878): *Sprach-Atlas der Rheinprovinz nördlich der Mosel sowie des Kreises Siegen.* [Language-Atlas of the Rhine Province north of the Mosel and of the district of Siegen]. Marburg, Germany: Self-published.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, F., Stocke, A., Yu, D., Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 25*(12), 2410–2423.

Yuan, J., Cieri, C., Liberman, M. (2006). Towards an integrated understanding of speaking rate in conversation. *Proceedings of Interspeech 2006, Pittsburgh, PA* (pp. 541–544). Pittsburgh, PA: ISCA.

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech 85*(3), 315–337.

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods, 50*, 2597–2605.