Scandinavian Studies in Language, 14(2), 2023 - https://tidsskrift.dk/sss

DIATOPIC VARIATION IN DIGITAL SPACE: WHAT TWITTER CAN TELL US ABOUT TEXAS ENGLISH DIALECT AREAS

Axel Bohmann
University of Freiburg
axel.bohmann@anglistik.uni-freiburg.de

Abstract
The availability of large amounts of social media text offers tremendous potential for studies of diatopic variation. A case in point is the linguistic geography of Texas, which is at present insufficiently described in traditional dialectological research. This paper summarises previous work on diatopic variation in Texas English on the basis of Twitter and presents an approach that foregrounds functional interpretability over a maximally clear geographical signal. In a multi-dimensional analysis based on 45 linguistic features in over 3 million tweets from across the state, two dimensions of variation are identified that pattern in geographically meaningful ways. The first of these relates to creative uses of typography and distinguishes urban centres from the rest of the state. The second dimension encompasses characteristics of interpersonal, spoken discourse and shows an East-West geographical divide. While the linguistic features of relevance for the dimensions are not generally considered in dialectological research, their geographic patterning reflects major tendencies attested in the literature on diatopic variation in Texas.[1]

Keywords: Texas English, diatopic variation, dialectology, Twitter, computer-mediated discourse, register, multi-dimensional analysis

[1] I am grateful to Alex Rosenfeld for sharing his data with me. This work was initially presented at a panel on Twitter in sociolinguistic research at NWAV 49, organised by Stef Grondelaers and Jane Stuart-Smith. I would like to thank both of them for giving me this opportunity and the attendees of the panel, especially Lars Hinrichs and Alex Rosenfeld, for fruitful discussion. Finally, my gratitude goes to Erling Strudsholm and Anita Berit Hansen for their invitation to participate on the Coseriu Symposium and their patience in organising this special issue.

1. Introduction

The availability of large amounts of social media text, geotagged with precise latitude-longitude information, offers tremendous potential for studies of what Coseriu (1955) refers to as diatopic variation, i.e. spatially stratified language use. Often, such research adopts the view of lexical variation, establishing how individual word use creates dialect areas. While such a perspective is statistically powerful in that it can draw on the full range of words contained in a data set (and their collocational behaviour), it sometimes sacrifices interpretability. Specifically, whether differences in word use reflect meaningful linguistic differences or are rather tied back to local discourse referents (such as sports teams, place names, etc.) is not always easy to establish.

The present paper proposes an analysis of diatopic variation in Texas English based on a corpus of Twitter messages from across the state, but focuses its attention on specific linguistic features attested in the literature to play an important role in distinguishing registers and lects. Based on frequency information for each of these 45 features in the subcorpus of tweets for each Texas voting precinct, a multi-dimensional analysis (MDA; Biber 1988) is run to identify fundamental dimensions of variation in the data. Variation along these dimensions is interpreted in geographic and social terms and shown to be meaningfully correlated with these factors. While the restriction of the analysis to a pre-selected set of features loses much of the word-level information other methods can draw on to establish dialect areas, the method gains in interpretability since each feature is associated directly with functional or stylistic motivations.

2. Regional variation in Texas English

Diatopic variation in Texas English has been the subject of as much controversy as it has received attention. Throughout the second half of the twentieth century, repeated attempts were made to delineate Texas English dialect areas, with a significant amount of disagreement among individual authors (Underwood 1990). The first dialect border drawn across the state of Texas appeared in Baugh's (1935:447) map of US dialect regions and divided the state into an Eastern and a Western part, by a line extending in a roughly northerly direction from Victoria on the Gulf Coast in the South to Sherman in the North (see Figure 1). This basic division between East and West Texas is retained in many other dialect maps, although the precise location of the border is subject to a great degree of variation.

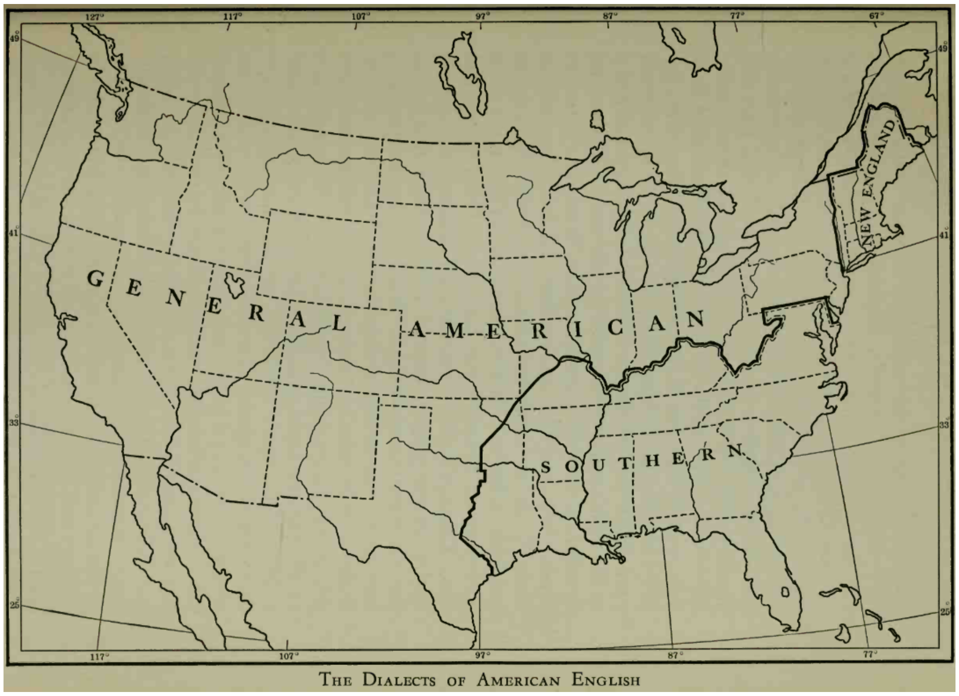

Figure 1: American English dialect regions, according to Baugh (1935: 447).

Most later attempts at identifying Texas English dialect areas were heavily indebted to the *Linguistic Atlas of the United States and Canada*, the first major project in North American dialectology. This project established a tripartite division of American English into a Northern, a Midlands, and a Southern dialect area, initially based on lexical evidence from the North East and the states along the Eastern Seaboard (Kurath 1949), but later replicated in further regional studies associated with the *Linguistic Atlas* project (see Grieve 2016:1–8 for an overview). Bagby Atwood, surveying Texas English for the *Linguistic Atlas*, came to the conclusion that Texas speech was characterised by a mixture of Southern and Midlands dialect words, with the addition of Spanish loans, and did not show clear indication of internal dialect boundaries: “I will not draw lines showing the limits of Southwestern or of any of its subareas. Far too many lines have been drawn already, probably by popular demand and certainly on insufficient evidence" (Atwood 1962:98–99). However, the extrapolation of dialect areas from the individual projects associated with the *Linguistic Atlas* to the entirety of the United States has often led to the postulation of a boundary between Midlands and Southern dialects extending across the state of Texas. A composite map produced by Grieve (2016:4) to represent the joint findings of the *Linguistic Atlas* (reproduced here as Figure 2), for instance, dissects Texas into a Southeastern (Southern dialect area) and a Western (Midlands) part.

