Rosenfeld, Alex, and Lars Hinrichs. 2023. Capturing Fine-Grained Regional Differences in Language Use through Voting Precinct Embeddings. *Computational Linguistics*, uncorrected proof.

Capturing Fine-Grained Regional Differences in Language Use through Voting Precinct Embeddings

Alex Rosenfeld[*]
Leidos
Innovations Center
alexbrosenfeld@gmail.com

Lars Hinrichs
The University of Texas at Austin
Department of English
TxE@utexas.edu

Linguistic variation across a region of interest can be captured by partitioning the region into areas and using social media data to train embeddings that represent language use in those areas. Recent work has focused on larger areas, such as cities or counties, to ensure that enough social media data is available in each area, but larger areas have a limited ability to find fine-grained distinctions, such as intracity differences in language use. We demonstrate that it is possible to embed smaller areas, which can provide higher resolution analyses of language variation. We embed voting precincts, which are tiny, evenly sized political divisions for the administration of elections. The issue with modeling language use in small areas is that the data becomes incredibly sparse, with many areas having scant social media data. We propose a novel embedding approach that alternates training with smoothing, which mitigates these sparsity issues. We focus on linguistic variation across Texas as it is relatively understudied. We developed two novel quantitative evaluations that measure how well the embeddings can be used to capture linguistic variation. The first evaluation measures how well a model can map a dialect given terms specific to that dialect. The second evaluation measures how well a model can map preference of lexical variants. These evaluations show how embedding models could be used directly by sociolinguists and measure how much sociolinguistic information is contained within the embeddings. We complement this second evaluation with a methodology for using embeddings as a kind of genetic code where we identify "genes" that correspond to a sociological variable and connect those "genes" to a linguistic phenomenon thereby connecting sociological phenomena to linguistic ones. Finally, we explore approaches for inferring isoglosses using embeddings.

[*] Research performed while attending The University of Texas at Austin.

Action Editor: Ekaterina Shutova. Submission received: 24 October 2022; revised version received: 28 March 2023; accepted for publication: 20 May 2023.

https://doi.org/10.1162/coli_a_00487

1. Introduction

Similar to embeddings that capture word usage, recent work in NLP has developed methods that generate embeddings for areas that represent language in those areas. For example, Huang et al. (2016) developed an embedding method for capturing language use in counties and Hovy and Purschke (2018) developed an embedding method for capturing language use in cities. These embeddings can be used for a wide variety of sociolinguistic analyses as well as downstream tasks.

Given the sheer volume available, social media data is often used to provide the text data needed to train the embeddings. However, one inherent problem that arises is the imbalance of population distribution across a region of interest, which leads to an imbalance of social media data across that region. For example, rural areas use Twitter less than urban areas (Duggan 2015). This could make it more difficult to capture language use in rural areas.

One solution to this issue is to use larger areas. For example, one could focus on cities and not explore the countryside, such as done in Hovy and Purschke (2018). Or one could divide a region of interest into large squares, such as done in Hovy et al. (2020). Or one could divide a region of interest into counties, such as done in Huang et al. (2016). While these solutions produce areas with more data, the areas themselves could be less useful for analysis as (1) there could be important areas that are not covered (e.g., only studying cities and missing the rest of the region), (2) the areas could have awkward boundaries (e.g., dividing regions into squares that ignore geopolitical boundaries), or (3) the resolution would be too low to be useful for certain analyses (e.g., using cities as areas prevents analyses of intracity language use).

We propose a novel solution to the data problem. We use smaller areas, voting precincts, that provide finer resolution analyses and propose a novel embedding approach to mitigate the specific data issues related to using smaller areas. Voting precincts are small, equally sized areas that are used in the administration of elections (in Texas, each voting precinct has about 1,100 voters). As they are well regulated (voting precincts are required to fit within county, congressional boundaries), monitored (voting precincts are a fundamental unit in censuses), compact (voting precincts need to be compact to make elections, polling, and governance more efficient), and cover an entire region, they form a perfect mesh to represent language use across a region. Unlike with using cities, voting precincts can also capture rural areas. Unlike with using squares, voting precincts follow geopolitical boundaries. Unlike with counties, voting precincts can better capture intracity differences in language use. Thus, by developing embedding representations of these precincts, we can find fine-grained differences in language use across a large region of interest.

While voting precincts are a great mesh to model language use across a region, the smaller sizes lead to significant data issues. For example, less populated areas use social media less, which can lead to voting precincts that have extremely limited data or no data at all. To counteract this, we propose a novel embedding technique where training and smoothing alternate to mitigate the weaknesses of both. Training has limited potential in voting precincts with little data, so smoothing will provide extra information to create a more accurate embedding. Smoothing can spread noise, so training afterwards can refine the embeddings.

We propose novel evaluations that explore how well embeddings can be used to predict information useful to sociolinguists. The first evaluation explores how well embeddings can be used to predict where a dialect is spoken using some specific features of the dialect. We use the Dictionary of American Regional English dataset

(DAREDS) (Rahimi, Cohn, and Baldwin 2017), which provides key terms for various American dialects. We evaluate how well embeddings can be used to predict dialect areas from those key terms.

The second evaluation explores how well embeddings can be used to predict lexical variation. Lexical variation is the choice between two semantically similar lexical items, for example, *fam* versus *family*, and is a good determiner of linguistic variation (Cassidy, Hall, and Von Schneidemesser 1985; Carver 1987). We evaluate how well embeddings can be used to predict choice in lexical variant across a region of interest.

As part of these evaluations, we perform a hyperparameter analysis that demonstrates that post-training retrofitting can have numerical issues when applied to smaller areas, so alternating is a necessary step with smaller areas. As mentioned, many smaller areas lack sufficient data, so retrofitting with these areas can cause the spreading of noise, which in turn can result in unreliable embeddings.

We then provide a novel methodology to extract novel sociolinguistic insights from social media data. Area embeddings capture language use in an area, and language use is connected to a wide swath of sociological factors. If we treat embeddings as the "genetic code" of an area, we can identify sections of the embeddings that act as genes for sociological phenomena. For example, we can find the "gene" that encodes how race and the urban–rural divide affect language use. Then by exploring the predictions of these "genes" we can then connect the sociological phenomenon with a linguistic one, for example, identify novel African American slang via analyzing the expressions of the "gene" corresponding to Black Percentage.

Finally, we use our embeddings to predict geographic boundaries of linguistic variation, or "isoglosses". Prior work has used principal component analysis to infer isoglosses, but with smaller areas, we find that PCA will focus on the urban–rural divide and ignore regional divides. Instead, we find that t-distributed stochastic neighbor embedding (Van der Maaten and Hinton 2008) is better able to identify larger geographic distinctions.

2. Prior Work

While there has been a wealth of work that has used Twitter data to explore lexical variation (e.g., Eisenstein et al. 2012, 2014; Cook, Han, and Baldwin 2014; Doyle 2014; Jones 2015; Huang et al. 2016; Kulkarni, Perozzi, and Skiena 2016; Grieve, Nini, and Guo 2018), the incorporation of distributional methods is a more recent trend.

Huang et al. (2016) apply a count-based method to Twitter data to represent language use in counties across the United States. They use a manually created list of sociolinguistically relevant variant pairs, such as *couch* and *sofa*, from Grieve, Asnaghi, and Ruette (2013) and embedded a county based on the proportion of each variant. They then used adaptive kernel smoothing to smooth the counts and used PCA for dimensionality reduction. They do not perform a quantitative evaluation and instead perform PCA of the embeddings. One limitation of their approach is that it requires a list of sociolinguistically relevant variant pairs. Producing such pairs is labor-intensive and such pairs are specific to certain language varieties (variant pairs that make sense for American English may not make sense for British English) and may lose relevance as language use changes over time.

Hovy and Purschke (2018) use document embedding techniques to represent language use in cities in Germany, Austria, and Switzerland. In this work, they collected social media data from Jodel,[1] a social media platform, and used Doc2Vec (Le and Mikolov 2014) to produce an embedding for each city. As their goal was to explore regional variation, they used retrofitting (Faruqui et al. 2015; Hovy and Fornaciari 2018) to have the embeddings better match the NUTS2 regional break down of those countries. We discuss these methods further in Section 4. For quantitative evaluation, they compare clusterings of their embeddings to a German dialect map (Lameli 2013). While this an excellent evaluation if you have such a map, the constantly evolving nature of language and the sheer difficulty of hand-creating such a dialect map make this approach difficult to generalize to analyses of new regions, especially a region as evolving and large as the state of Texas, which is our focus. The authors also evaluated their embeddings by measuring how well they could predict the geolocation of the Tweet. While geolocation is a laudable goal in and of itself, our focus is on linguistic variation specifically and geolocation is not necessarily a measure of how well the embeddings capture linguistic variation. For example, a list of business names in each area would be fantastic for geolocation, but of less use for analyzing variation.

Hovy et al. (2020) followed up this work by extending their method to cover entire continents/countries and not just the cities. They did this by dividing their region of interest into a coordinate grid of 11 km (6.8 mi.) by 11 km squares and training embeddings for each square. They then retrofitted the square embeddings. They did not perform a quantitative evaluation of their work.

An alternative approach to generating regional embeddings is through using linguistic features as the embedding coordinates. For example, Bohmann (2020) embedded Twitter linguistic registers into a space based on 236 linguistic features. They then use factor analysis on these embeddings to generate 10 dimensions of linguistic variation. While these kinds of embeddings are more interpretable, they require more a priori knowledge about relevant linguistic features and the capability to calculate them. While we do not explore linguistic feature–based embeddings in our work, we do perform a similar task in extracting smaller dimensional representations when analyzing theoretic linguistic hypotheses.

Clustering is a well-explored topic in computational dialectology (e.g., Grieve, Speelman, and Geeraerts 2011; Pröll 2013; Lameli 2013; Huang et al. 2016). To this effect, we largely follow the clustering approach in Hovy and Purschke (2018). We also explore this topic while incorporating newer clustering techniques, such as t-SNE (Van der Maaten and Hinton 2008). Like Hovy et al. (2020), we do not do hard clustering (like *k*-means) and only do soft clustering.

There has been work that has analyzed non-conventional spellings (Liu et al. 2011 and Han and Baldwin 2011, for example), but recent work has explored the use of word embeddings to study lexical variation through non-conventional spelling (Nguyen and Grieve 2020). In that work, the authors explored the connection between conventional and non-conventional forms and found that word embeddings do capture spelling variation (despite being ignorant of orthography in general) and discovered a link between the intent of the different spelling and the distance between the embeddings. While we do not directly interact with this work, their exploration of the connection between non-conventional spelling and lexical variation may be useful for future work.

There is a wealth of work that uses computational linguistic methods to connect sociological factors with word use (See Nguyen et al. [2016] for a review of work in this area as well as computational sociolinguistics in general). One such approach is

[1] https://jodel.com/.

that from Eisenstein, Smith, and Xing (2011), which uses a regression model to connect word use with demographic features. By using a regularization method to focus on key words, they show which words are connected to specific sociological factors. While we don't connect word A with demographic B, we use a similar technique to extract sections of embeddings that are related to specific demographic differences.

3. Texas Twitter and Precinct Data Collection

Our focus is on language use across the state of Texas. It is large, populous, and has been researched only lightly in sociolinguistics and dialect geography, compared with other large American states. Both Thomas and Bailey have contributed quantitative studies of variation in Mainstream (not ethnically specific) Texas English: Thomas (1997) describes a rural/urban split in Texas dialects, driven by the much-accelerated migration of non-southerners into Texas and other southern U.S. states since the latter decades of the twentieth century, a trend that effectively creates "dialect islands in Texas where the large metropolitan centers lie" (Thomas 1997, page 309) and relegating canonical features of southern U.S. speech (Thomas's focus is on the monophthongization of PRICE and the lowering of the nucleus in FACE vowels) to rural areas and small towns. Bailey et al. (1991), by tracking nine different features of phonetic innovation/conservativeness in Texas English and resolving findings at the level of the county, identify the most

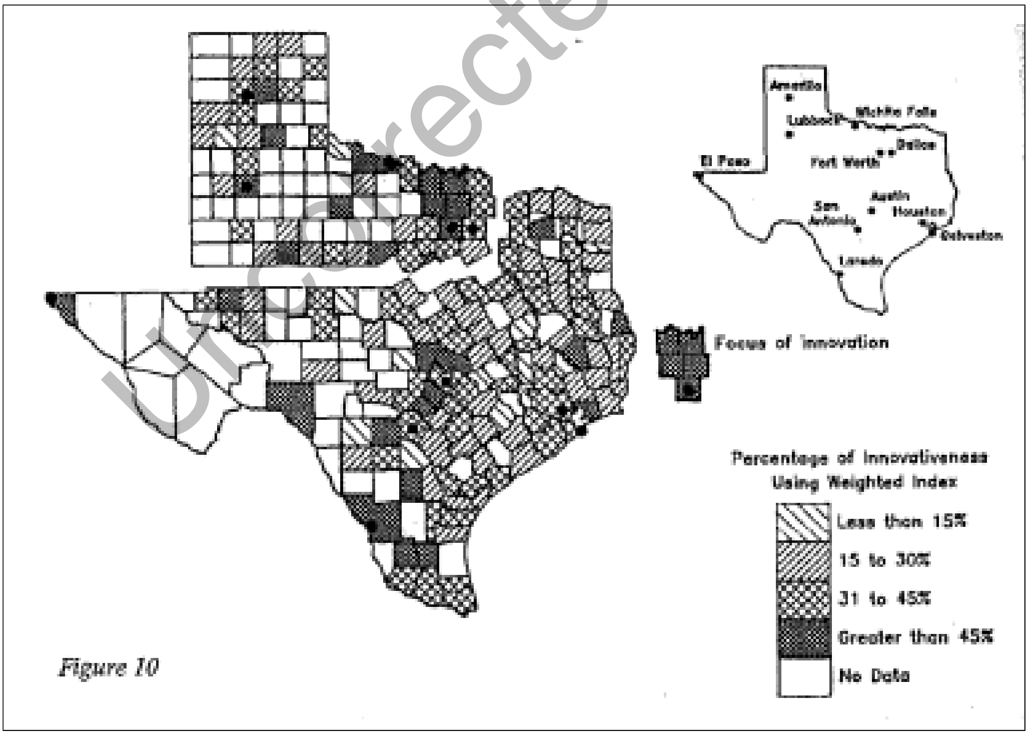


Figure 1
Weightedindex for innovative forms, aggregated at the county level. (Reprinted from Bailey, Wikle, and Sand 1991, withpermission of Johns Benjamin Publishing Co.). Q1 Q2

linguistically innovative areas driving change in Texas English as a cluster of five counties in the Dallas/Fort Worth area.