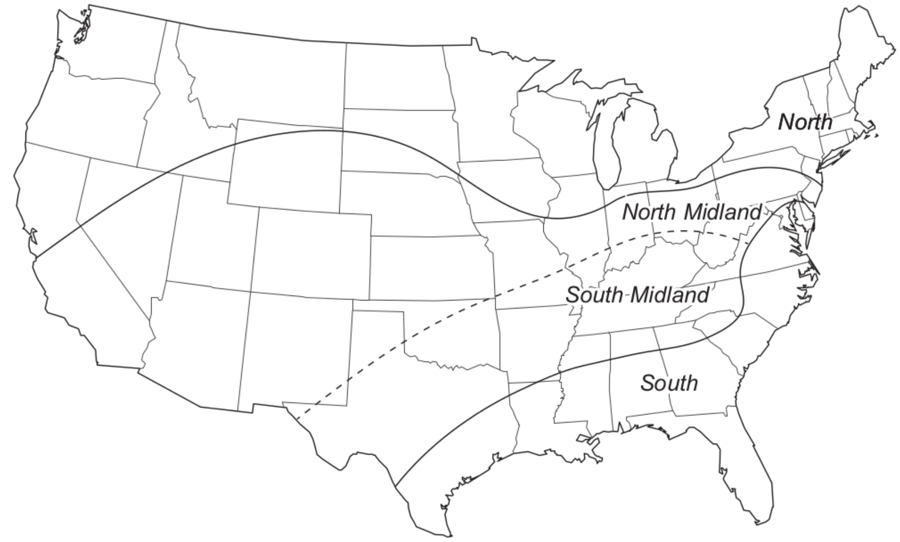

Figure 2: Dialect areas established by the Linguistic Atlas project (composite map from Grieve 2016:4).

In his survey of fifteen different attempts to sub-divide Texas into dialect areas, Underwood (1990) takes a pessimistic, if not cynical, stance. Neither do the individual sources converge on any common pattern, he argues, nor do they give an indication of any reliable evidence for the lines they establish:

> Some of these representations of Texas dialects are, at best, based upon scant and sometimes dubious evidence. More commonly, they are based on nothing more than conjecture, speculation, or personal whim. And even a few scholars have justified their dialect maps of Texas on the basis of earlier research by linguists who had previously concluded that their evidence did

> not warrant establishment or even postulation of dialect boundaries in Texas. (Underwood 1990:96)

Underwood's point, crucially, is not that it is implausible per se to map dialect areas of Texas. Rather, he takes issue with the paucity of evidence on whose basis such mapping is conducted and the lack of accountability in the literature he cites. Since the publication of his article in 1990, major empirical advances have been made, however. Chief among these is the *Atlas of North American English: Phonetics, Phonology and Sound Change* (ANAE; Labov et al. 2005), which maps phonological variation across the nation based on telephone interviews with 762 informants. The phonologically based dialect regions established therein see Texas squarely belonging to the South, an area characterised by monophthongization of the PRICE vowel and the Southern Vowel Shift. Grieve (2016:210, 218) as well, in his study of regional grammatical variation in American English, based on a large corpus of letters to the editor, produces maps which leave Texas largely intact as a unified whole, bar its westernmost tip around El Paso.

It would appear, then, that recent evidence converges on the uniformity of Texas English. Atwood's reluctance to draw dialect borders within the state, Underwood's criticism, and the tendency for Texas English to be subsumed under one dialect region in recent, large-scale studies all appear to point in this direction. It has to be remembered, however, that diatopic variation is an inherently scalar notion: far from being limited to a finite set of large-scale dialect regions, variation is pervasive in language down to small-scale individual differences (Kretzschmar 2015). Indeed, this variability and the complex settlement history, rather than any simple notion of uniformity, is what inspired Atwood's scepticism of clear borders. It should also be noted that, while the individual studies derided by Underwood (1990) differ in the precise borders they draw, there is overwhelming convergence between them in regard to a general East-West division of Texas English.

There certainly is evidence for internal linguistic differentiation within the state that warrants closer attention. Bailey (1991), providing the kind of careful empirical evidence whose absence Underwood (1990) laments in other studies, shows two important dimensions of distinction in ongoing changes in Texas English. The first is a regional distinction between East and West Texas and the second an increasing tendency for urban and rural areas to take diverging courses of linguistic development. A study by Cukor-Avila and colleagues (2012) complements this picture from a perceptual perspective. Based on dialect maps drawn by 367 informants from across the state, they establish a complex set of dimensions along which Texans perceive linguistic differences within the state. They conclude that “Texans do not view themselves as a homogeneous speech community, nor do they consider Texas to be the land of cowboys and hillbillies”. (Cukor-Avila et al. 2012:18)

While the integrity of Texas English from the coarse perspective of large-scale US dialect regions may be taken as consensus, then, there remains a good deal of internal variability at a finer level of granularity that invites further exploration. The associated challenges are: a) the complexity of the socio-linguistic fabric of the state, where complex settlement histories, urbanisation, language contact with Spanish, etc. are likely to engender variation along multiple dimensions of differentiation and b) the necessity of large amounts of data, ideally with good geographical resolution and additional social information to model this complexity. The following section explores how data from social media, and especially Twitter, may help to address these challenges and summarises extant research in this direction.

3. Mapping diatopic variation with social media data

A fundamental problem in mapping diatopic variation is achieving good geographic coverage. The foundational dialectological surveys reviewed in the previous section, such as the *Linguistic Atlas* and the ANAE, involved a staggering amount of resources to achieve this task. Yet, even ANAE only represents 145 cities systematically, with additional data points opportunistically incorporated during the data collection procedure (Labov et al. 2005:23).

Data from the social media service Twitter offer the potential to achieve much higher geographic resolution – in principle down to the precise latitude and longitude coordinates a tweet was sent from, at a fraction of the cost. With 821 million daily tweets in 2020 (a number that has steadily increased from 340 million in 2012; Yaqub M. 2022), many of which are geotagged, it has become possible to use big-data approaches to investigate linguistic developments in real time and across space at a level of detail dialectologists of the twentieth century could not have foreseen.

The fact that tweets can be directly harvested from Twitter's application programming interface (API) in a structured format that lends itself to downstream analysis further enhances the possibilities of doing Twitter-based dialectological research.

Data of this kind have been used in a number of previous studies. Takhteyev and colleagues (2012) find that Twitter networks tend to coincide with geographical units (such as metropolitan areas) to a high degree, and that translocal network ties on the service are best predicted by variables that indicate the spatial connectedness among places, such as frequency of air travel between them. These observations lend credibility to treating Twitter not as a locally undifferentiated virtual space, but as an arena in which regionally specific linguistic processes are likely to find articulation. Eisenstein and colleagues (2010) as well as Russ (2012) both report important geographical patterning of lexical variation on Twitter in the USA. The former authors identify new sets of topic-sensitive words with regional specificity, whereas Russ applies known indicators of dialect variation (such as *soda*, *pop*, and *coke* as generic terms for carbonated sweet beverages) to Twitter data and proves that corresponding findings replicate those derived from linguistic atlas data. Lexical variation between British and American terms is further studied on a global level in Gonçalves et al. (2018). Eisenstein and colleagues (2014) further comment on the utility and methodological challenges of large-scale social media data for identifying regional dialects. A number of studies have mapped English dialect areas based on Twitter data. This is done for British English on a lexical basis by Grieve and colleagues (2019) and on the basis of a well-known grammatical alternation, the ditransitive, by Stevenson (2016). Strelluf (2020) maps one particular, low-frequency grammatical feature (NEED/WANT + past participle) in tweets from cities across the English-speaking world. The diatopic perspective may also productively be combined with a diastratic one, e.g. in research focusing on regional variation in African American English (Austen 2017; Jones 2015). A useful overview of the potentials of social-media dialectology is provided by Eisenstein (2018).

Before discussing some of the potential issues in using Twitter data to establish dialect areas, it is useful to summarise one particular approach that has been successfully applied for analysing diatopic variation in Texas English. Rosenfeld (2019) employs a modified variant of the method proposed in Hovy and Purschke (2018) to model differences in word usage across Texas voting districts. The central element of the method is a distributional model of word usage, a so-called word embedding model. Such models learn information about word usage based on large corpora of text, which they use to create similar vector representations for words with similar usage properties in a high-dimensional space. Lexical items that occur in similar contexts, i.e. with similar collocates, are located in close proximity in the vector space. Hovy and Purschke apply a version of such embeddings, the doc2vec algorithm (Le & Mikolov 2014), which learns representations for (in their case) different cities in the same space in which the words are represented, based on the words that occur in tweets associated with a given city. For instance, the four nearest neighbours in embedding space to the city of Vienna are the words *leiwand*, *ur*, *bissi*, and *oida*, all of which are clearly recognisable Viennese dialectal forms (Hovy & Purschke 2018:4390). This allows a) for similarity relationships among cities to be expressed mathematically and b) for the association of particular words with particular regions to emerge. The innovation in Hovy and Purschke (2018), a study of German dialect areas, lies in their application of retrofitting. In order to achieve a smoother geographic signal, they update the vector for each city with information from surrounding cities. This allows them to model German dialect regions in good accordance with independent dialectological work and with the potential for analysis "at a finer granularity than was previously possible" (Hovy & Purschke 2018:4383).

Rosenfeld makes important methodological contributions to further refine this method, which are, however, not the main focus here (for details, see Rosenfeld 2019:87–94). What is important are his results for Texas English. Based on a corpus of 2.3 million tweets, each mapped onto one of the 8,000 voting precincts in the state, Rosenfeld establishes a (smoothed) vector representation for each voting precinct in word/document embedding space and subsequently subjects this data set to agglomerative clustering (Ward 1963). The result are the nine different dialect regions shown in Figure 3.

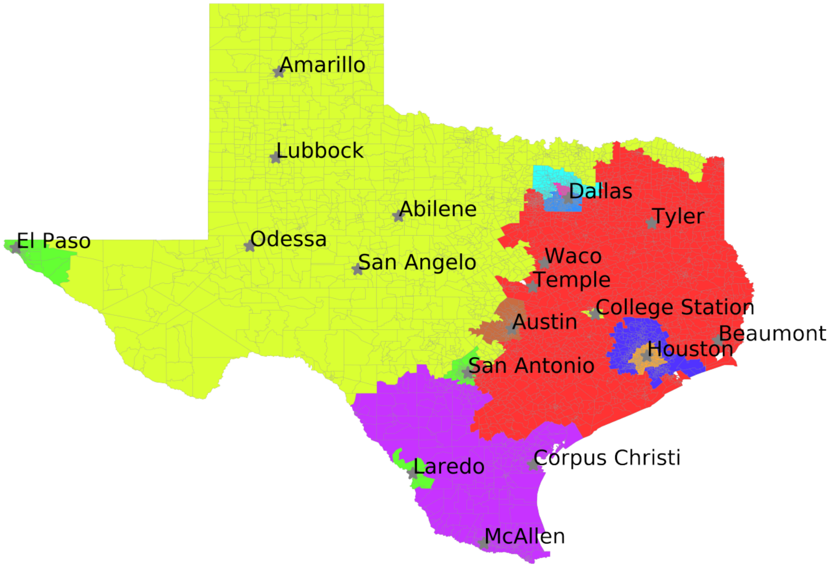

Figure 3: Texas English dialect regions established by Rosenfeld (2019: 111).

The map both reflects and gives a more precise shape to many observations that have been made about variation in Texas English in the past. For instance, the pervasive notion of an East-West division is reflected in the line separating the yellow (West Texas) precincts from the large areas of red (East Texas) and purple (South Texas). Bailey's (1991) observations regarding the increasing role of urbanisation are also directly reflected in the map, with each of the major cities (Dallas, Houston, Austin, San Antonio) being identified as a separate cluster, and additional clusters for the suburbs of Dallas and Houston. San Antonio is interesting because it clusters together with El Paso and Laredo. All three cities are quite distant from each other, so geographical proximity cannot be the reason for their clustering together. Rosenfeld (2019:101) explains this behaviour with respect to the high percentage of Hispanic populations in all three cities, showing that the method not only capitalises on diatopic but also diastratic variation.

Figure 3 amply demonstrates the power of combining large amounts of geo-tagged social media data with advanced machine-learning methods to achieve high-resolution dialectological maps. In interpreting the results linguistically, however, some issues arise that traditional, survey-based dialectology does not run up against. Some of these have to do with the drawbacks of social-media data in general. First, while careful selection of participants is an important step in more traditional work, and precise information about many socio-demographic variables pertaining to each informant may be collected, Twitter data is "demographically lean" (Iorio 2009). Information about the typical social characteristics considered in twenty-first-century dialectology – age, gender, ethnicity, socio-economic status – is not systematically represented for Twitter users. The most common response has been to use census-level statistics for each district or voting precinct as a stand-in for individual users' information (Eisenstein 2015; Rosenfeld 2019).

Next, there is the issue of representativeness. It is obvious that use of social media skews towards certain segments of the population and that Twitter discourse is thus not fully representative of offline, vernacular language use. However, the precise relationship between these two concepts is rarely explored (although see van Halteren 2021 for a critical perspective), in part due to the fact that a full description – or even faith in the existence – of such a thing as offline, vernacular usage is anything but self-evident. The response to both these points of criticism has often been that the proof is in the pudding: in big corpora of Twitter language, attested patterns of offline use are replicated (e.g. Hovy & Purschke 2018) and district-level social information acts as a meaningful predictor of variation (Eisenstein 2015).

There is, however, a further issue that is more germane to the precise method used in Rosenfeld (2019). Unlike traditional dialectological surveys, which work with closed lists of (lexical, phonological, less often: grammatical) features that are independently considered to be relevant for diatopic variation, word embeddings draw on all of the words in the data. This is not strictly a disadvantage. Labov, Ash, and Boberg (2005:7), for instance, point out the problem inherent in fixed sets of elicitation items, which often show "a heavy concentration of rural and agricultural terms and other words and phrases that are no longer current. The stream of evidence for dialect differentiation therefore shrinks over time and contributes to the widespread impression that regional dialects are disappearing." A word embedding method does not suffer from such shrinkage and is arguably able to identify relevant distinctions without a researcher's bias towards certain forms contaminating the analysis (Hovy & Purschke 2018:4383).

On the other hand, carefully designed lists of features retain interpretability that is lost in distributional models. It is not easy to account for the precise patterns of usage that cause the clustering of dialect areas in Figure 3 in general. Neither, and this is perhaps a more fundamental issue with such methods, is it clear to what extent the models capitalise on meaningful linguistic – compared to relatively trivial – referential variation. To what extent is the clustering driven by words with local referents, such as sports teams, the names of streets, neighbourhoods, local celebrities, or venues? Traditional dialectological methods are able not only to identify dialect regions but also, by considering the individual features that contribute to them, to come up with linguistic explanations for these regions. Recall, for instance, that the *Linguistic Atlas* project was able to identify settlement history as a key factor leading to different US dialect regions. In distributional approaches, post-hoc analyses may uncover similar patterns (Rosenfeld 2019:98–102), but the question of interpretability of the clusters, and the extent to which they are based on meaningful structural variation, largely remains.