In addition to these geographic approaches to variation in Texas, there have been a number of studies focusing on selected features (Bailey and Dyer 1992; Atwood 1962; Bailey et al. 1991; Bernstein 1993; Di Paolo 1989; Hinrichs, Bohmann, and Gorman 2013; Koops 2010; Koops, Gentry, and Pantos 2008; Walsh and Mote 1974; Tarpley 1970; Wheatley and Stanley 1959) and/or variation and change in minority varieties (Bailey and Maynor 1989, 1987, 1985; Bayley 1994; Galindo 1988; Garcia 1976; Bailey and Thomas 2021; McDowell and McRae 1972).

Outside of computational sociolinguistics, attempts to geographically model linguistic variation in Texas English have been made as part of the established, large initiatives in American dialect mapping. These include:

- Kurath's linguistic atlas project (LAP; see Petyt [1980] for an overview) that produced the Linguistic Atlas of the Gulf States (Pederson 1986), based on survey data;
- Carver's (1987) "word geography" atlas of American English dialects, which visualizes data from the Dictionary of American Regional English (Cassidy, Hall, and Von Schneidemesser 1985) on the geographic distribution of lexical items; and
- the Atlas of North American English (Labov et al. 2006), which maps phonetic variation in phone interview data from speakers of of American English.

3.1 Data Collection

In this section, we will describe how we collected Texas Twitter data for our analysis. Twitter data has allowed sociolinguists new ways to explore how society affects language (Mencarini 2018). This data is composed of a large selection of natural uses of language that cut across many social boundaries. Additionally, tweets are often geotagged, which allows researchers to connect examples of language use with location.

We draw our Twitter data from two sources. The first is from archive.org's collection of billions of tweets (Archive Team 1996–) that were retrieved between 2011 and 2017. This collection represents tweets from all over the world and not Texas specifically. The second source is a collection of 13.6 million tweets that were retrieved using the Twitter API between February 16, 2017, and May 3, 2017. We only retrieved tweets that originate in a rectangular bounding box that contains Texas.

Our preprocessing steps are as follows. First, we remove all tweets that do not have coordinate information nor a city name in its metadata. Any tweet that does not have coordinate information, but a city name, we use the simplemaps.org United States city database[2] to give these tweets coordinates based upon its city's coordinates. We then remove tweets that were not sent from Texas. We then remove all tweets that have a hashtag (#) to help remove automatically generated tweets, like highway accident reports. We then use the ekphrasis Python module to normalize the tweets

[2] https://simplemaps.com/data/us-cities.

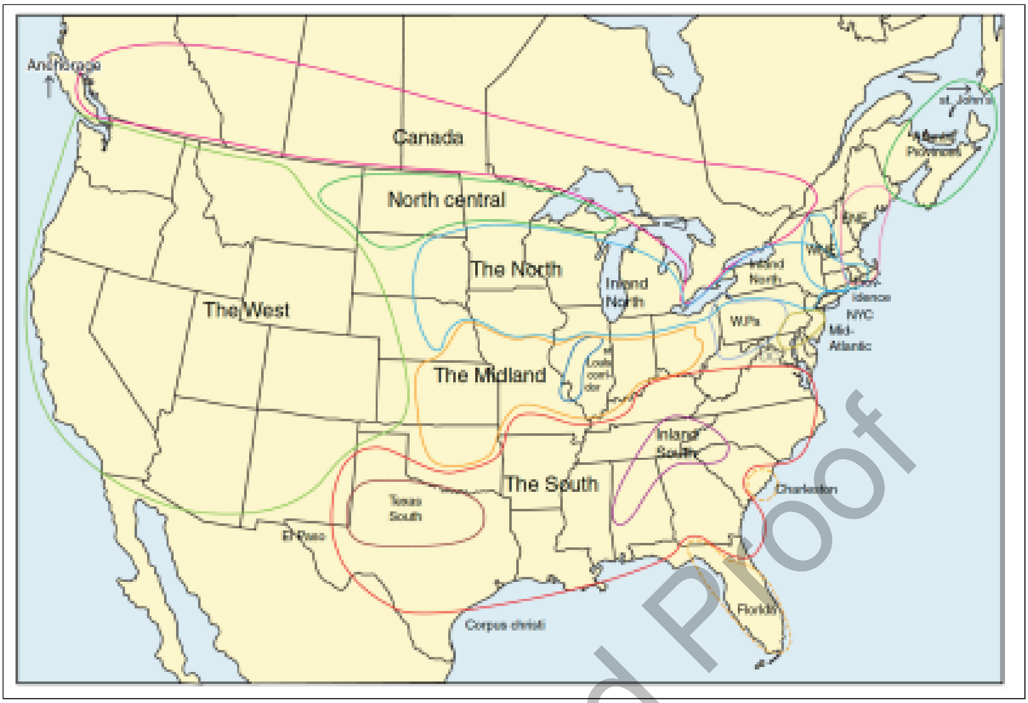


Figure 2
Major dialects of North American English. (Reprinted from Labov et al. 2006, p 148, by permission.)

(Baziotis, Pelekis, and Doulkeridis 2017). We do not remove mentions or replace them with a named entity label. Together, this results in 2.3 million tweets (1.7 million from archive.org and 563 thousand from the Twitter API).

In Figure 3, we visualize number of tweets in each voting precinct (left) and the voting precincts that have 10 or fewer tweets (right). We see that quite a few voting precincts have 10 or fewer tweets, especially rural and West Texas. This indicates that

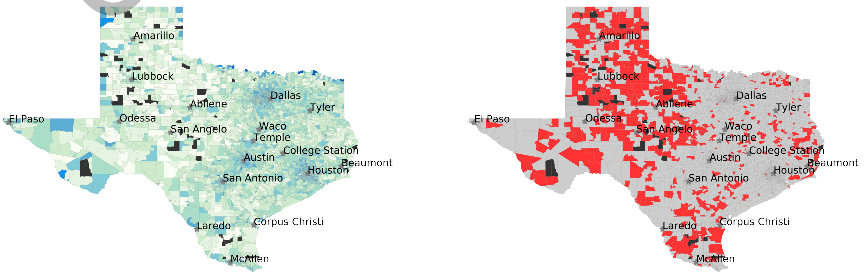


Figure 3
The left image visualizes the number of tweets per voting precinct. The right image shows which voting precincts have 10 or fewer tweets (red) or no tweets (black).

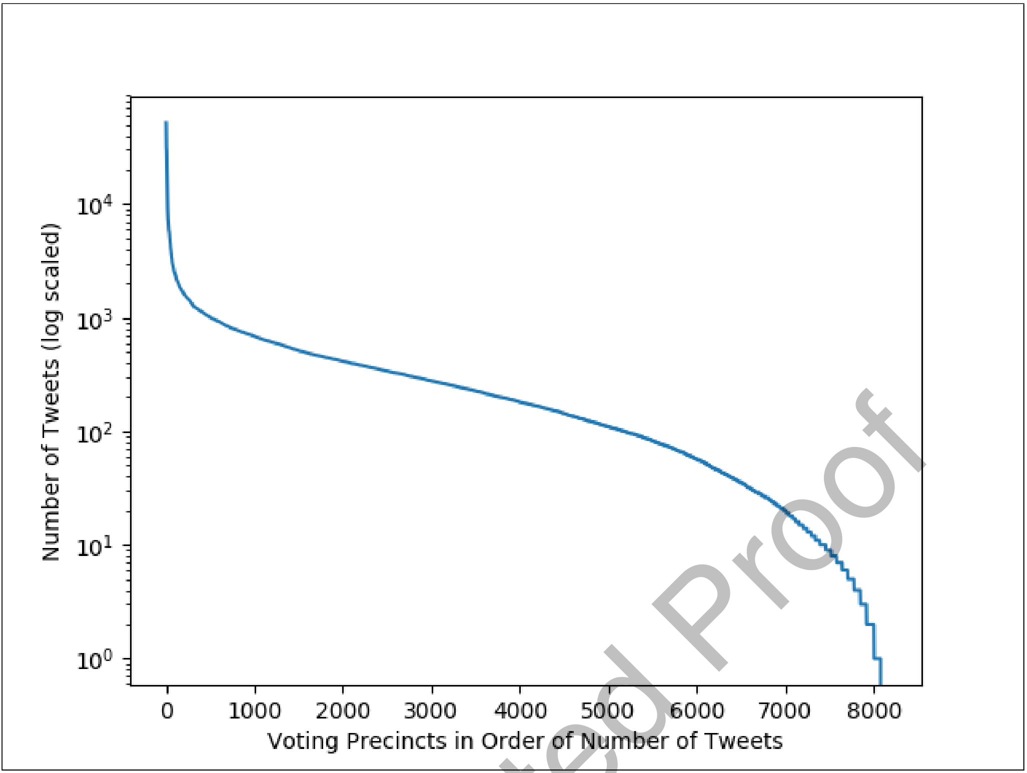


Figure 4
Distribution of tweets among voting precincts.

many precincts do not have enough tweets to generate accurate representations on their own and thus require some from of smoothing. In Figure 4, we show how the tweets are distributed across voting precincts. The voting precincts are ranked by number of tweets. We see that there is a few that have a vast amount of tweets, but most voting precincts have a number of tweets in the hundreds.

3.2 Voting Precincts

Our goal is to represent language use across the entirety of Texas (including rural Texas) as well as capture fine-grained differences in language use (including within a city). In prior work, researchers either only used cities (e.g., Hovy and Purschke 2018), or used a coordinate grid (e.g., Hovy et al. 2020). The former does not explore rural areas at all and does not explore within-city divisions. The latter uses boundaries that do not reflect the geography of the area and are difficult to use for fine-grained analyses.

To achieve our goals, we operate at the voting precinct level. Voting precincts are relatively tiny political divisions that are used for the efficient administration of elections. Each voting precinct usually has one polling place and, in the 2016 election, each voting precinct contained on average 1,547 registered voters nationwide (U.S. Election Assistance Commission 2017). These voting precincts are generally relatively tiny (on average containing 3,083 people), cohesive (each voting precinct must reside entirely within an electoral district/county), and balanced (generally, voting precincts

Table 1
Population Demographics of the 8,148 voting precincts in Texas.

Variable	Pop/Area Per VP	Demo % of VP
Land Area	76.08km^2 (± 18.55km^2)	
Population	3083.0 (± 2601.2)	100.0% (± 0.0%)
Asian	116.2 (± 309.1)	2.60% (± 5.48%)
Black	354.1 (± 681.6)	10.6% (± 16.8%)
Hispanic	1160.5 (± 1677.5)	33.7% (± 27.6%)
Multiple	39.1 (± 50.9)	1.15% (± 0.90%)
Native American	9.8 (± 12.9)	0.36% (± 1.09%)
Other	4.1 (± 7.6)	0.11% (± 0.22%)
Pacific Islander	2.1 (± 10.7)	0.06% (± 0.66%)
White	1396.8 (± 1384.4)	51.3% (± 29.4%)

are designed to contain similar population sizes). Additionally, states record meticulous detail on the demographics of each voting precinct (See Table 1 for descriptive statistics). Thus, these voting precincts act as perfect building blocks.[3]

We note that gerrymandering has very little influence on voting precinct boundaries. It is true that congressional districts (and similar) can be heavily gerrymandered and voting precincts are bound by congressional district boundaries. However, the practical pressures of administration and the relatively small size of the voting precincts minimize these effects. Voting precincts are used to administer elections, which means that significant effort is needed to coordinate people to run polling stations and identify locations where people can vote. Additionally, voting precincts are often used to organize polling and signature collection. Due to these factors, there is a strong need for all parties involved to make voting precincts as compact and efficient as possible. In contrast, voting precinct boundaries only decide where you vote and not who you vote for, so there is not the pressure to gerrymander in the first place. Voting precincts are also generally small enough to fit into the nooks and crannies of congressional districts. Congressional districts have dozens of voting precincts, so voting precincts are small enough to be compact despite any boundary issues of the larger congressional district. It is for these reasons that voting precincts are often used as atomic units in redictricting efforts (e.g., Baas n.d.).

The voting precinct information comes from the United States Census and is compiled by the Auto-Redistrict project (Baas n.d.). Each precinct in this data comes with the coordinate bounds of the precinct along with the census demographic data. Further processing of the demographic data was done by Murray and Tengelsen (2018).

In order to map tweets to voting precincts, we first extract a representative point for each voting precinct using the Shapely Python module (Gillies et al. 2007). Representative points are computationally efficient approximations to the center of a voting precinct. We then associate a Tweet to the closest voting precinct by distance from the Tweet's coordinates to the representative points.

[3] While voting precincts were a better fit for our needs, similar analyses could be done with Census tracts, Census block groups, or any fine-grained sectioning of a region.

4. Voting Precinct Embedding Methods

In this section, we describe the area embedding methods we will analyze. Area embedding methods generally have two parts: a training part and a smoothing part. The training part takes text and uses a machine learning or counting based model to produce embeddings. The smoothing part averages area embeddings with their neighbors to add extra information.

4.1 Count-Based Methods

The first approach we explore is a count-based approach from Huang et al. (2016). The training part counts the relative frequencies of a manually curated list of sociolinguistically relevant lexical variations. The smoothing part takes a weighted average of the area embedding and enough nearest neighbors to meet some data threshold.

4.1.1 Training: Mean-Variant-Preference. Grieve, Asnaghi, and Ruette (2013) and Grieve and Asnaghi (2013) have manually collected sets of lexical variants where the choice of variant is indicative of local language use. For example, *soda*, *pop*, and *Coke* are a set of lexical variants for "soft drink" and regions have a variant preference. Huang et al. (2016) count the relative frequency of variants and use these counts as the embedding.

More specifically, they begin with a manually curated list of sociolinguically-relevant sets of lexical variants. They designate the most frequent variant as the "main" variant. In the soft drink example, *soda* would be the main variant as it is the most frequent variant among all variants.

Given an area and a set of lexical variants, Huang et al. (2016) take the relative frequency of the "main" variant across Twitter users in the area:

$$MVP(area, variants) = \frac{1}{U(area)} \sum_{\text{users } u \text{ in the area}} \frac{\text{times user } u \text{ used main variant}}{\text{times user } u \text{ used any variant}}$$

where $U(area)$ is the number of Twitter users in that area. The embedding for an area would be each MVP value for set of variants in the list of sets of variants.

As baseline in our analysis, we just use the relative frequency over all tweets:

$$MVP(area, variants) = \frac{\text{total times main variant was used in the area}}{\text{times times any variant was used}}$$

Huang et al. (2016) derived their list of sets of variants from those in Grieve, Asnaghi, and Ruette (2013). They then filter this list by removing any sets that appear in less than 1,000 areas or that have a p-value less than 0.001 according to Moran's I test (Moran 1950).

For our count based model, we use the publicly available list of 152 sets in Grieve and Asnaghi (2013). We similarly use Moran's I to filter by p-value and remove any sets that appear in less than 1000 voting precincts. The original list of pairs and our final list can be found in Table A1.

4.1.2 Smoothing: Adaptive Kernel Smoothing. One issue with working with area embeddings is that there is an uneven distribution of tweets and many areas can lack Tweet data. Huang et al. (2016) do smoothing by creating neighborhoods that had enough data then taking a weighted average of the embeddings in the neighborhood.