In the remainder of this paper, I discuss a method that draws on the same corpus of Twitter messages used in Rosenfeld (2019), but takes as its point of departure a set of pervasive linguistic features established previously to indicate important differences in discourse structure. I show that such a method, while lacking the sophistication and fine granularity of distributional models, is able to uncover dimensions of diatopic and diastratic variation in Texas English, and that these dimensions are amenable to linguistic interpretation. This method is proposed not as a competitor to distributional approaches such as Rosenfeld (2019), whose statistical sophistication far surpasses my proposal here. Rather, the two approaches usefully complement each other: distributional models excel at uncovering regional structure in the data, whereas the method below has its strengths in identifying fundamental aspects of linguistic variation of relevance for such regional structure.

4. Data and methods

The corpus for the present analysis is the one used in Rosenfeld (2019). It consists of 3,511,253 individual tweets, each of which is associated with one of the over 8,000 Texas voting precincts. Locating these tweets in geographical space is done by precise latitude-longitude coordinates, wherever available. In addition, tweets without such information but with an identifiable Texan town in the "place" field were also included. The coordinates for locating the latter tweets were derived from the town via data from simplemaps.org (Rosenfeld 2019:85). The data has been pre-processed such that individual elements of the tweets – primarily user handles and urls – have been replaced with generic placeholders (<url>, <user>, etc.). Other than the texts of the tweets themselves and their tweet IDs, no user-specific information is retained in the corpus. However, detailed information for each voting precinct, collected from the US Census Bureau, is available as an important source of meta-data. Of primary importance for the present study are the following by-precinct pieces of information: the coordinate bounds of the district (required for spatial analysis), the percentage of different ethnicities (White, Hispanic, Black), the percentage of people who voted Republican in the most recent election, and the precinct's population density.

The choice of voting precincts as the socio-geographical unit of analysis over alternatives such as counties or voting districts are motivated by several factors. First, voting precincts are the smallest geographical units for which detailed demographic information is available, thus providing the highest possible spatial resolution while retaining aggregate social information. Second, as opposed to counties, voting precincts are designed in such a way as to keep population size, rather than area, comparable (Rosenfeld 2019:86). Consequently, they allow for inferences about relatively small and specific demographic groups, even in densely populated areas of the state.

The data for each precinct are treated as one corpus text. The resulting corpus of 8,080 texts is subjected to a multi-dimensional analysis (MDA; Biber 1988; Bohmann 2019). This method measures frequency profiles for many linguistic features in each corpus text. In contrast to many approaches in computational sociolinguistics, which automatically detect distinctive features based on a large set of candidate items (often: words; e.g. Louf et al. 2023), MDA follows a different design. Each linguistic feature under consideration is theoretically motivated on the basis of its stylistic and/or discourse-structuring properties. As such, the goal is not finding features that maximise a certain task, such as predicting regional differences, but putting individual texts into linguistically interpretable relations.

MDA rests on the assumption that linguistic variation is, in large parts, functionally motivated and that, consequently, the behaviour of many different linguistic variables is subject to a small set of underlying situational properties. To give a concrete example: personal pronouns are found at higher frequencies in spontaneous face-to-face conversations, which also feature higher-than-average frequencies of private verbs (such as *think*, *believe*, etc.) and contractions. All of these features, in turn, are under-represented in academic writing, where nominalisations, passive-voice constructions, and complex noun phrases are highly frequent. Given a corpus of personal conversations and academic articles, an MDA is able to detect these commonalities and express all six features as part of one dimension of variation. In this case, the dimension expresses a difference between conceptually oral and conceptually literate texts (Koch & Oesterreicher 1985) and might be labelled involved versus informational production (Biber 1988; Bohmann 2019).

Mathematically, such dimensions are established by subjecting the matrix of texts and measured features to exploratory factor analysis (Thompson 2004; Gorsuch 2015). Similar to principal components analysis, this method reduces the dimensionality of the data. Whereas the latter takes common as well as item-specific variance into account, factor analysis only considers common variance, with the aim of establishing generalisable latent dimensions. Based on the covariance profiles of the features, factor analysis identifies groups of linguistic features that behave similarly across the corpus, i.e. that have a tendency to be over- or under-represented in the same texts. In the present study, this means the method finds clusters of functionally related features that occur with greater- or lesser-than-average frequencies in tweets from the same Texas voting precincts.

The MDA procedure expresses the relationship between each feature and each dimension of variation by virtue of a measure of correlation, a so-called structure coefficient. Features with a structure coefficient whose absolute value is above a pre-defined threshold are considered to be relevant for a particular dimension. The method also, importantly, scores all corpus texts along each dimension, allowing for an analysis of how the multidimensional space of variation structures the corpus material itself.

This later part will be central in the present analysis, as relations in multidimensional space can be mapped onto geographical space. The utility of MDA for analysing Twitter discourse has been demonstrated by Clarke and Grieve (2019).

Space limitations prevent any detailed discussion of the individual linguistic features extracted from the Texas English Twitter corpus. I restrict myself to listing the 100 features here (see Table 1) and stating in general terms that they were selected based on the role they have been found to play in previous research on register variation, variation in computer-mediated discourse, or as markers of Texas/Southern speech. Only those features that occurred with a frequency > 0 in at least 75% of the corpus texts/precincts under analysis were retained. The less frequent features that did not meet this criterion are enclosed in parentheses in the list below.

Category	**Items**	**Primary role in structuring variation**
Pronominal forms	first-person pronouns, second-person pronouns, third-person personal pronouns, pronoun *it*, indefinite pronouns	General register variation
Modality and stance devices	possibility modals, prediction modals, (amplifiers, downtoners)	General register variation
Subordination	*if* and *unless*, *although* and *though*, *because*	General register variation
Prepositions and adverbials	all prepositions, time adverbials, place adverbials	General register variation
Features associated with orality	contractions, (*gotta, gonna, wanna, g-dropping*),	General register variation
Prefixation	Prefixes *re-* , *un-* , (*anti-, co-, counter-, ex-, inter-, dis-, mis-, under-/over-, pre-, pro-, semi-, sub-, super-, trans-, uni-, with-)*	General register variation
Suffixation	Suffixes *-ic, -ion, (-able/-ible, -age, -ance, -ant, -ary, -ation, -dom, -ful, -hood, -ial, -ical, -ican, -ify, -ism, -ist, -ity, -ize, -ive, -less,*	General register variation

	-ment, -ness, -ory, -ous, -ship, -tor, -ture)	
Punctuation	*&, !, ?, :, %, *, @* (other than as part of usernames)	General register variation /CMD
Twitter-specific discourse conventions and CMD devices	hahstags, usernames, (*smh, yolo)*	CMD
Emojis and emoticons	tears of joy, heart eyes, loudly crying, ok hand, unamused, heart, kiss, smiling face eyes, weary, raising hands, *:)/:-)*	CMD
Features associated with regionally/socially specific usage	*fixing to*, possessive pronoun + *ass,* (*yall*, *holler*, *th*-stopping, word-final *t/d* deletion*)*	Dialectal and sociolectal variation
Reference to specific entities	date, money, number, time, url	Variation in discourse topic
Further features	Negator *not*	General register variation

Table 1: Measured linguistic features.

An important step of the analysis concerns the decision about how many dimensions to retain, i.e. how aggressively to reduce the original dimensionality of the data. This decision is not a matter of strict mathematical criteria. Theoretical considerations as well as heuristic tools (considering the amount of variance that is still explained while reducing the data to given number of dimensions) play a role in this step. Since all corpus texts in the present study represent one situational variety, Twitter discourse, the register variation found in this data set is likely to be less pronounced than in previous MDA research working with diverse text types. Therefore, and following the heuristic method of a scree plot analysis (Cattell 1966), a four-dimensional solution was selected to represent the variation in the present data set. Together, the four factors account for 45% of the total variance in the measured variables. Factors were established with the principal axes method and rotated to the oblimin criterion, which allows for moderate degrees of inter-factor correlation. The factor scores for individual precincts were calculated via regression.

The analysis below considers the variation along the first two dimensions in the data, which account for 19 and 13% of the total variance in all features, respectively. Dimensions 3 and 4 are less informative, accounting for 7 and 6% of variance, respectively. It should also be noted that the first two dimensions show a significant degree of correlation, with a coefficient of 0.337. They are not to be understood as fully orthogonal, independent dimensions, then, but as inter-related to a relatively large degree.

Each dimension is discussed separately below. First, the relevant linguistic features are explained and interpreted in functional terms. Next, the dimension's geographic patterning is explored by constructing a map of Texas in which each voting precinct is color-coded on a gradient that corresponds to its dimension score. This allows for an intuitive understanding of the diatopic profile of each dimension. Finally, the following demographic variables available for each precinct are considered as predictors in a regression model fitted over the data, with the by-precinct dimension score as the outcome to be predicted: population density (representing the degree of urbanisation), the percentage of Hispanic and Black people in the population, and the percentage of the Republican vote. All predictors were logarithmically transformed in order to arrive at more normally distributed values. Additional factors of potential relevance, such as age structure of median household income, were unfortunately not available for this data set. As such, the discussion is restricted to the above variables.

5. Results

5.1. Dimension 1

The first dimension, accounting for almost half of the total feature variance, shows structure coefficients greater than 0.5 for 25 different features, and no salient negative coefficients. The relevant features and their structure coefficients are listed in Table 2. What is immediately striking in this list is the predominance of different emojis. The top seven items are all of this kind, and they altogether make up ten of the 25 features listed. Other strategies drawing on individual orthographic symbols – the "happy" emoticon *:)*, asterisks, the ampersand, the at-sign (where it is not part of a username), question marks, and colons – make up a significant portion of the remaining fifteen features. Only a handful of lexico-grammatical features are associated strongly with the first dimension. These are subordinators *because*, *if*, and *unless*, place adverbials, prediction modals, and indefinite pronouns as well as the pronoun *it*. Two suffixation devices (adjectival -*ic* and nominal -*ion*) round up the list.

emoji_loudly_crying (0.69)	emoji_kiss (0.65)	emoji_raising_hands (0.64)	& (0.61)	Pronoun *it* (0.56)
emoji_heart_eyes (0.69)	emoji_weary (0.65)	-*ic* (0.64)	@ (0.61)	? (0.56)
emoji_unamused (0.67)	*:)/:-)* (0.65)	-*ion* (0.64)	Place adverbials (0.60)	Indefinite pronouns (0.56)
emoji_tears_of_joy(0.67	*because* (0.65)	emoji_smiling_face_eyes (0.63)	Prediction modals (0.59)	*:* (0.54)
emoji_ok_hand (0.66)	*if/unless* (0.64)	* (0.63)	emoji_heart (0.59)	Money (0.51)

Table 2: Salient features for Dimension 1 (structure coefficients in parentheses).

The primary functional interpretation for this first dimension is relatively easily established with reference to the typographic elements that predominate the list. Texts scoring highly on this dimension are characterised by a particular kind of multimodal online writing that draws on spelling and pictographic strategies as additional semiotic resources in the construction of a tweet. This focus on semiotic creativity is, however, not to be equated with a generally interpersonal, “conversational” style. While third person pronouns (*it* as well as indefinite ones) reach salient structure coefficients, this is not true for second- and first-person pronouns, the latter of which actually show a negative structure coefficient (-0.04). Other markers of a colloquial and interpersonal style, such as user mentions and contractions, are also ranked relatively low in terms of their structure coefficients (37 and 40 out of 45, respectively).

How is this creative style of online writing associated with the geography of Texas? Figure 4 illustrates the answer to this question in the form of two maps. Each shows the state of Texas, with every voting precinct represented as a separate polygon and receiving colour shading based on two properties. In the left panel of Figure 4, each precinct is shaded according to log-transformed population density. Darker shades of brown indicate higher population density, whereas darker shades of blue indicate more sparsely populated precincts. The right panel shows shading according to the dimension score along the first dimension calculated during the MDA, again with darker browns indicating higher-scoring precincts and darker blues indicating particularly low dimension scores. The similarities between the two maps are striking, although the left panel appears to include more room for intermediate light zones not clearly shaded blue or brown. These represent the wider suburban areas surrounding the big cities in East Texas. The right panel shows high dimension scores clearly concentrated around the city centres. The northernmost cluster of predominantly brown shades is the Dallas-Fort Worth metropolitan area. Going South from there, a string of brown cells extends along ~~the~~ Interstate 35, with concentrations in Austin and San Antonio. Further East, towards the northern end of the Gulf Coast, the city of Houston is also clearly marked in brown.

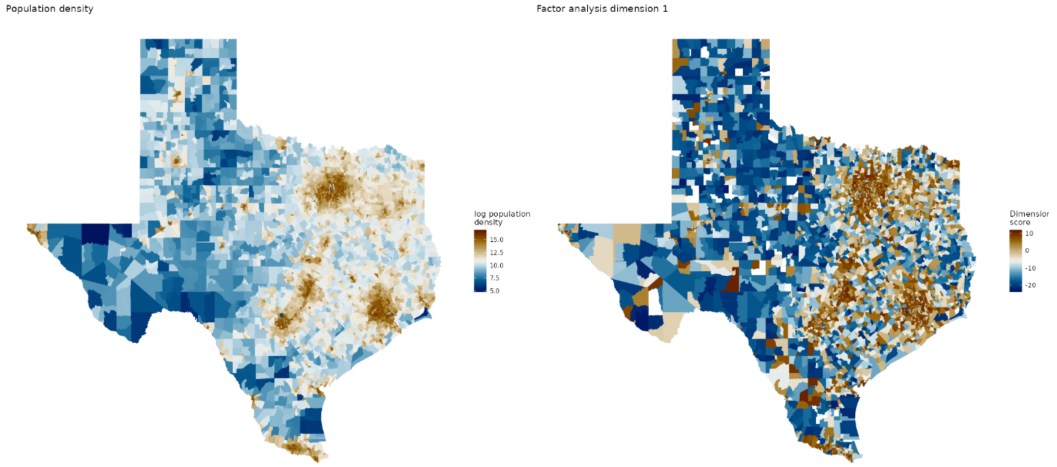

Figure 4: Distribution of log population density (left) and Dimension 1 score distribution (right).

The importance of urbanity for Dimension 1 is further confirmed by a regression model fitted over the data to predict a precinct's Dimension 1 score from its population density, percentage of Black, percentage of Hispanic inhabitants, and percentage of people who voted Republican. The coefficients of the model are shown in Figure 5, created with the dotwhisker (Solt & Hu 2018) package in R (R Core Team 2022). The black vertical line is the intercept, which is simply the model's base prediction and of no immediate value for interpreting the coefficients. Each dot is the estimated coefficient for the effect of a predictor variable. Positive values, i.e. dots located to the right of the intercept, indicate an increase in predicted dimension 1 scores as the value for the predictor variable increases. Negative values, to the left of the intercept line, indicate an inverse relationship between increase in the predictor and outcome variable. The whiskers extending horizontally from the coefficient dots represent 95% uncertainty intervals. If these do not intersect the vertical line, this can be taken as evidence for a predictor's statistical significance in the traditional sense.