For an area A, a neighborhood is the smallest set of geographically closest areas to A that have data above a certain threshold. For a set of lexical variants, this is some multiple B times the average frequency of those variants across all areas. For *soda*, *pop*, and *Coke*, this would be B times the average number of times someone used any of those variants. Huang et al. (2016) explore B values of 1, 10, and 100.

Huang et al. (2016) then use adaptive kernel smoothing (AKS) with a Gaussian kernel to get a weighted average of all embeddings in a neighborhood. The weight of a neighbor embedding is e to the negative distance between the area and the neighbor. The new area embedding is calculated as follows:

$$\overrightarrow{area} \leftarrow \frac{\sum_{N(area,\ B,\ altpair)} e^{-dist(area,\ neighbor)} \overrightarrow{neighbor}}{\sum_{N(area,\ B)} e^{-dist(area,\ neighbor)}}$$

where $N(area, B, variants) =$ the neighborhood around *area* such that the total usage of the pair is at least B times the average. Huang et al. (2016) after this smoothing process use PCA to reduce the dimension of the embeddings to 15.

As we will also explore more traditional embedding models, such as Doc2Vec, we adapt this smoothing approach for unsupervised machine learning models. Instead of average counts of variants, we use average number of tweets. In that way, each neighborhood will have a sufficient number of tweets to mitigate the data sparsity issue.

4.2 Post-training Retrofitting

The approach Hovy and Purschke (2018) and Hovy et al. (2020) took in their analysis is one where embeddings are first trained on social media data then altered such that adjacent areas have more similar embeddings. The first step uses Doc2Vec (Le and Mikolov 2014), while the second step uses retrofitting (Faruqui et al. 2015).

4.2.1 Training: Doc2Vec. The first part in their approach is to train a Doc2Vec model (Le and Mikolov 2014) for 10 epochs to obtain an embedding for each German-speaking city (Hovy and Purschke 2018) or coordinate square (Hovy et al. 2020). Doc2Vec is an extension of word2vec (Mikolov et al. 2013) that also trains embeddings for document labels (or in this case, the city/square/voting precinct where the post was written).

In Doc2Vec, words, contexts, and document labels are represented by embeddings and these embeddings are modeled through the following distribution:

$$P(word|context, documentlabel) = softmax(word \cdot (context + label))$$

By maximizing the likelihood of this probability relative to a dataset, the model will fit the word, context, and document label embeddings so that the above distribution best reflects the statistics of the data.

Doc2vec provides a vector $\overrightarrow{doc}$ for each document label *doc* (similarly with voting precincts and cities). The loss function is similar to word2vec as follows:

$$loss = \sum_{(w,c,d)\in D} \log(\sigma((\vec{w} + \vec{d}) \cdot \vec{c})) + \sum_{c' \sim P_D} \log(1 - \sigma((\vec{w} + \vec{d}) \cdot \vec{c'}))$$

where D is the collection of target word–context word–document label triples extracted from a corpus and P_D is the unigram distribution. We use the gensim implementation of Doc2Vec (Řehůřek and Sojka 2010).

The result of this process is that we have an embedding for each voting precinct (in our case) or coordinate square/German-speaking city (in Hovy and Purschke's case).

4.2.2 Smoothing: Retrofitting. One key insight from Hovy and Purschke (2018) is that Doc2Vec alone can produce embeddings that capture language use in an area, but not in a way that captures regional variation as opposed to city specific artifacts. For example, an embedding for the city of Austin, Texas, might capture all of the language use surrounding specific bus lines in the Austin Public Transportation system, but that information is less useful for understanding differences in language use across Texas.

The solution, proposed by Hovy and Purschke, is to use retrofitting to modify the embeddings so that that they better reflect regional information. Retrofitting (Faruqui et al. 2015) is an approach where embeddings are modified so that they better fit a lexical ontology. In Hovy and Purschke's case, their "ontology" is a regional categorization of German cities or, for their later paper, the adjacency relationship between coordinate squares. An embedding is averaged with the mean of its adjacent neighbors to smooth out any data-deficiency issues. This averaging is repeated 50 times to enhance the smoothing. This process is reflected in the following formula:

$$\overrightarrow{area} \leftarrow ½\, \overrightarrow{area} + ½ \frac{1}{\text{number of adjacent neighbors}} \sum_{neighbor \text{ of } area} \overrightarrow{neighbor}$$

4.3 Proposed Models

Given that our divisions are much smaller than those in previous work, we propose several area embedding methods that may perform better under our circumstances.

4.3.1 Geography Only Embedding. In this section, we describe a novel baseline that reflects embeddings that effectively only contain geographic information and no Twitter data, which we call Geography Only Embedding. In this approach, embeddings are randomly generated (we use a Doc2Vec model that is initialized, but not trained) and then retrofit the embeddings using the same process above.

Despite its simple description, this approach can be seen as one where embeddings capture solely geographic information. To see this, note that the randomization process provides each precinct its own completely random embedding. In effect, the embedding acts as a kind of unique identifier for the precinct as it is incredibly unlikely for two 300 dimensional random vectors to be similar. By retrofitting (i.e., averaging these unique identifiers precincts), you form unique identifiers for larger subregions. Thus, each precinct and each area has an embedding that directly reflects where it is located on the map. In this way, these embeddings capture the geographic properties, while simultaneously containing no Twitter information.

4.4 Smoothing: Alternating

One issue with the Post-training Retrofitting approach in our setting is that it relies on a large body of tweets per area. In our case, the voting precincts are too small. Despite having 2.3 million tweets, each voting district only contains about 400 tweets on average and hundreds of precincts have fewer than 10 tweets. Thus, the initial Doc2Vec step would lack sufficient data to create quality embeddings. The retrofitting step would then just be propagating noise.

In order to alleviate this issue, we propose to alternate the Doc2Vec and retrofitting steps to mitigate the weaknesses of both. In our setting, training injects Tweet information into the embeddings, but voting precincts often lack enough data to be used on its own. In contrast, retrofitting can send information from adjacent neighbors to improve an embedding, but can also overwhelm the embedding with noise or irrelevant information, for example, the Austin embedding (a major metropolis) could overwhelm the Round Rock embedding (a suburb of Austin) even though language use is different between those areas. If we train after retrofitting, we can correct any wrong information from the adjacent neighbors. If we retrofit after training, we can provide information where its lacking. Thus, alternating these steps can mitigate each step’s weakness.

4.5 Training: BERT with Label Embedding Fusion

Since the prior work, there have been advances in document embedding approaches, such as those that use contextual embeddings. We explore BERT with Label Embedding Fusion (BERTLEF) (Xiong et al. 2021), which is a recent paper in this area. BERT LEF combines the label and the document as a sentence pair and trains BERT for up to 5 epochs to predict the label and the document. This is similar to the Paragraph Vectors flavor of Doc2Vec as it is using the label and document to predict the context. A diagram showing how this approach works in Figure 5.

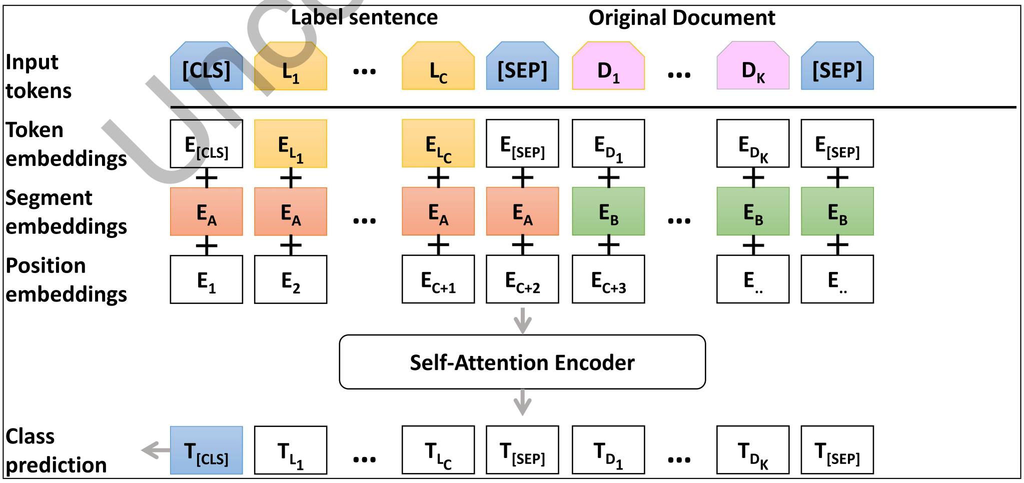


Figure 5
Diagram demonstrating the BERT with Label Embedding Fusion architecture (adapted from Xiong et al., 2021).

4.6 Approach Summary

We summarize the different approaches we will explore in Table 2. “Model” is the training part and “Smoothing” is the smoothing part. “Data” indicates if the underlying data is a manually crafted set of features (“Grieve List”), raw text, or some other data. “Train epochs” is the number of epochs the models were trained in total. “Smooth Iter” is the number of smoothing iterations in total. “Dim” is the final dimension size of the embeddings.

Table 2
Different embedding methods we explore in our analysis. “Model” is the training approach. “Smoothing” is the smoothing approach. “Data” is the data used in this approach, specifically raw text or otherwise. “Train Epochs” is the number of train epochs. Doc2vec approaches have 10 epochs and BERTLEF approaches have 5 epochs to follow previous work. “Smooth Iter” is the number of smoothing iterations. “Dim” is the dimension of the embeddings.

Model	Smoothing	Data	Train Epochs	Smooth Iter	Dim
Static	None	Ones	None	None	1
Coordinates	None	Lat–Long	None	None	2
MVP	AKS B=1	Grieve list	None	1	45
MVP + PCA	AKS B=1	Grieve list	None	1	15
MVP	AKS B=10	Grieve list	None	1	45
MVP + PCA	AKS B=10	Grieve list	None	1	15
MVP	AKS B=100	Grieve list	None	1	45
MVP + PCA	AKS B=100	Grieve list	None	1	15
Random 300	None	None	None	None	300
Random 300	Retrofitting	None	None	50	300
Doc2Vec	None	Raw text	10	None	300
Doc2Vec	AKS B=1	Raw text	10	1	300
Doc2Vec + PCA	AKS B=1	Raw text	10	1	15
Doc2Vec	AKS B=10	Raw text	10	1	300
Doc2Vec + PCA	AKS B=10	Raw text	10	1	15
Doc2Vec	AKS B=100	Raw text	10	1	300
Doc2Vec + PCA	AKS B=100	Raw text	10	1	15
Doc2Vec	Retrofitting	Raw text	10	50	300
Doc2Vec	Alternating	Raw text	10	50	300
Random 768	None	None	None	None	768
Random 768	Retrofitting	None	None	50	768
BERTLEF	None	Raw text	5	None	768
BERTLEF	AKS B=1	Raw text	5	1	768
BERTLEF + PCA	AKS B=1	Raw text	5	1	15
BERTLEF	AKS B=10	Raw text	5	1	768
BERTLEF + PCA	AKS B=10	Raw text	5	1	15
BERTLEF	AKS B=100	Raw text	5	1	768
BERTLEF + PCA	AKS B=100	Raw text	5	1	15
BERTLEF	Retrofitting	Raw text	5	50	768
BERTLEF	Alternating	Raw text	5	50	768

We have six baselines. The first is “Static” which is just a single constant value and emulates the use of static embeddings. The second is “Coordinates”, which uses a representative point[4] of the voting precinct as the embedding. “Lat–Long” refer to latitude and longitude. “Random 300 None” and “Random 768 None” are random embeddings with no smoothing. “Random 300 Retrofitting” and “Random 768 Retrofitting” are random vectors where retrofitting is applied. As discussed in Section 4.3.1, these correspond to embeddings that capture geographic information and do not contain any linguistic information.

We then have the count-based approached by Huang et al. (2016). “MVP” is Mean-Variant-Preference (Section 4.1.1). “AKS” is adaptive kernel smoothing, “B” is the multiplier, and “PCA” is applying PCA after AKS (Section 4.1.2). “Grieve list” is a list of sets of sociologically-relevant lexical variants described in Section 4.1.1.

Finally, we have the machine learning and iterated smoothing methods. “Doc2Vec” is Doc2Vec (Section 4.2.1). “BERTLEF” is BERT with Label Embedding Fusion (Section 4.5). “Retrofitting” applies smoothing after training (Section 4.2.2) and “Alternating” alternates smoothing with training (Section 4.4). “Raw text” means that the model is trained on text instead of manually crafted features.

5. Quantitative Evaluation

5.1 Prediction of Dialect Area from Dialect-specific Terms

Our first evaluation measures how well embeddings can be used to map a dialect when provided some words specific to that dialect. We use the dialect divisions in DAREDS (Rahimi, Cohn, and Baldwin 2017), which divides the United States into 99 dialect regions, each with their own set of unique terms. These regions and terms were compiled from the Dictionary of American Regional English (Cassidy, Hall, and Von Schneidemesser 1985). As our focus is on the state of Texas, we only use the “Gulf States”, “Southwest”, “Texas”, and “West” dialects, each of which include cities in Texas. The list of terms that are specific to those regions can be found in Section Appendix B.

We measure the efficacy of an embedding by how well it can be used to predict how often dialect specific terms are used in a given voting precinct. Given that we have a set number of tweets in each voting precinct and are trying to predict the amount of times dialect specific terms are used, we assume that the underlying process is a Poisson distribution as we are counting the number of times an event is seen (dialect term) in a specific exposure period (number of tweets). A Poisson distribution with rate parameter λ is a probability distribution on $\{0, \ldots, \infty$ with the following probability mass function:

$$Pois(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

If an embedding method captures variational language use, then a Poisson regression fit on those embeddings should accurately emulate this Poisson distribution. Poisson regression is like regular linear regression except it assumes that errors follow a Poisson distribution around the mean instead of a Normal distribution.

One particular issue that is faced with performing Poisson regression with large embeddings is that models may not converge due to data separation (Mansournia et al. 2018). To correct this, we use bias-reduction methods (Firth 1993; Kosmidis and

[4] The representative point is produced by Shapely’s (Gillies et al. 2007) representative_point method.

Firth 2009), which are proven to always produce finite parameter estimates (Heinze and Schemper 2002). We use R's brglm2 package (Kosmidis 2020) to do this.

To evaluate the fit, we use two metrics: Akaike information criterion (AIC) and McFadden's pseudo-R^2. AIC is an information theoretic measure of goodness of fit. We choose AIC as its robust to number of parameters and, assuming we are correct about the underlying distribution being Poisson, it is asymptotically equivalent to Leave One Out Cross Validation (Stone 1977). AIC is given by the following formula:

$$AIC = 2 * \text{number of model parameters} - 2 * \text{maximum likelihood of model}$$

Table 3
Results of dialect area prediction evaluation for relevant DAREDS regions. The values are AIC for each region (lower is better).