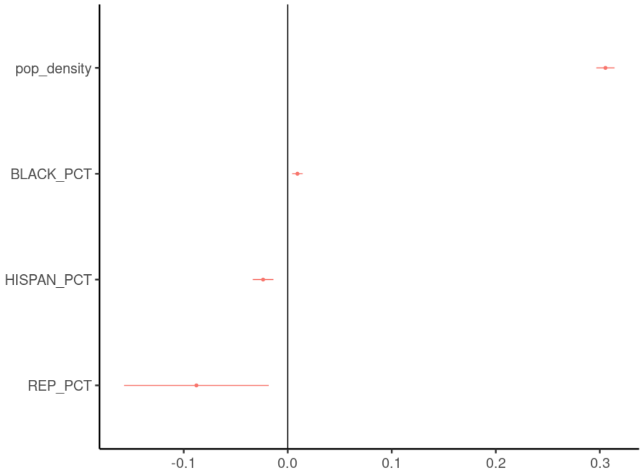

Figure 5: Coefficient estimates for a linear regression predicting Dimension 1 score.

While all predictors in the model reach significance, it is clearly population density which dominates in terms of effect size. More densely populated areas, such as the urban centres mentioned above, are predicted to show drastically higher values along dimension 1. In terms of ethnicity, higher proportions of Black, but lower proportions of Hispanic speakers, also increase the predicted dimension scores, although with a much smaller effect size. Political leaning is the weakest predictor in the model, with higher percentages voting Republican slightly favouring lower dimension 1 scores.

As far as diatopic variation in Texas is concerned, dimension 1 lends support to the diverging paths urban and rural linguistic developments may be taking (Bailey 1991). Of course, Bailey's work initially established such a divergence with respect to general Southern and Standard features, mostly at the level of phonology. The data underlying dimension 1 are of a fundamentally different kind: semiotic strategies in written computer-mediated discourse, with no initial relation to any particular dialect areas. As such, reading the results of the present section as direct confirmation of Bailey's (1991) point would be far-fetched. What the results do show very clearly, however, is that the rural-urban divide has important consequences for linguistic variation in the state of Texas. Findings from related research in computational sociolinguistics also corroborate this importance at the national level (Louf et al. 2023). Together, these findings point to the importance of the categories rural and urban in linguistic identity construction, potentially replacing traditional regional distinctions as Bailey (1991) observes. This topic deserves continued attention in sociolinguistic and dialectological research (e.g. Hinrichs et al. 2013).

5.2. Dimension 2

The second dimension comprises 16 features with a structure coefficient greater than 0.5, listed in Table 3. Five of these can also be found in Table 2 above: indefinite pronouns and *it*, question marks, colons, modals of prediction, and subordinators *if* and *unless*. However, the seven features with the highest structure coefficients along dimension 2 are unique to that dimension. Contractions and second-person pronouns indicate an involved, colloquial style, as do user mentions. Negator *not* is also associated with characteristics of interpersonal conversation (Bohmann 2019:94–95). The inclusion of additional pronominal forms (first- and third-person personal pronouns, *it*, indefinite pronouns) speaks both to an interpersonal and a less nominal, abstract style of discourse. The focus on modality, indicated by the high structure coefficients for both possibility and prediction modals, is not easy to interpret in the abstract, as a differentiation between epistemic and deontic uses would be required for a meaningful interpretation. Given the patterning of all other features, it may be speculated that it is predominantly deontic modality, used to indicate rights and obligations in interpersonal communication, that is captured by dimension 2. Finally, the inclusion of exclamation points and question marks in the list attests to a level of involvement captured by the second dimension.

Contractions (0.76)	Time adverbials (0.65)	*!* (0.63)	*:* (0.56)
Second-person pronouns (0.71)	Possibility modals (0.65)	*?* (0.61)	Prediction modals (0.54)
Negator *not* (0.67)	Third-person pronouns (0.65)	Indefinite pronouns (0.59)	Prefix *re-* (0.53)
User mentions (0.67)	Pronoun *it* (0.65)	First-person pronouns (0.58)	*if/unless* (0.51)

Table 3: Salient features for Dimension 2 (structure coefficients in parentheses).

The geographic distribution of dimension 2 scores can, once again, be analysed visually in Figure 6. This time, no second map is included for comparison, since the main distinguishing feature of dimension 2 can be read from the distribution itself. Whereas Figure 4 showed clear concentrations of brown cells against a predominantly blue background, i.e. high scores for dimension 1 being clearly concentrated in a limited set of urban districts, Figure 6 shows a much wider spread of brown. The distribution of high dimension 2 scores is clearly not limited to the city centres; and, in fact, the downtown areas of Houston, Dallas, San Antonio, and Austin all show lighter brown than the surrounding suburbs. What can be seen, as well, is a noticeable East-West divide, with patches of blue being more common in the West, whereas the East is more uniformly brown. The pattern is far from conclusive, as many deeper shades of brown can also be found in West Texas, particularly along ~~the~~ Interstate 20, moving West from Dallas. Conversely, there are pockets of blue on the Southern Gulf coast and close to the border to Louisiana in the East. Nonetheless, it is justified to speak of a general East-West split, with East Texas generally scoring higher along the dimension 2 continuum.

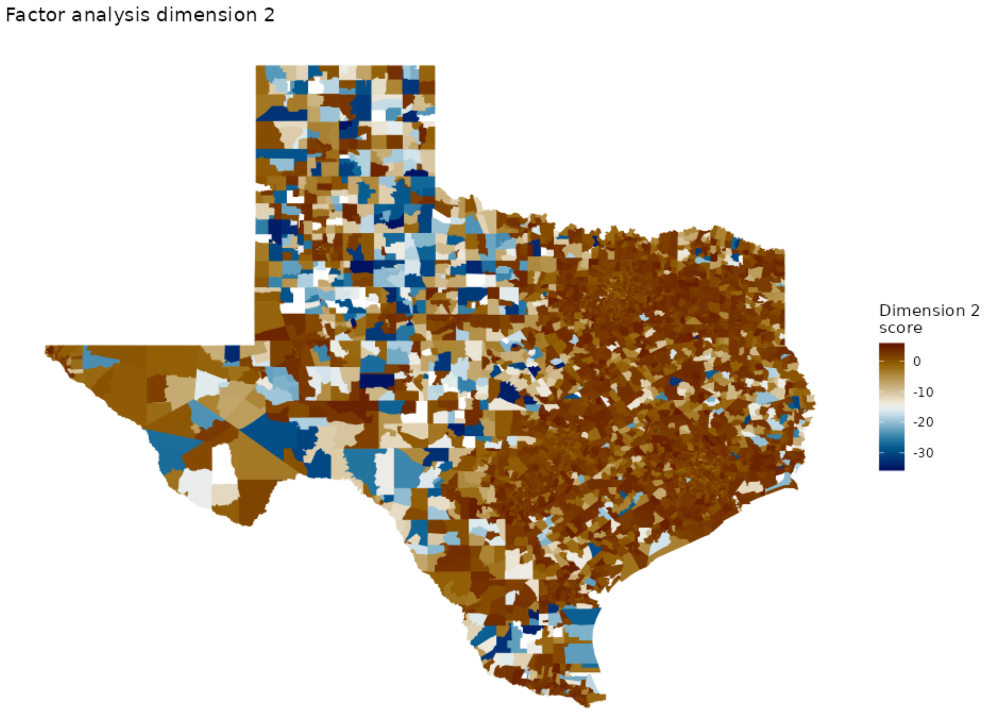

Figure 6: Distribution of scores along Dimension 2. Darker shades of brown indicate higher values, dark shades of blue low values.

A look at the coefficients from a regression model predicting dimension 2 scores (Figure 7) highlights the interrelatedness of both dimensions. All terms, with one exception, show an effect in the same direction as in Figure 5. Political affiliation is the only coefficient that has switched sides and can now be found considerably to the right of the intercept line. Whereas higher percentages of Republican vote predicted marginally lower dimension 1 scores, they predict quite substantially higher dimension 2 scores, although with a wide margin of uncertainty. I would suggest that population density and political preference need to be considered together in these models. Densely populated urban centres tend to be strongholds of the Democratic party, whereas other areas with relatively high, but not extreme, population density where the political affiliation is more towards Republican represent the smaller towns of East Texas. Figure 8 illustrates this point by showing the correlation between each dimension's scores and population density in the data.

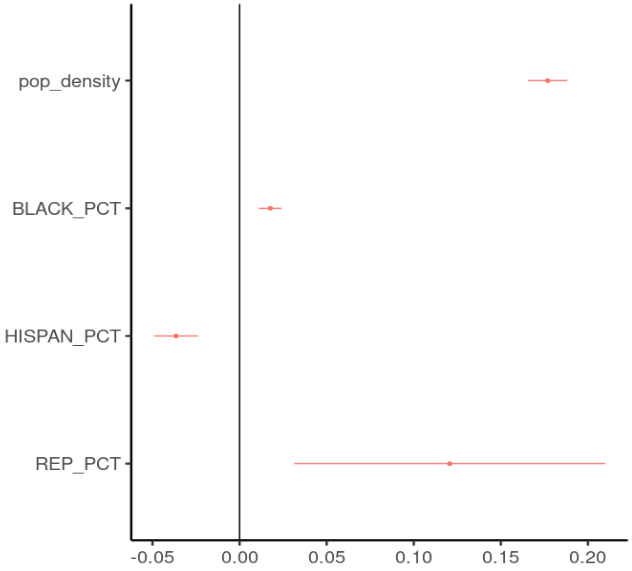

Figure 7: Coefficient estimates for a linear regression predicting Dimension 2 score.

As can be seen from Figure 8, both Dimension 1 and Dimension 2 show a general tendency towards higher scores with higher (log-transformed) population density. However, whereas the relationship is approximately linear for Dimension 1, this is not the case for Dimension 2. The latter peaks at logged population densities of around 11 and plateaus there. For particularly densely populated areas, there is even a slight dip in Dimension 2 scores. This, in combination with the patterns identified in Figure 6 and the coefficient for REP_PCT in Figure 7, supports the interpretation above, namely: that Dimension 1 is related to urban centres and Dimension 2 to the generally more densely populated eastern part of the state. Interestingly, then, the two dimensions together converge on the two most important aspects of diatopic variation identified for Texas in the dialectological literature.

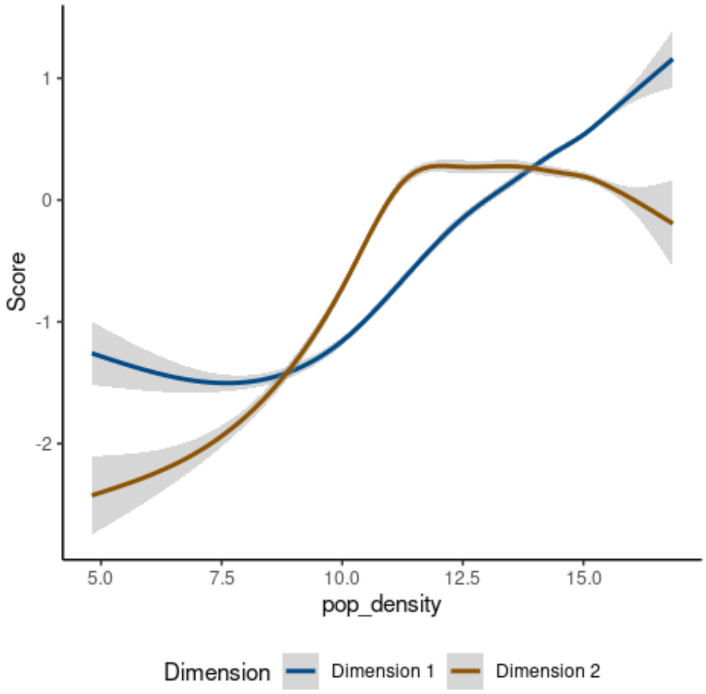

Figure 8: Correlation between population density and scores along Dimensions 1 and 2.

6. Discussion

The analysis presented above was introduced as a complement rather than a competitor to the distributional method used in Rosenfeld (2019). At this point, it is worth returning to the latter and relating the two methods to each other. It is obvious from a comparison of Figure 3 with Figures 4 and 6 that the clean representation of diatopic relations achieved in Rosenfeld (2019) is out of reach for the MDA-based approach taken here. There are at least two reasons for this fact. First, distributional methods can capitalise on a lot more information, namely: each word and its collocational behaviour in each sub-corpus. Second, and importantly, the smoothing procedure employed by Rosenfeld increases the similarity between voting precincts which are in close geographical proximity to each other. Therefore, the geographical signal is strengthened and made to appear more regular by default.

It would have been possible, in principle, to introduce such a smoothing procedure in the present study as well, either by updating the frequency information for each feature in each precinct with information from the surrounding precincts or by updating the calculated dimension scores. The reason this was not done lies in the purpose of the analysis itself, which focused on interpretable linguistic patterns rather than clear and plausible regional distinctions. The primary aim of an MDA, in other words, is not to construct maps as shown in Figures 4 and 6, but the kind of feature bundles shown in Tables 2 and 3. When taking these as a starting point, the interpretability of Dimensions 1 and 2 – creative, multimodal writing and an interpersonal, colloquial style, respectively – is obvious and can be taken as a testament to the method's success.

The question, perhaps, is not why Dimensions 1 and 2 show weaker geographic signals, but why they do show such signals at all. The relevant features are not generally considered in dialectological research, and they apply either to the context of CMD in particular or to very general situational variation. The traditional dialectological variants – often agricultural terms that are exceedingly rare in naturally occurring discourse – are not represented at a sufficient rate, even in a multi-million tweet corpus, to play any statistical role. Phonological variation, which is of course much more pervasive and has been the basis of more recent approaches to American dialectology (Labov et al. 2005), is not directly represented in the written medium of Twitter. While some approaches have been made in the past to combine the perspectives of register and (social) dialectology (Biber & Finegan 1994), the cross-fertilisation has been minimal. What the present study shows is that register variation indeed has a diatopic (and diastratic) component.

Having established the patterns of regional variation for Dimensions 1 and 2, the task remains to interpret them. Why is it that people in the urban centres draw decisively more actively on the full range of semiotic affordances on Twitter? One explanatory factor may be the age structure of the communities: younger users are expected to be found in these urban centres at higher rates than elsewhere, and these may simply be more accustomed to the modal affordances of emoji usage. It is unfortunate that data for each voting precinct's age structure is not available to further explore this hypothesis. Another line of interpretation may have to do with visibility as a valuable resource on Twitter (Squires 2015:245). In highly populated areas, more tweets are produced in general, and the competition for attention may foster the increasing use of multimodal and creative resources to enrich one's discourse.

This line of interpretation may be extended to Dimension 2 as well. As was argued above, it is the populated areas outside the urban centres in particular that show high scores along this dimension. Here, the competition for attention may be less fierce, and predominant modes of sociality less anonymous than in the city centres. These properties may contribute to a more interpersonal style. Put briefly, tweets from the cities may be situated more towards the "public" end of the cline between private and public that Twitter on the whole occupies (Bruns & Moe 2014) and share characteristics of advertising language, whereas tweets from the more rural parts of the state (particularly in the East) may be more private in conceptualisation, sharing properties of spoken conversation. One aspect the present study is not optimally suited to address is precinct-internal variation, especially in the urban areas. The reduction of each precinct's linguistic profile to two-dimension scores is not able to capture the internal heterogeneity of the discourse in each. This element of internal heterogeneity deserves further attention in future research.