		DAREDS AIC by Region			
Method	Alternation	Gulf States	Southwest	Texas	West
Static	None	4890.32	8793.00	7885.50	6236.38
Coordinates	None	4859.89	8159.15	7681.31	6090.05
MVP	AKS B=1	4713.70	8251.73	7214.86	6078.22
MVP + PCA	AKS B=1	4713.31	8492.32	7523.04	6110.55
MVP	AKS B=10	4696.95	7697.70	7011.86	5933.71
MVP + PCA	AKS B=10	4725.05	8324.49	7483.78	6060.23
MVP	AKS B=100	4581.97	7421.84	7123.18	5861.19
MVP + PCA	AKS B=100	4584.86	7710.95	7382.14	5950.82
Random 300	None	4878.53	7441.02	6780.70	6065.14
Random 300	Retrofitting	4778.34	7196.95	6372.70	5797.75
Doc2Vec	None	4599.22	6746.71	6145.31	5511.69
Doc2Vec	AKS B=1	4945.14	7940.38	7498.78	6088.75
Doc2Vec + PCA	AKS B=1	4859.17	8706.27	7819.10	6187.54
Doc2Vec	AKS B=10	4907.23	7589.73	7211.45	6058.02
Doc2Vec + PCA	AKS B=10	4874.47	8662.70	7827.59	6153.67
Doc2Vec	AKS B=100	5017.93	7916.88	7038.32	6093.19
Doc2Vec + PCA	AKS B=100	4880.77	8689.66	7869.85	6182.27
Doc2Vec	Retrofitting	4814.15	7164.03	6433.94	5802.43
Doc2Vec	Alternating	4689.96	6919.24	6192.12	5659.31
Random 768	None	5345.06	7211.48	6609.13	6029.10
Random 768	Retrofitting	5366.13	7349.66	6534.66	6221.10
BERTLEF	None	5299.95	7211.09	6521.57	6260.76
BERTLEF	AKS B=1	5292.91	7217.49	6828.36	6212.75
BERTLEF + PCA	AKS B=1	4870.77	8601.52	7860.10	6208.87
BERTLEF	AKS B=10	5286.53	7390.63	6793.89	6172.18
BERTLEF + PCA	AKS B=10	4870.26	8647.27	7847.80	6215.73
BERTLEF	AKS B=100	5382.80	7538.72	6630.50	6176.40
BERTLEF + PCA	AKS B=100	4894.13	8639.23	7858.67	6230.27
BERTLEF	Retrofitting	5450.53	7619.40	6875.99	6355.34
BERTLEF	Alternating	5308.68	7377.52	6511.52	6124.20

We show the AIC scores for the various precinct embedding approaches in Table 3. See Section 4.6 for a reference for the method names. In the Gulf States region, we see that methods that use manually crafted lists of lexical variants (MVP models) are competitive with machine learning–based models applied to raw text with the largest neighborhood size outperforming these methods. However, in the other regions, the Doc2Vec approaches that use Retrofitting and Alternating smoothing greatly outperform those approaches. What this indicates is that if we have a priori knowledge of sociolinguistically relevant lexical variants then we can accurately predict dialect areas. However, machine learning methods can achieve similar or greater results with just raw text. Thus, even when lexical variant information is unavailable, we can still make accurate predictions.

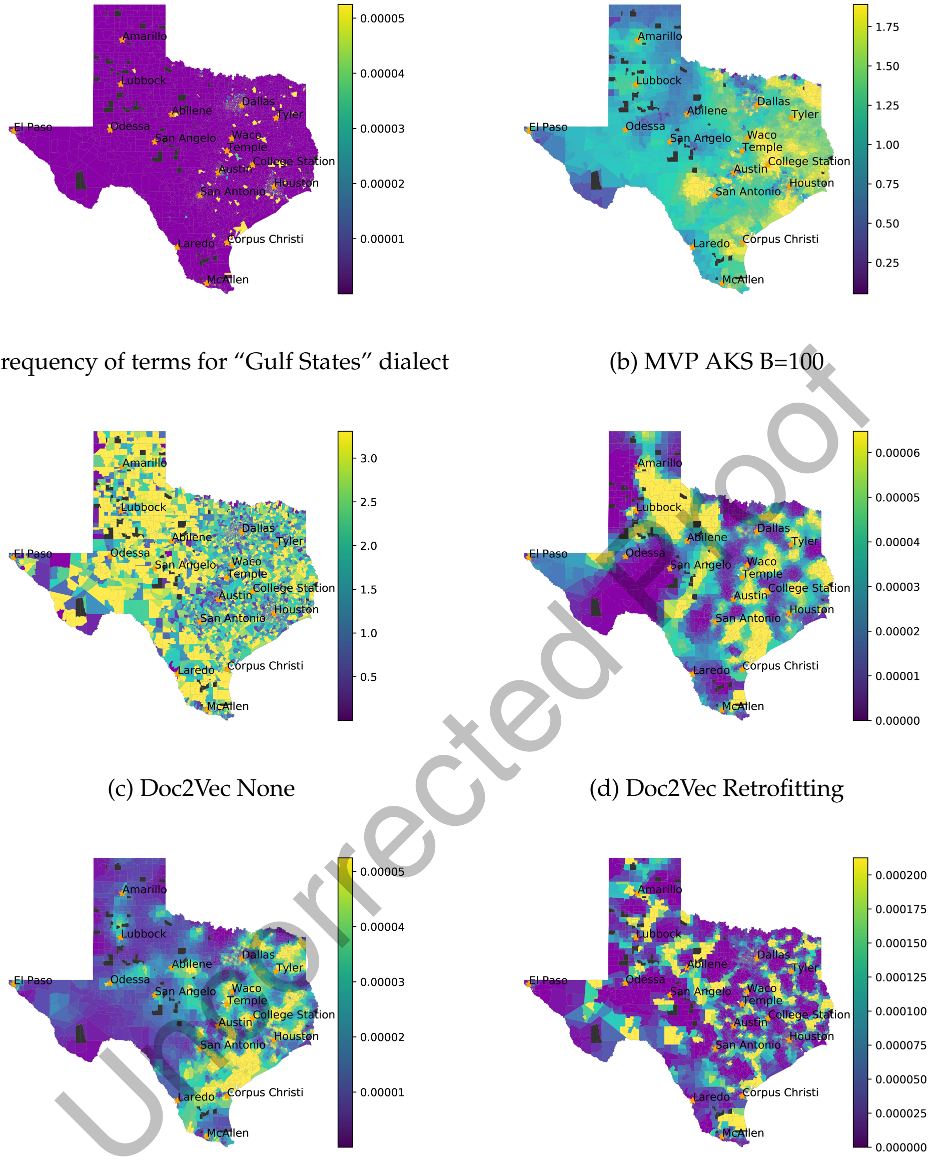
Among the Doc2Vec approaches, we see that Alternating smoothing does better than all other forms of smoothing. More than that, Alternating smoothing is the only one that consistently beats the geography only baseline (Random 300 Retrofitting). In other words, the other smoothing approaches may not be leveraging as much linguistic information as they could and may be overpowered by the geography signal. In contrast, alternating smoothing and training produces embeddings that provide more than what can be provided by geography alone.

In the table, we see that Doc2Vec without smoothing outperforms Doc2Vec with smoothing. We see similar phenomenon with the BERTLEF models. The nature of the task may benefit Doc2Vec without smoothing as counts in an area are going to be higher in places with more data. However, we see that Doc2Vec Alternating smoothing does better than every other smoothing variant across the board. In particular, Alternating smoothing outperforms the AKS approaches. What that indicates is that the effectiveness of MVP models is due to the manually crafted list of lexical variants and less due to the smoothing approach.

In Figures 6–9, we visualize the predictions of a select set of methods for the relevant DAREDS regions.[5] In each one, we see that Doc2Vec None produces a noisy, largely indiscernable pattern, indicating that the high score may be related to the model learning the artifacts of the dataset. In contrast, the Doc2Vec Alternating (panel e) and MVP AKS B=100 (panel b) produce patterns that make sense, for example, the prediction of the "Gulf States" region is near the Gulf of Mexico (southeast of Texas) for which the region is named. Similarly, these models predict the "Southwest" and "West" regions are to the southwest and west, respectively. Of particular note, these predictions match the locations of where the words were used, as shown in subfigure a. In contrast, the Doc2Vec Retrofitting (panel d) and BERTLEF Alternating (panel f) show some appropriate regional patterns, but are much messier than Doc2Vec Alternating, which corroborates their score.

BERT based models generally do worse than their Doc2Vec counterparts. One possibility is that the added value of using a BERT model doesn't outgain the increase in parameters (768 parameters in BERT to 300 parameters in Doc2Vec). What this indicates is that the added pretraining done with BERT may not provide the obvious boost in analyzing lexical variation as is seen in other kinds of tasks. Additionally, while we see that Alternating smoothing does better than Retrofitting, both are worse than the AKS smoothing methods and Retrofitting smoothing is worse than the random vector

[5] As Poisson regressions can go to infinity, we cap the values to a standard deviation above the mean to prevent particularly large predictions hiding other predictions.



(a) Frequency of terms for "Gulf States" dialect

(b) MVP AKS B=100

(c) Doc2Vec None

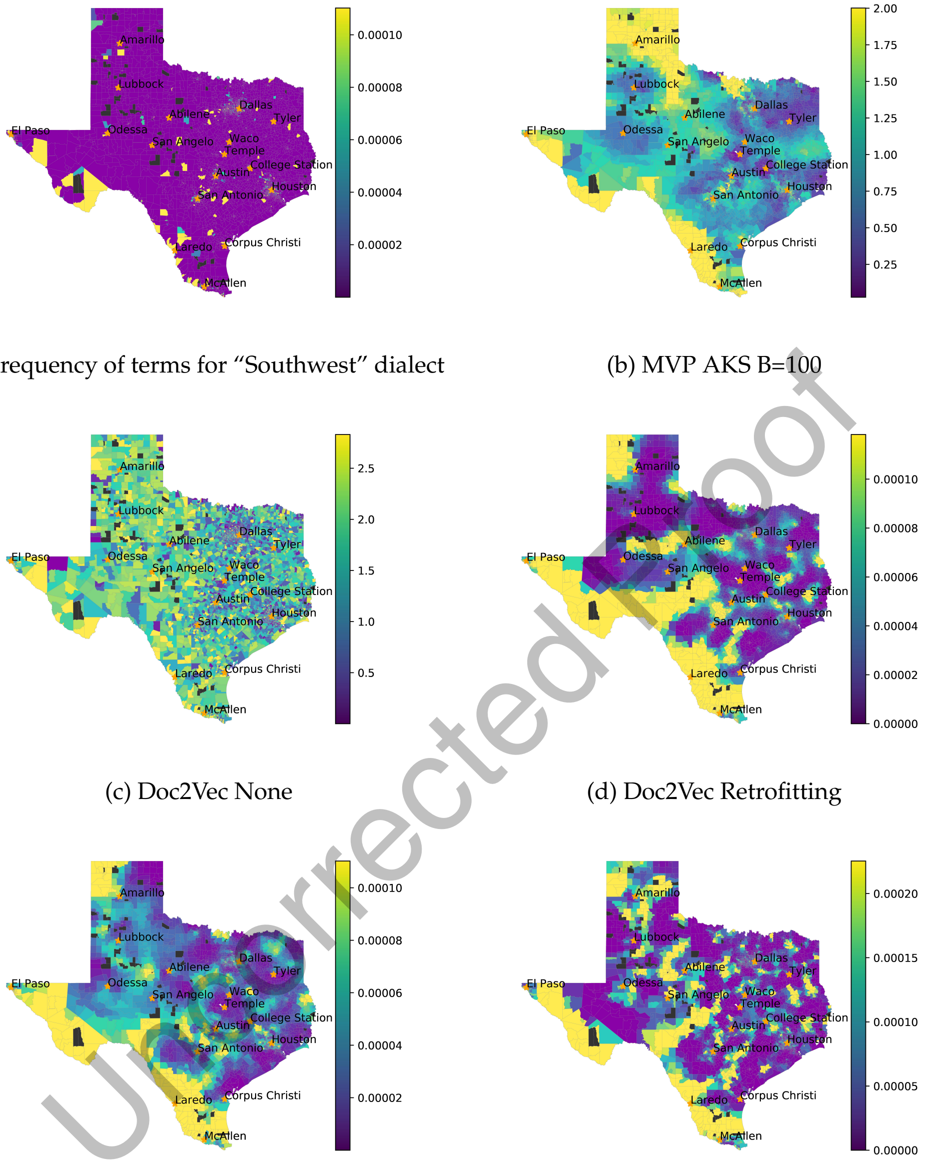
(d) Doc2Vec Retrofitting

(e) Doc2Vec Alternating

(f) BERTLEF Alternating

Figure 6
Predicted location of "Gulf States" dialect using various embedding approaches.

baseline. In Figure 10, we show a possible explanation and explore this phenomenon in more detail in the next evaluation. The figure shows the tradeoff between number of smoothing iterations and AIC. Generally, Retrofitting increases in AIC with more iterations, which is bad. Thus, for our data, retrofitting may actually be detrimental and therefore fewer iterations would be less harmful. In contrast, with Alternating



(a) Frequency of terms for “Southwest” dialect

(b) MVP AKS B=100

(c) Doc2Vec None

(d) Doc2Vec Retrofitting

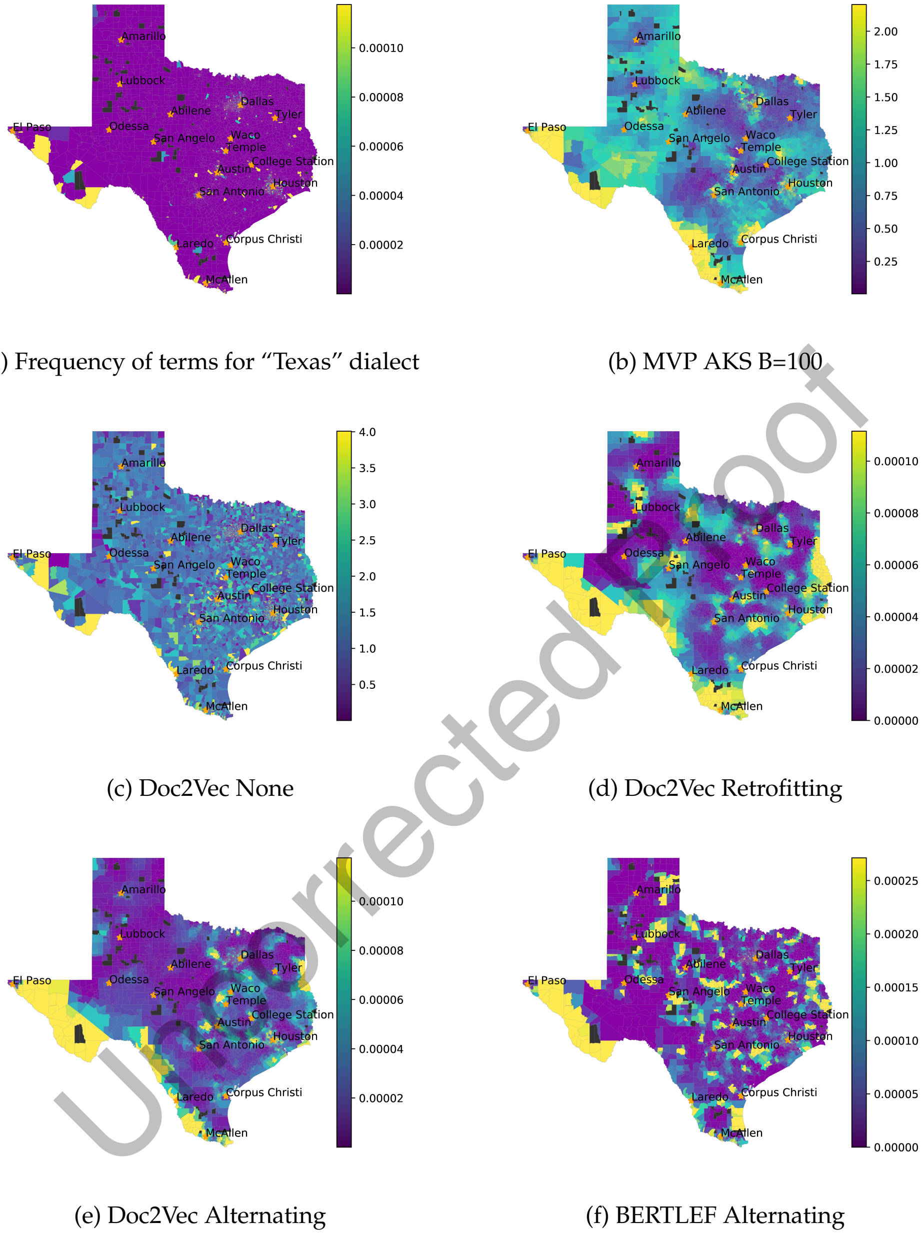
(e) Doc2Vec Alternating

(f) BERTLEF Alternating

Figure 7
Predicted location of “Southwest” dialect using various embedding approaches.

smoothing, we do not see an increase in AIC, which indicates that alternating training and smoothing may mitigate any harm that could be brought from smoothing the data.

The other metric we explore is McFadden’s pseudo-R^2 (McFadden et al. 1973). McFadden’s pseudo-R^2 is a generalization of the coefficient of determination (R^2) that is more appropriate for generalized linear models, such as Poisson regression. Whereas



(a) Frequency of terms for "Texas" dialect

(b) MVP AKS B=100

(c) Doc2Vec None

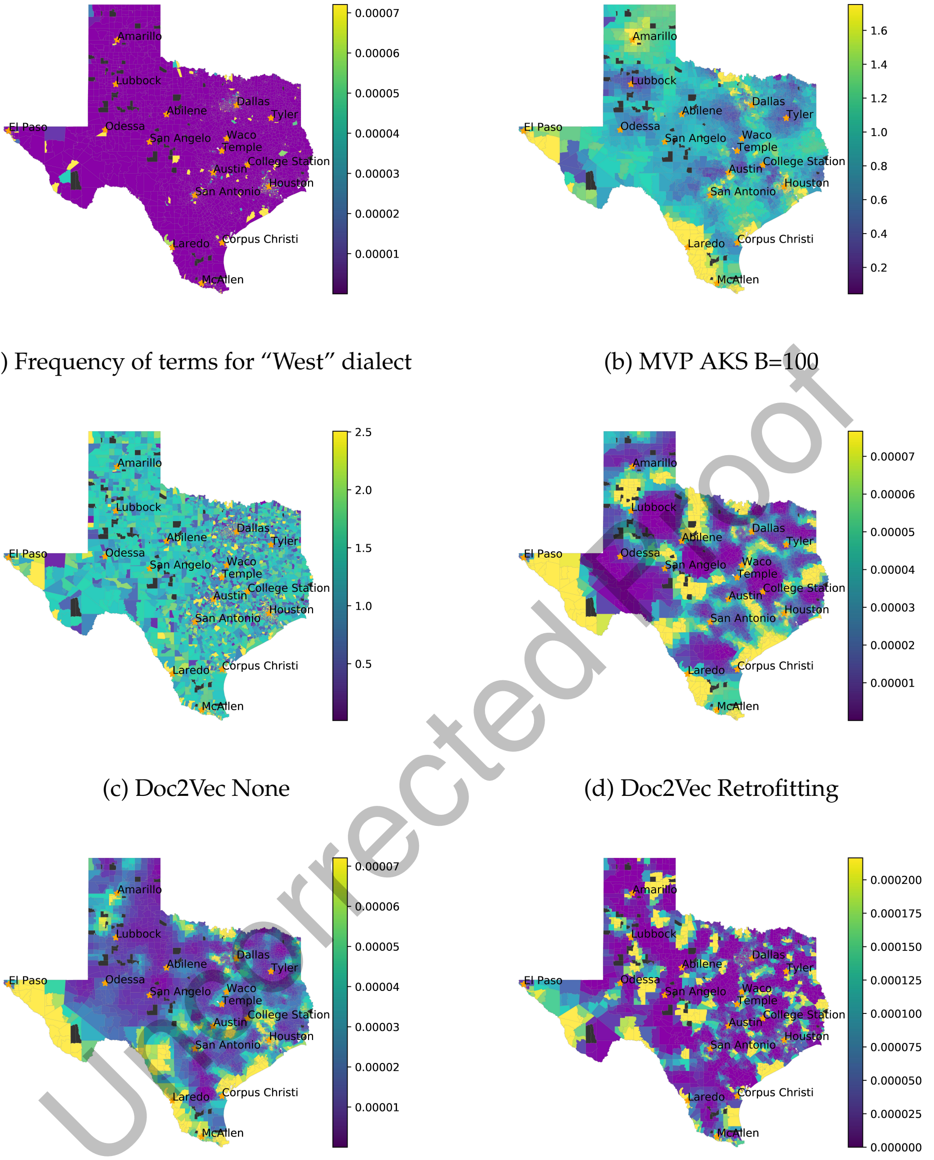
(d) Doc2Vec Retrofitting

(e) Doc2Vec Alternating

(f) BERTLEF Alternating

Figure 8
Predicted location of "Texas" dialect using various embedding approaches.

the coefficient of determination is 1 minus the residual sum of squares divided by the total sum of squares, McFadden's pseudo-R^2 is 1 minus the residual deviance over the null deviance. The deviance of a model is the log-likelihood of the predicted values of the model minus the log-likelihood of the actual values of the model. The residual deviance is the deviance of the model in question and the null deviance is the deviance



(a) Frequency of terms for "West" dialect

(b) MVP AKS B=100

(c) Doc2Vec None

(d) Doc2Vec Retrofitting

(e) Doc2Vec Alternating

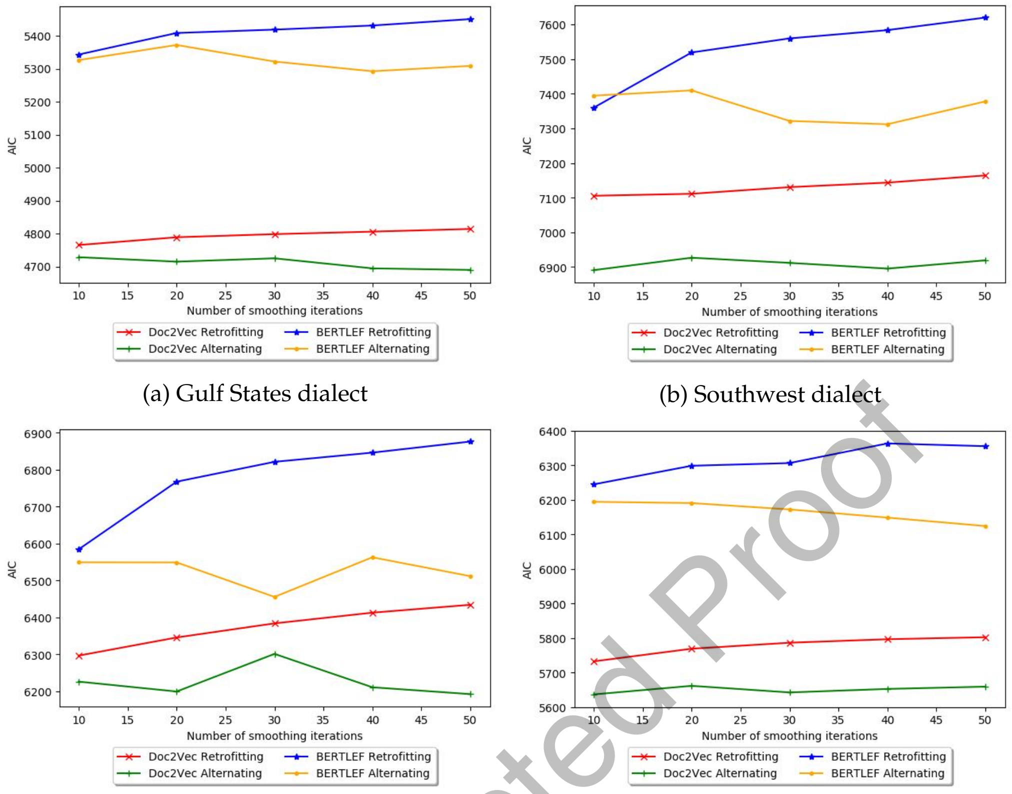
(f) BERTLEF Alternating

Figure 9
Predicted location of "West" dialect using various embedding approaches.

of a model where the probability is the same for every voting precinct (only has an intercept and no embedding information).

McFadden's pseudo-$R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}}$

We chose this metric as well as it produces easier to understand values (1 is the best, 0 means the model is just as good as a constant model, negative numbers indicate that



(a) Gulf States dialect

(b) Southwest dialect

(c) Texas dialect

(d) West dialect

Figure 10
Hyperparameter analysis that compares number of smoothing iterations with AIC.

the model is worse than just using a constant model). However, it does not have many of the nice properties that AIC has.

We provide the corresponding evaluation scores in Table 4 and hyperparameter analysis graphs in Figure 11. R^2 values are largely connected to number of parameters (MVP scores are lower than Doc2Vec scores, which are lower than BERTLEF scores), so comparing models with different parameter sizes is of limited help. What the pseudo-R^2 do tell us is that the embeddings are useful for capturing dialect areas as they are positive (as in, more useful than a constant model). More than this, as values between 0.2 and 0.4 are seen as indicators of excellent fit (McFadden 1977), we see that the Doc2Vec and BERTLEF approaches with Retrofitting and Alternating smoothing provide excellent fits for the data.

5.2 Prediction of Lexical Variant Preference

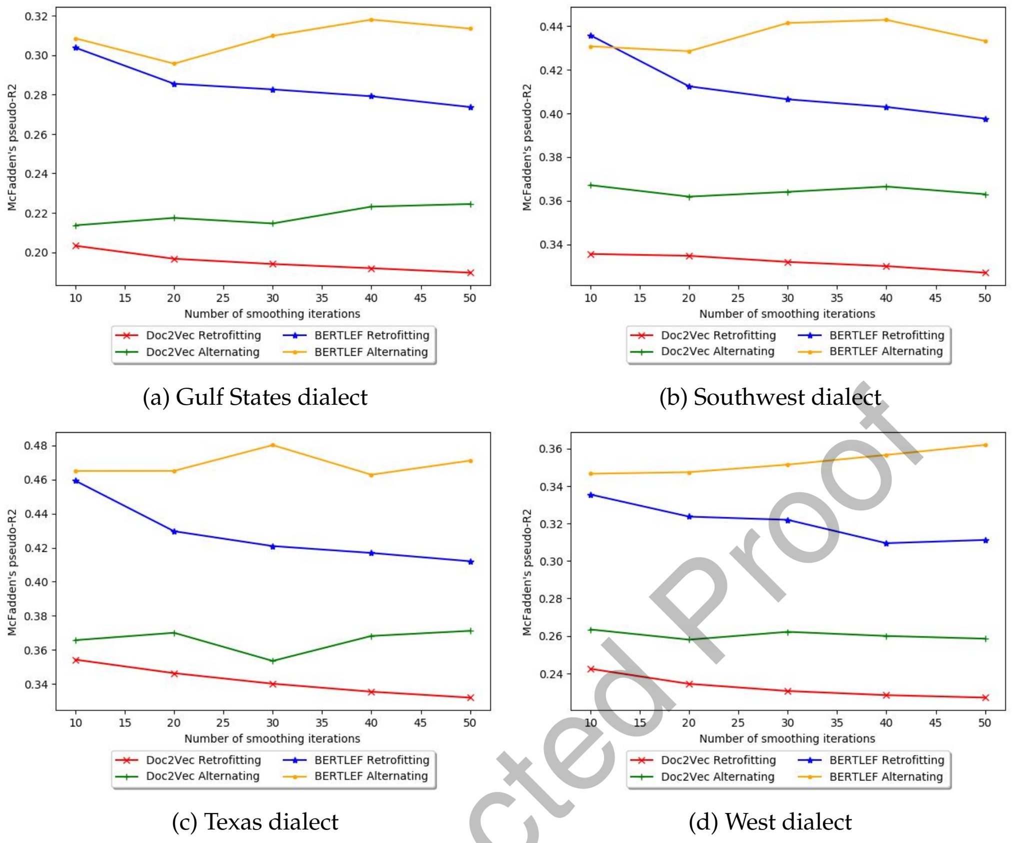
In this section, we evaluate embeddings based on their ability to predict lexical variant preference. Lexical variation is the choice between two semantically similar lexical items, such as *pop* versus *soda*. Lexical variation is a good determiner of linguistic variation (Cassidy, Hall, and Von Schneidemesser 1985; Carver 1987). Thus, if a voting

Table 4
Results of dialect area prediction evaluation for relevant DAREDS regions.
The value is McFadden's pseudo-R^2 for each region (higher is better).