Returning to Underwood's (1990:96) scepticism of "wordcarvers" and their attempts to draw precise linguistic boundaries across the state of Texas, it should be noted that any such line is indeed an arbitrary simplification. No actual border exists, obviously, which separates two different yet internally homogenous dialects. The task of the sociolinguist and the dialectologist is not to draw rigid borders but to identify patterns. In this regard, multiple sources of data can and should be considered alongside each other without the need to reduce all their complexity to simple oppositions. Twitter data have their place in this endeavour, not as a better, but as an additional source for studying diatopic and diastratic variation.

7. Conclusion

The present paper has addressed the potential of Twitter data to explore aspects of diatopic, as well as diastratic, variation in Texas English. This variation, it was argued, is currently not fully understood in large parts due to an absence of robust empirical data. Twitter discourse was introduced as one potential source of such data and its utility demonstrated with reference to Rosenfeld's (2019) study of regional variation in Texas English. The methodological contribution the present study sought to make was not a better dialect map of Texas, but the identification of linguistically interpretable dimensions of variation. To this purpose, a multi-dimensional analysis was run on Rosenfeld's data: over 2 million tweets, mapped to the Texas voting precincts.

The first two dimensions identified in this procedure proved interpretable in both linguistic and geographic terms. Dimension 1 comprised emojis and other creative uses of typography to enhance computer-mediated discourse and was found at particularly high scores in the urban centres of the state, such as Houston, Dallas-Fort Worth, San Antonio, and Austin. The second dimension showed clear characteristics of interpersonal, oral discourse and was found at higher rates in East Texas than West Texas. Regression models for both dimensions, with demographic information for the voting precincts as predictors, confirmed and refined these interpretations.

It was argued that MDA is a useful perspective to investigate meaningful diatopic variation, perhaps surprisingly so since the method itself is designed to identify rather general aspects of situational language use. The extent to which register and diatopic variation are interrelated remains a fruitful area of future research.

References

Atwood, Bagby. 1962. *The regional vocabulary of Texas*. Austin, TX: University of Texas Press.

Austen, Martha. 2017. Put the groceries up: Comparing Black and White regional variation. *American Speech* 92 (3), 298–320.

Bailey, Guy. 1991. Directions of change in Texas English. *Journal of American Culture* 14 (2), 125–134.

Baugh, Albert. 1935. *A history of the English language*, first edition. New York; London: D. Appleton-Century.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge; New York: Cambridge University Press.

Biber, Douglas and Edward Finegan (eds.). 1994. *Sociolinguistic perspectives on register*. New York; Oxford: Oxford University Press.

Bohmann, Axel. 2019. *Variation in English worldwide: Registers and global varieties* (Studies in English Language). Cambridge: Cambridge University Press.

Bruns, Axel and Hallvard Moe. 2014. Structural layers of communication on Twitter. *Twitter and society* (Digital Formations vol. 89), ed. by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, 15–28. New York: Peter Lang.

Cattell, Raymond. 1966. The scree test for the number of factors. *Multivariate Behavioral Research* 1 (2), 245–276.

Clarke, Isobelle and Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PloS One* 14(9).

Coseriu, Eugenio. 1955. *La geografía lingüística*. Montevideo: Universidad de la República, Facultad de Humanidades y Ciencias.

Cukor-Avila, Patricia, Lisa Jeon, Patricia Rector, Chetan Tiwari, and Zak Shelton. 2012. "Texas – It's like a whole nuther country": Mapping Texans' perceptions of dialect variation in the Lone Star State. *Texas Linguistics Forum* 55, 10–19.

Eisenstein, Jacob. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics* 19 (2), 161–188.

Eisenstein, Jacob. 2018. Identifying regional dialects in on-line social media. *The Handbook of Dialectology*, ed. by Charles Boberg, John Nerbonne, and Dominic Watt, 368–383. Hoboken, N.J.: Wiley-Blackwell.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (EMNLP '10), 1277–1287. Stroudsburg, PA, USA: Association for Computational Linguistics.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(11).

Gonçalves, Bruno, Lucía Loureiro-Porto, José Ramasco, and David Sánchez. 2018. Mapping the Americanization of English in space and time. *PLoS ONE* 13(5): e0197741.

Gorsuch, Richard L. 2015. *Factor analysis*. Classic edition. New York; London: Routledge.

Grieve, Jack. 2016. *Regional variation in written American English* (Studies in English Language). Cambridge: Cambridge University Press.

Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2.

Hinrichs, Lars, Axel Bohmann, and Kyle Gorman. 2013. Real-time trends in the Texas English vowel system: F2 trajectory in GOOSE as an index of a variety's ongoing delocalization. *Rice Working Papers in Linguistics* 4.

Hovy, Dirk and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4383–4394. Brussels, Belgium: Association for Computational Linguistics.

Iorio, Josh. 2009. Effects of audience on orthographic variation. *Studies in the Linguistic Sciences: Illinois Working Papers,* 127–140.

Jones, Taylor. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter." *American Speech* 90 (4), 403–440.

Koch, Peter and Wulff Oesterreicher. 1985. Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit um Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36 (85), 15–43.

Kretzschmar, William A. 2015. *Language and complex systems*. Cambridge: Cambridge University Press.

Kurath, Hans. 1949. *Word geography of the Eastern United States*. Ann Arbor, MI: University of Michigan Press.

Labov, William, Sharon Ash, and Charles Boberg. 2005. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin; New York: Mouton de Gruyter.

Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv:1405.4053 [cs]*. http://arxiv.org/abs/1405.4053.

Louf, Thomas, Bruno Gonçalves, José Ramasco, David Sánchez, and Jack Grieve. 2023. American cultural regions mapped through the lexical analysis of social media. *Humanities and Social Sciences Communications* 10, Article 133.

R Core Team. 2022. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosenfeld, Alex B. 2019. *Computational models of changes in language use*. Austin, TX: The University of Texas at Austin doctoral dissertation.

Russ, Brice. 2012. Examining large-scale regional variation through online geotagged corpora. Conference presentation at the 2012 ADS Meeting. Portland, OR, 06 January 2012.

Solt, Frederick and Yue Hu. 2018. dotwhisker: Dot-and-whisker plots of regression results. https://CRAN.R-project.org/package=dotwhisker.

Squires, Lauren. 2015. Twitter: Design, discourse, and the implications for public text. *The Routledge Handbook of Language and Digital Communication,* ed. by Alexandra Georgakopoulou; Tereza Spilioti, 239–256. London: Routledge.

Stevenson, Jonathan. 2016. *Dialect in digitally mediated written interaction: a survey of the geohistorical distribution of the ditransitive in British English using Twitter*. York, UK: University of York MA dissertation.

Strelluf, Christopher. 2020. *Needs* + PAST PARTICIPLE in regional Englishes on Twitter. *World Englishes* 39 (1), 119–134.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman. 2012. Geography of Twitter networks. *Social Networks* 34 (1), 73–81.

Thompson, Bruce. 2004. *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Underwood, Gary. 1990. Scholarly responsibility and the representation of dialects: The case of English in Texas. *Journal of English Linguistics*. 23 (1–2), 95–113.

Van Halteren, Hans. 2021. Pitfalls in tweet-based variation studies. Conference presentation at NWAV 49, The University of Texas at Austin. Austin, TX, 22 October 2021.

Ward, Joe H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58 (301), 236–244.

Yaqub M. 2022. How many tweets per day 2022 (number of tweets per day). Renolon. *Smart Insights*. https://www.renolon.com/number-of-tweets-per-day/. (12 November, 2022).