		DAREDS R2 by Region			
Method	Alternation	Gulf States	Southwest	Texas	West
Static	None	0.00	0.00	0.00	0.00
Coordinates	None	0.01	0.09	0.03	0.03
MVP	AKS B=1	0.07	0.09	0.12	0.05
MVP + PCA	AKS B=1	0.06	0.05	0.06	0.03
MVP	AKS B=10	0.08	0.17	0.16	0.09
MVP + PCA	AKS B=10	0.05	0.07	0.07	0.05
MVP	AKS B=100	0.11	0.21	0.14	0.10
MVP + PCA	AKS B=100	0.09	0.16	0.09	0.07
Random 300	None	0.17	0.29	0.28	0.17
Random 300	Retrofitting	0.20	0.32	0.34	0.23
Doc2Vec	None	0.25	0.39	0.38	0.29
Doc2Vec	AKS B=1	0.15	0.21	0.16	0.16
Doc2Vec + PCA	AKS B=1	0.02	0.02	0.02	0.02
Doc2Vec	AKS B=10	0.16	0.26	0.21	0.17
Doc2Vec + PCA	AKS B=10	0.01	0.02	0.01	0.02
Doc2Vec	AKS B=100	0.13	0.22	0.23	0.16
Doc2Vec + PCA	AKS B=100	0.01	0.02	0.01	0.02
Doc2Vec	Retrofitting	0.19	0.33	0.33	0.23
Doc2Vec	Alternating	0.22	0.36	0.37	0.26
Random 768	None	0.30	0.46	0.46	0.38
Random 768	Retrofitting	0.30	0.44	0.47	0.34
BERTLEF	None	0.32	0.46	0.47	0.33
BERTLEF	AKS B=1	0.32	0.46	0.42	0.34
BERTLEF + PCA	AKS B=1	0.01	0.03	0.01	0.01
BERTLEF	AKS B=10	0.32	0.43	0.43	0.35
BERTLEF + PCA	AKS B=10	0.01	0.03	0.01	0.01
BERTLEF	AKS B=100	0.29	0.41	0.45	0.35
BERTLEF + PCA	AKS B=100	0.01	0.03	0.01	0.01
BERTLEF	Retrofitting	0.27	0.40	0.41	0.31
BERTLEF	Alternating	0.31	0.43	0.47	0.36

precinct embedding approach can be used to predict lexical variation, the embeddings should be reflective of linguistic variation.

We model lexical variation as a binomial distribution. We suppose a population can choose between two variants *lex*1 and *lex*2, for example, *pop* and *soda*. Each voting precinct acts like a weighted coin where heads is one variant and tails is the other. Given *n* mentions of soft drinks, this corresponds to *n* flips of the weighted coin. Thus, the number of times a voting precinct uses one form over the other is a binomial distribution.



(a) Gulf States dialect

(b) Southwest dialect

(c) Texas dialect

(d) West dialect

Figure 11
Hyperparameter analysis that compares number of smoothing iterations with McFadden's pseudo-R^2.

If voting precinct embedding approach captures linguistic variation, then they should be able to predict the probability of a voting precinct choosing *lex*1 over *lex*2. In other words, we use binomial regression to predict the probability of a lexical choice from the embeddings. The benefit of this approach is that it naturally handles differences in data size (less data in a precinct just means smaller n) and reliability of the probability (a probability of 50% is more reliable when $n = 500$ than when $n = 2$).

We derive our lexical variation pairs from two Twitter lexical normalization datasets from Han and Baldwin (2011) and Liu et al. (2011). The Han and Baldwin (2011) dataset was formed from three annotators normalizing 1,184 out of vocabulary tokens from 549 English tweets. The Liu et al. (2011) dataset was formed from Amazon Turkers normalizing 3,802 nonstandard tokens (tokens that are rare and diverge from a standard form) from 6,150 tweets. In both cases, humans manually annotated what appears to be "non standard" uses of tokens with their "standard" variants. These pairs therefore reflect lexical variation[6]. We filter out pairs that have data in less than 500 voting

[6] We note that these pairs contain pairs that do not necessarily reflect lexical variation, such as typos. However, drawing the line between typo and variation is a difficult question of its own and beyond the scope of our analysis.

precincts. This leads to a list of 66 pairs from Han and Baldwin (2011) and 110 pairs from Liu et al. (2011). See Sections Appendix C and Appendix D in the Appendix for the list of pairs and statistics. For each voting precinct, we derive the frequency of each variant in a pair directly from our Twitter data.

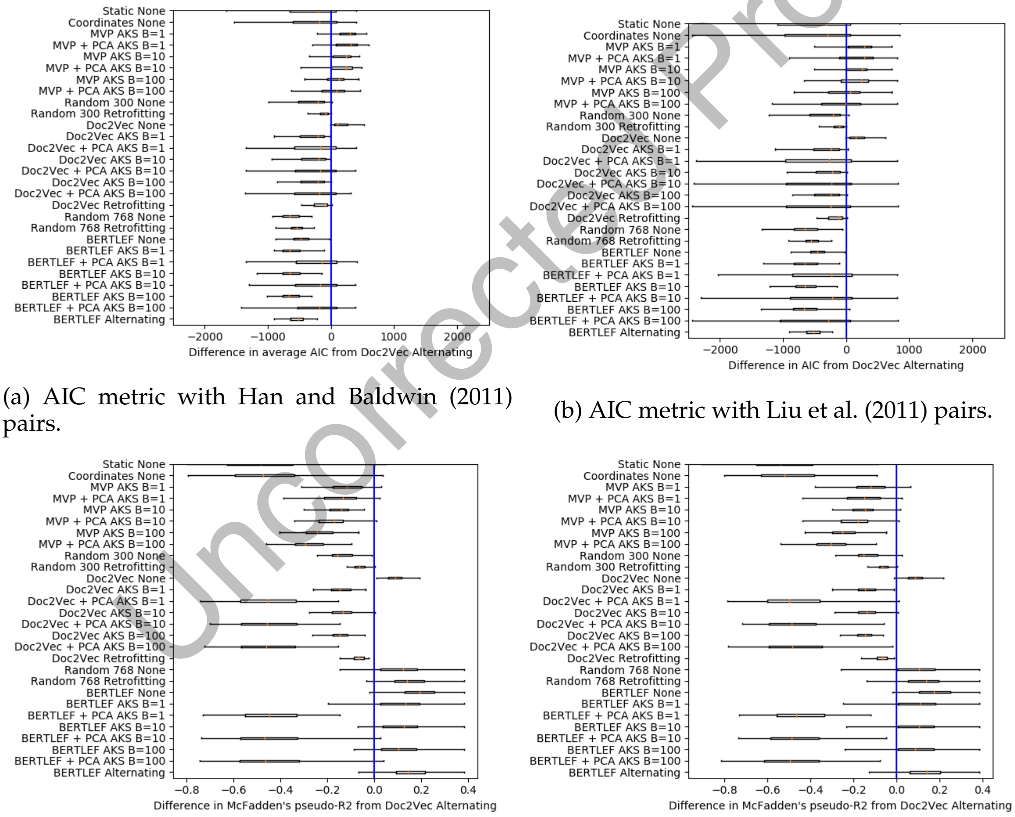
Table 5
Results of lexical variation evaluation for the Han and Baldwin (2011) and Liu et al. (2011) pairs. "AIC" and "R2" are average AIC and McFadden's pseudo-R^2 across pairs. Lower AIC is better and higher pseudo-R^2 is better. "Pairs" are the number of lexical pairs where the binomial regression was fit successfully. "Shared number of pairs" are the number of pairs that succeeded on all models. As BERTLEF with Retrofitting succeeded very few times, we remove it from our analysis.

		Han and Baldwin			Liu et al.		
Method	Alternation	AIC	R2	Pairs	AIC	R2	Pairs
Static	None	5037.90	−0.00	66	7332.17	−0.00	109
Coordinates	None	4820.86	0.02	66	7242.46	0.01	110
MVP	AKS B=1	3968.56	0.37	66	5855.48	0.38	110
MVP + PCA	AKS B=1	4100.76	0.34	66	6248.76	0.34	110
MVP	AKS B=10	3946.91	0.34	66	5810.90	0.35	110
MVP + PCA	AKS B=10	4108.08	0.30	66	6199.99	0.32	110
MVP	AKS B=100	4160.22	0.25	66	5948.60	0.28	110
MVP + PCA	AKS B=100	4263.89	0.21	66	6495.72	0.22	110
Random 300	None	4469.52	0.34	66	5614.97	0.26	110
Random 300	Retrofitting	4173.60	0.42	66	6033.76	0.40	110
Doc2Vec	None	3720.66	0.57	66	4274.39	0.53	110
Doc2Vec	AKS B=1	4601.33	0.33	66	5785.18	0.35	110
Doc2Vec + PCA	AKS B=1	4953.07	0.03	66	7038.40	0.05	110
Doc2Vec	AKS B=10	4460.91	0.34	66	5905.68	−0.35	110
Doc2Vec + PCA	AKS B=10	4914.14	0.04	66	7102.57	−0.10	110
Doc2Vec	AKS B=100	6322.71	−0.86	66	13100.68	−1.34	110
Doc2Vec + PCA	AKS B=100	5247.45	−1.00	66	7139.56	0.05	110
Doc2Vec	Retrofitting	10318.41	−3.26	66	12927.14	−2.94	110
Doc2Vec	Alternating	3991.38	0.48	66	5064.28	0.46	110
Random 768	None	4652.19	0.56	66	5570.99	0.45	110
Random 768	Retrofitting	4501.30	0.59	66	8982.39	0.00	110
BERTLEF	None	4446.72	0.63	66	5360.23	0.51	110
BERTLEF	AKS B=1	4675.30	0.56	62	5576.14	0.46	103
BERTLEF + PCA	AKS B=1	4896.52	0.05	66	6860.40	0.07	110
BERTLEF	AKS B=10	4639.71	0.56	64	5579.60	0.46	107
BERTLEF + PCA	AKS B=10	4922.05	0.04	66	7055.13	0.06	110
BERTLEF	AKS B=100	4698.94	0.56	64	5679.19	0.46	103
BERTLEF + PCA	AKS B=100	4942.70	0.03	66	7269.16	−0.13	110
BERTLEF	Retrofitting	N/A	N/A	22	N/A	N/A	35
BERTLEF	Alternating	4488.41	0.59	66	5880.80	0.49	110
Shared Number of pairs				60			96

With the frequency data, we fit binomial regression models for each pair of words with each voting precinct as a datapoint. Models that have a stronger fit indicate that the corresponding embeddings better capture the choice of variant in the voting precincts.

We present the results of this evaluation in Table 5. See Section 4.6 for a reference for the method names. We see many of the same insights as in the dialect area prediction analysis. We see that MVP approaches are competitive with Doc2Vec Alternating on the Han and Baldwin (2011) and underperform Doc2Vec Alternating on the Liu et al. (2011) dataset. We see that Doc2Vec does better with Alternating smoothing than other approaches and BERTLEF approaches can do worse than baseline.

In Figure 12, we present the difference in AIC and McFadden's pseudo-R^2 across pairs. As different pairs may naturally easier or harder to predict, we compare the Doc2Vec Alternating to provide a more neutral comparison of methods. We see that the MVP approaches tend to have more rightward AIC boxes. Together with the averages



(a) AIC metric with Han and Baldwin (2011) pairs.

(b) AIC metric with Liu et al. (2011) pairs.

(c) McFadden's psuedo-R^2 metric with Han and Baldwin (2011) pairs.

(d) McFadden's psuedo-R^2 metric with Liu et al. (2011) pairs.

Figure 12
Box and whisker plots that show the difference in AIC and pseudo-R^2 between the various methods and Doc2Vec Alternating across lexical variant pairs. The blue line is where the method has an equal AIC/R^2 to Doc2Vec Alternating. Points right of the blue line are pairs where the model outperformed Doc2Vec Alternating.

being close, this indicates that MVP approaches do better than Doc2Vec Alternating more often, but perform much worse when they do perform worse. For the approaches that are applied to raw text (and use smoothing), we see that the boxes are to the left of the blue line, which indicates that they do worse than Doc2Vec Alternating. What this indicates is that among approaches that do not requires manually crafted features, Doc2Vec Alternating performs the best.

Table 5 does also highlight some very different conclusions than the previous evaluation. In the previous evaluation, all methods had a positive McFadden's pseudo-R^2, whereas here we see that many approaches have a negative R^2, which is a sign that predictions are extremely off the mark. We also see that some models, especially Doc2Vec Retrofitting, have AICs that are nearly double the others, which is also a sign of poor prediction. Additionally, we see issues in fitting the binomial regression models in the first place. The "Pairs" column indicates how many of the 66 Han and Baldwin (2011) pairs and 110 Liu et al. (2011) pairs were fit successfully and did not throw collinearity errors. For example, BERTLEF AKS B=1 only had 62 pairs with complete fitting, which means 4 pairs failed to fit. The BERTLEF Retrofitting model succeeded on only about a third of the pairs, so was thrown out. In other words, we see that several models have severe issues in this evaluation.

In Figure 13, we compare the number of smoothing iterations to the average AIC (top graphs), average McFadden's pseudo-R^2 (middle graphs), and number of pairs that were successfully fit. We see that Retrofitting approaches get substantially worse with more iterations. BERTLEF approaches are particularly susceptible to this issue.[7] In contrast, the Alternating smoothing approaches do not have these issues. The Doc2Vec Alternating approach is stable from start to finish and the BERTLEF Alternating approach has more minor deviations.

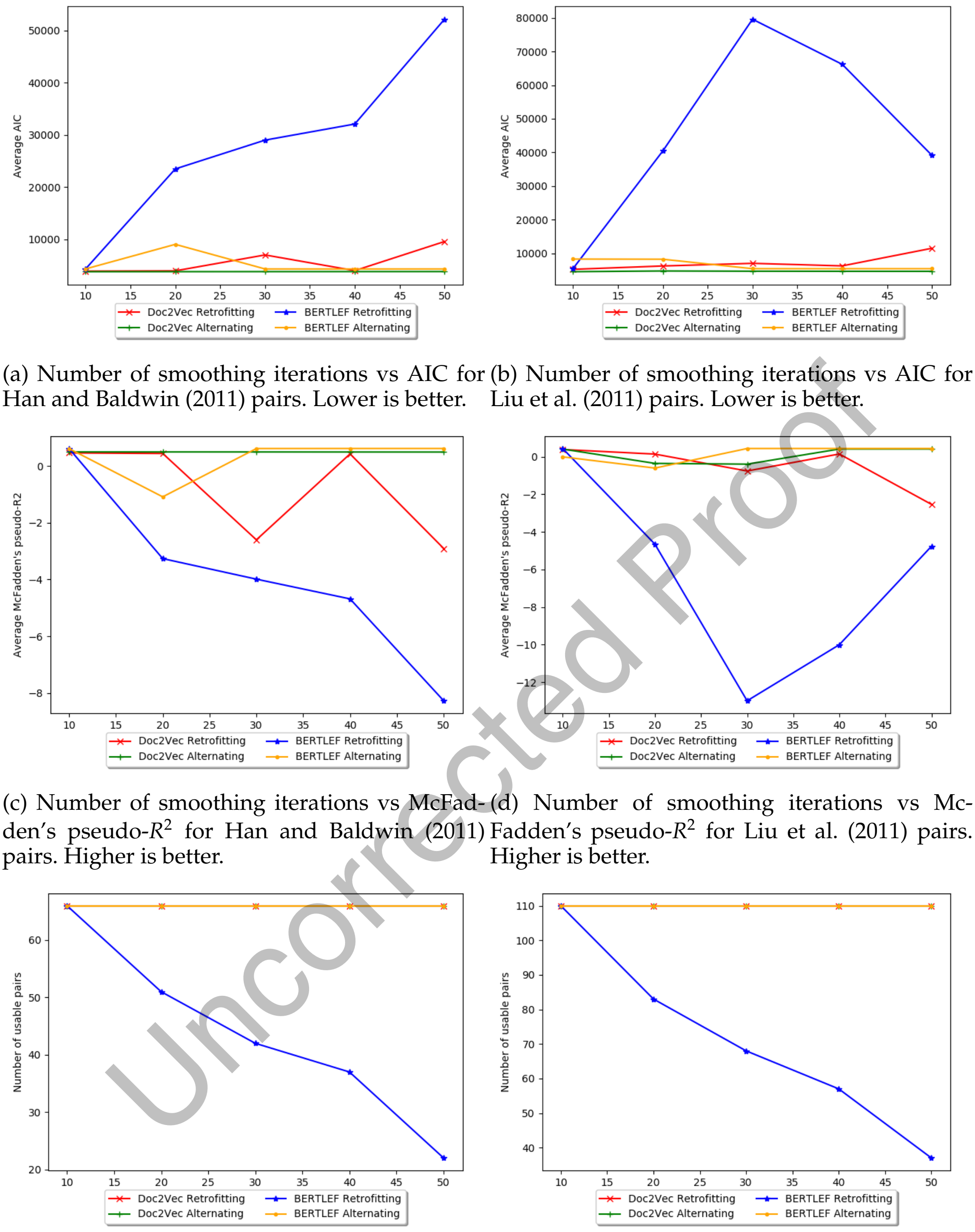
We believe the cause of these problems is that retrofitting, with voting precinct level data, causes the embeddings to become collinear and thus susceptible to modeling issues. In Figure 14, we compare number of smoothing iterations to the column rank of the embedding matrix (as calculated by NumPy's matrix_rank method). The gray lines are the desired rank. Doc2Vec approaches have a dimension of 300 so should have a column rank of 300. BERTLEF have a dimension of 768 so should have a column rank of 768. In the figure, we see that, for Retrofitting approaches, the rank sharply declines, which indicates that smoothing after training causes the embedding dimensions to rapidly become collinear and thus have limited predictive value. In contrast, the Doc2Vec Alternating approach does not suffer any decrease in column rank and the BERTLEF Alternating approach only suffers minor loss in column rank.

The lesson to draw from this is that, for working with fine-grained areas like voting precincts, alternating training and smoothing is not just a model improvement, but a necessary part to prevent severe numerical issues. With large areas like cities, retrofitting has enough data to prevent the kinds of issues seen here. However, to gain insight at a much smaller resolution, alternating is not just a nice to have, but a necessity.

5.3 Finer Resultion Analyses Through Variant Maps

As with dialect area prediction, we can generate maps that predict where one variant of a word is chosen over another. This may allow sociolinguists to better explore

[7] While BERTLEF Retrofitting results do appear to climb back up, the number of pairs that are being averaged over are decreasing, so may indicate survivor bias and not improvement.



(a) Number of smoothing iterations vs AIC for Han and Baldwin (2011) pairs. Lower is better.

(b) Number of smoothing iterations vs AIC for Liu et al. (2011) pairs. Lower is better.

(c) Number of smoothing iterations vs McFadden's pseudo-R^2 for Han and Baldwin (2011) pairs. Higher is better.

(d) Number of smoothing iterations vs McFadden's pseudo-R^2 for Liu et al. (2011) pairs. Higher is better.

(e) Number of smoothing iterations vs number of successfully fit pairs for Han and Baldwin (2011) pairs. Higher is better.

(f) Number of smoothing iterations vs number of successfully fit pairs for Liu et al. (2011) pairs. Higher is better.

Figure 13
Hyperparameter analysis of lexical variation evaluation.

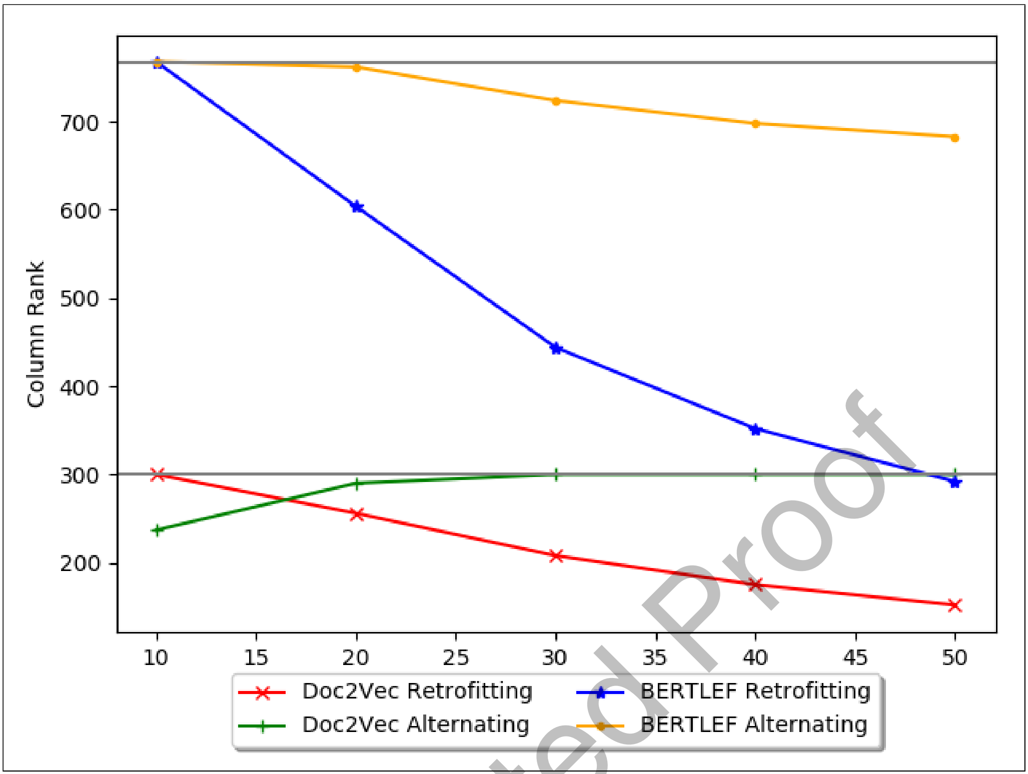
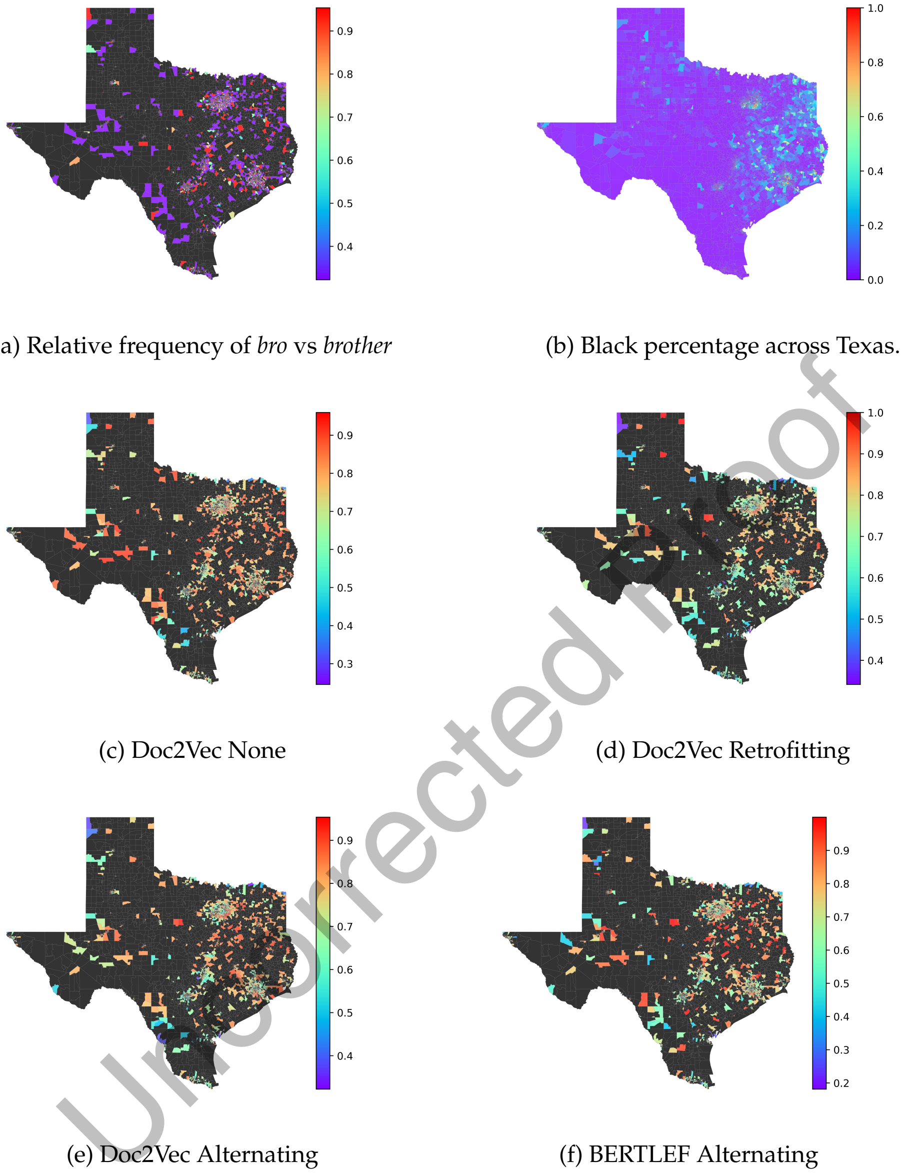


Figure 14
Number of smoothing iterations vs embedding matrix rank. The top gray bar is 768 (full rank for BERT-based methods) and the bottom gray bar is 300 (full rank for Doc2Vec-based methods). Higher is better.

sociolinguistic phenomena. We show an example of this with *bro* vs *brother* in Figure 15.

In panel (a), we have the percentage of times *bro* was used. In panel (b), we have the Black percentage throughout Texas. We include this as *bro* has been recognized as African American slang (Widawski 2015). The bottom four panels are the predicted percentages from various models. We see that both the gold values and Black Percentage have an East–West divide. We also see that the models predict a similar divide with the Retrofitting/Alternating models having a clearer distinction.

A more interesting facet appears when we focus on the divide in *bro* vs *brother* around Houston, Texas (Figure 16). In panel (a), we show the Black Percentage demographics around Houston and see that Black people are not uniformly distributed throughout the city and that there are sections of the city where Black people are more concentrated (highlighted with a red ellipse is one such section). In panel (b), we show our predictions for *bro* vs *brother* from the Doc2Vec Alternating model and see that the predictions are also not uniformly distributed throughout the city and instead are concentrated in the same areas that the Black population are (also highlighted with an ellipse). What this indicates is that using voting precincts as our subregions, we are able to narrow down our analyses to specific, relatively tiny areas.



(a) Relative frequency of *bro* vs *brother*

(b) Black percentage across Texas.

(c) Doc2Vec None

(d) Doc2Vec Retrofitting

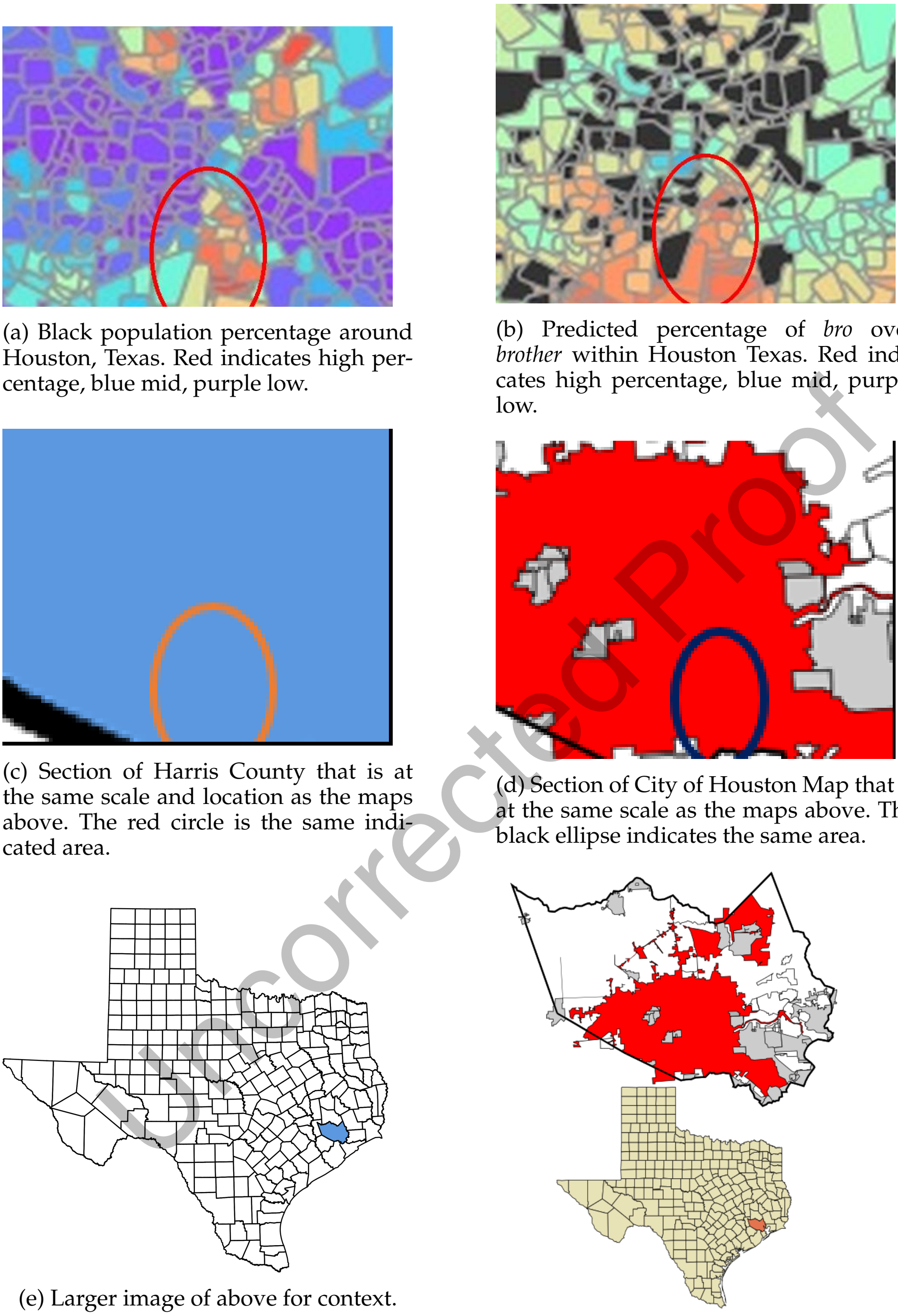
(e) Doc2Vec Alternating

(f) BERTLEF Alternating

Figure 15
Predicted location of *bro* vs *brother* using various embedding approaches. Values are min–max scaled. Black shaded precincts are where neither *bro* nor *brother* are used.

In contrast, larger areas, such as cities and counties, cannot capture these insights. If we use counties instead of voting precincts, as in Huang et al. (2016), we see in panel (c)[8] that the *bro*–*brother* distinction we identified would be enveloped by a single area. If we use cities instead of voting precincts, as in Hovy and Purschke (2018), we see

[8] Images come from US News and World Report and Wikipedia.



(a) Black population percentage around Houston, Texas. Red indicates high percentage, blue mid, purple low.

(b) Predicted percentage of *bro* over *brother* within Houston Texas. Red indicates high percentage, blue mid, purple low.

(c) Section of Harris County that is at the same scale and location as the maps above. The red circle is the same indicated area.

(d) Section of City of Houston Map that is at the same scale as the maps above. The black ellipse indicates the same area.

(e) Larger image of above for context.

(f) Larger image of above for context.

Figure 16
Section of Houston to highlight need for more fine grained areas.

in panel (d) that we would also envelop that area and similarly be completely unable to make any finer-grained analyses. Thus, we have shown that finer-grained subregions can produce finer-grained insights. However, as discussed in previous sections, one needs to use a different modeling approach in order to be able to gain these insights and not run into the data issues.

5.4 Embeddings as Linguistic Gene to Connect Language Use with Sociology

The previous sections describe various embedding methods for representing language use in a voting precinct. Language use in any area is connected to race, socioeconomic status, population density, among many, many other factors and these factors are all represented within the embedding. In this section, we explore how we can extractions of these embeddings that correlate to sociological factors and use these extractions to make sociolinguistic analyses.

Our proposed methodology is similar to how genes are used as a nexus to connect two different biological phenomena. For example, consider the HOX genes. HOX genes are common throughout animal genetic sequences and are responsible for limb formation (such as determining whether a human should grow an arm or a leg out of their shoulder) (Grier et al. 2005). By looking at expressions of HOX genes, researchers have found a connection between HOX genes and genetic disorders related to finger development—for example, synpolydactyly and brachydactyly. From this, researchers identified a possible connection between limb formation and finger development via the HOX gene link.

We use a similar strategy to link sociological phenomena with linguistic phenomena. We have embeddings for each voting precinct (genetic sequences for each species). We can identify what portion of these embeddings correspond to a sociological variable of interest (find the genes for limb formation). We can use these portions to predict a linguistic phenomenon (use gene expressions to predict a separate physiological phenomenon). Then, if successful, we can then link the sociological phenomenon with the linguistic phenomenon (connect limb formation and finger disorders through the HOX genes).

To extract the section of the embedding that corresponds to a sociological variable, we use Orthogonal Matching Pursuit (OMP) which is a linear regression that zeros out all but a fixed number of weights. We can train an OMP model to predict the sociological variable from the voting precinct embeddings. The coordinates with non-zero weights are the section of the embedding that correspond to how the sociological phenomenon interacts with language use in an area. For example, if we use the embeddings to predict Black Percentage in a voting precinct, the extracted section should correlate with how race intersects with language use.

More formally, OMP is a linear regression model where all but a fixed upper bound of weights is zero. For input matrix X, for example, where each row is a voting precinct embedding, output vector y, for example, the corresponding variable, and number of non-zero weights n, OMP minimizes the following loss:

$||y - Xw||$ where w are the regression weights, $||w||_0 \leq n$ and $n > 0$.

We use OMP to extract the 10 coordinates in the precinct embeddings that most correspond to a sociological variable of interest. For example, if our sociological variable was Black Percentage, OMP would give us the 10 coordinates that more correlate with Black Percentage. We can connect Black Percentage to other linguistic phenomenon by how well those 10 coordinates predict a linguistic phenomenon of interest as well as identify new linguistic phenomena that could be related to the sociological variable.

First, we explore what insights we can derive from the Black Percentage "gene" in voting precincts' language "genetic code". We use OMP to identify 10 coordinates that highly correlate with Black Percentage. We can connect this "gene" to linguistic phenomena by using it to predict lexical variation. We can then look at how increase in accuracy by using the gene than the entire genetic code. If we find a lexical variant pair that is better modeled with the gene than the entire embedding, that is an indication that the pair is connected to the sociological variable, here Black Percentage.

We measure increase in accuracy by percent decrease in AIC or percent increase in McFadden's pseudo-R^2. We use percentage increase/decrease to account for different pairs having natural ease of modeling. If a pair has a high percentage increase/decrease, then they are likely to be connected to the underlying sociological variable. We also compare to using the sociological variable directly and the percentage improvement.

In Tables 6 and 7 we show the top 30 lexical variant pairs from Han and Baldwin (2011) and Liu et al. (2011). The Gene columns are the rankings as derived from using the extracted embedding section and the SV columns are using the sociological variables alone. From these, a sociolinguist can look at the rankings and possibly identify insights that were previously missed.

To produce an estimate of the accuracy of these lists, we use the African American slang dictionary in Widawski (2015) as our gold labels and use them to calculate the average precision (AP). We see that using McFadden's pseudo-R^2 provides the best results, with using the "gene" performing slightly better than using the sociological variable on its own. We also see that the "gene" approach provides different predictions from solely using the sociological variable, such as the prediction that the *til* versus *until* distinction was possibly connected to Black Percentage.

This indicates that our approach can provide lexical variants that are connected to sociological variables and thus can be used by sociologists to find new variants that could be useful in research. Our approach is completely unsupervised, so novel changes and spread in different communities can be monitored and continually updated with new data, which is not feasible for traditional methods.

We perform a similar experiment with the Population Density variable. We show the top ranked pairs in Tables 8 and 9. As *g*-dropping is a well explored phenomenon for rural vs urban divide Campbell-Kibler (2005), we use this as our gold data. Here, we see that AIC performs best overall with the "gene" approach slightly outperforming the sociological variable. From these lists, it appears that there is a connection between shortening words and population density, for example, convo vs conversation, gf vs girlfriend, bf vs boyfriend, txt vs text, and prolly vs probably. By using genes, we might be able to identify new connections that we may not found otherwise.

6. Dialect Map Prediction via Visualization

In this section, we use dimensionality reduction techniques applied to the precinct embeddings to geographic boundaries of linguistic variation, or "isoglosses". The precinct embeddings are reduced to RGB color values and hard transition in colors indicate a boundary. To project embeddings into RGB color coordinates, we explore two approaches. The first is principal component analysis (PCA), which is previously used in prior work (Hovy et al. 2020). The second is t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton 2008), which is a probablistic approach often used for visualizing word embedding clusters.

Table 6
Ranking of lexical variation pairs when using extractions from embeddings (Gene) versus using the sociological variable directly (SV). The ranking is done by percentage increase in R2/percentage decrease in AIC from the original embedding to the extraction/sociological variable. AP is the average precision. Bold pairs are pairs that previous research has identified to being relevant to the sociological variable.

Dataset: Han and Baldwin (2011)
Sociological Variable: Black Percentage

Rank	Gene AIC	SV AIC	Gene R2	SV R2
1	umm-um	umm-um	til-until	lil-little
2	convo-conversation	convo-conversation	lil-little	**bro-brother**
3	freakin-freaking	freakin-freaking	**bro-brother**	umm-um
4	gf-girlfriend	gf-girlfriend	convo-conversation	tha-the
5	sayin-saying	sayin-saying	tha-the	gon-gonna
6	chillin-chilling	chillin-chilling	fb-facebook	da-the
7	yess-yes	bf-boyfriend	hrs-hours	yu-you
8	playin-playing	txt-text	comin-coming	fb-facebook
9	lawd-lord	yess-yes	playin-playing	**cuz-because**
10	bf-boyfriend	lawd-lord	**fam-family**	bs-bullshit
11	txt-text	bs-bullshit	btw-between	ppl-people
12	cus-because	ohh-oh	lookin-looking	dat-that
13	ahh-ah	cus-because	de-the	**dawg-dog**
14	prolly-probably	pics-pictures	**dawg-dog**	kno-know
15	ohh-oh	ahh-ah	yu-you	chillin-chilling
16	bs-bullshit	prolly-probably	thx-thanks	til-until
17	nothin-nothing	hahah-haha	**cuz-because**	jus-just
18	hahah-haha	hahahaha-haha	**def-definitely**	bday-birthday
19	naw-no	talkin-talking	da-the	wat-what
20	tht-that	til-till	jus-just	goin-going
21	pics-pictures	naw-no	bday-birthday	de-the
22	talkin-talking	nothin-nothing	ahh-ah	prolly-probably
23	hahahaha-haha	playin-playing	mis-miss	gettin-getting
24	doin-doing	hahaha-haha	mins-minutes	nd-and
25	bb-baby	tht-that	gettin-getting	fuckin-fucking
26	til-till	gon-gonna	kno-know	lookin-looking
27	fb-facebook	doin-doing	doin-doing	naw-no
28	comin-coming	fuckin-fucking	gon-gonna	**fam-family**
29	thx-thanks	bb-baby	soo-so	cus-because
30	kno-know	goin-going	yr-year	mis-miss
AP	0.055	0.057	0.252	0.237

6.1 Principal Component Analysis

PCA is widely used in the humanities for descriptive analyses of data. If we have a collection of continuous variables, PCA essentially creates a new set of axes that captures the greatest variance in the original variables. In particular, the first axis captures

Table 7
Ranking of lexical variation pairs when using extractions from embeddings (Gene) versus using the sociological variable directly (SV). The ranking is done by percentage increase in R2/percentage decrease in AIC from the original embedding to the extraction/sociological variable. AP is the average precision. Bold pairs are pairs that previous research has identified to being relevant to the sociological variable.

Dataset: Liu et al. (2011)

Sociological Variable: Black Percentage

Rank	Gene AIC	SV AIC	Gene R2	SV R2
1	wheres-whereas	wheres-whereas	**homies-homes**	trippin-tripping
2	quiero-query	quiero-query	cali-california	lil-little
3	max-maximum	max-maximum	re-regarding	**bro-brother**
4	tv-television	tv-television	mo-more	tha-the
5	**homies-homes**	bbq-barbeque	trippin-tripping	wit-with
6	re-regarding	**homies-homes**	lil-little	yo-you
7	bbq-barbeque	cali-california	**bro-brother**	bout-about
8	cali-california	trippin-tripping	convo-conversation	tho-though
9	convo-conversation	convo-conversation	fa-for	da-the
10	trippin-tripping	freakin-freaking	wit-with	yea-yeah
11	freakin-freaking	gf-girlfriend	tha-the	cause-because
12	mines-mine	mines-mine	th-the	yu-you
13	gf-girlfriend	sayin-saying	fb-facebook	fb-facebook
14	sayin-saying	chillin-chilling	bout-about	dis-this
15	chillin-chilling	txt-text	hrs-hours	gon-going
16	yess-yes	cutie-cute	tho-though	**cuz-because**
17	playin-playing	yess-yes	comin-coming	bs-bullshit
18	lawd-lord	nun-nothing	fr-for	ppl-people
19	txt-text	lawd-lord	playin-playing	dat-that
20	cus-because	bs-bullshit	dis-this	sum-some
21	cutie-cute	ohh-oh	**fam-family**	fr-for
22	nun-nothing	cus-because	fml-family	kno-know
23	wen-when	wen-when	fav-favorite	quiero-query
24	wut-what	pics-pictures	yo-you	chillin-chilling
25	prolly-probably	wut-what	hwy-highway	tv-television
26	ohh-oh	prolly-probably	app-application	jus-just
27	thot-thought	sis-sister	thru-through	thang-thing
28	nada-nothing	thot-thought	sum-some	mo-more
29	turnt-turn	feelin-feeling	lookin-looking	bday-birthday
30	sis-sister	talkin-talking	yu-you	wat-what
AP	0.080	0.077	0.264	0.110

the greatest variance in the data, the second axis captures the second greatest variance, and so on. By quantifying the connection between the original variables and the axes, researchers can explore what variables have the most impact in the data. For example, Huang et al. (2016) use this approach to explore the geographic information contained inside area embeddings.

Table 8
Ranking of lexical variation pairs when using extractions from embeddings (Gene) versus using the sociological variable directly (SV). The ranking is done by percentage increase in R2/percentage decrease in AIC from the original embedding to the extraction/sociological variable. AP is the average precision. Bold pairs are pairs that previous research has identified to being relevant to the sociological variable.

Dataset: Han and Baldwin (2011)

Sociological Variable: Population Density (log scaled)

Rank	Gene AIC	SV AIC	Gene R2	SV R2
1	umm-um	umm-um	de-the	til-until
2	convo-conversation	convo-conversation	til-until	**fuckin-fucking**
3	**freakin-freaking**	**freakin-freaking**	convo-conversation	hahaha-haha
4	gf-girlfriend	gf-girlfriend	dawg-dog	**lookin-looking**
5	**sayin-saying**	**sayin-saying**	mis-miss	hahah-haha
6	yess-yes	txt-text	hrs-hours	btw-between
7	**chillin-chilling**	**chillin-chilling**	mins-minutes	hahahaha-haha
8	bf-boyfriend	bf-boyfriend	yu-you	yess-yes
9	txt-text	yess-yes	fb-facebook	**talkin-talking**
10	cus-because	lawd-lord	**comin-coming**	naw-no
11	lawd-lord	cus-because	tha-the	cus-because
12	ahh-ah	ohh-oh	**playin-playing**	de-the
13	**playin-playing**	bs-bullshit	**lookin-looking**	prolly-probably
14	ohh-oh	hahah-haha	bro-brother	mis-miss
15	prolly-probably	ahh-ah	ahh-ah	fam-family
16	bs-bullshit	prolly-probably	cus-because	**freakin-freaking**
17	hahah-haha	pics-pictures	gon-gonna	til-till
18	pics-pictures	hahahaha-haha	fam-family	**goin-going**
19	**nothin-nothing**	**talkin-talking**	congrats-congratulations	lil-little
20	naw-no	naw-no	pic-picture	hrs-hours
21	hahahaha-haha	til-till	nd-and	bs-bullshit
22	**talkin-talking**	**nothin-nothing**	thx-thanks	pls-please
23	tht-that	hahaha-haha	lil-little	nah-no
24	mis-miss	**playin-playing**	cuz-because	congrats-congratulations
25	til-till	tht-that	prolly-probably	def-definitely
26	**doin-doing**	**fuckin-fucking**	**fuckin-fucking**	da-the
27	hahaha-haha	bb-baby	yess-yes	**sayin-saying**
28	bb-baby	**doin-doing**	da-the	tht-that
29	**fuckin-fucking**	**goin-going**	yr-year	dawg-dog
30	gon-gonna	pic-picture	wat-what	txt-text
AP	0.293	0.278	0.164	0.264

Hovy et al. (2020) use PCA to produce variation maps by reducing area embeddings to three dimensions and then standardizing these dimensions to between 0 and 1 to be used as RGB values. We perform a similar analysis for a select set of methods in the left images in Figures 17 and 18. We see that the geography only approach (Random 300 Retrofitting) produces a mostly random pattern of areas while the Doc2Vec None approach produces some regionalization, but is rather noisy.

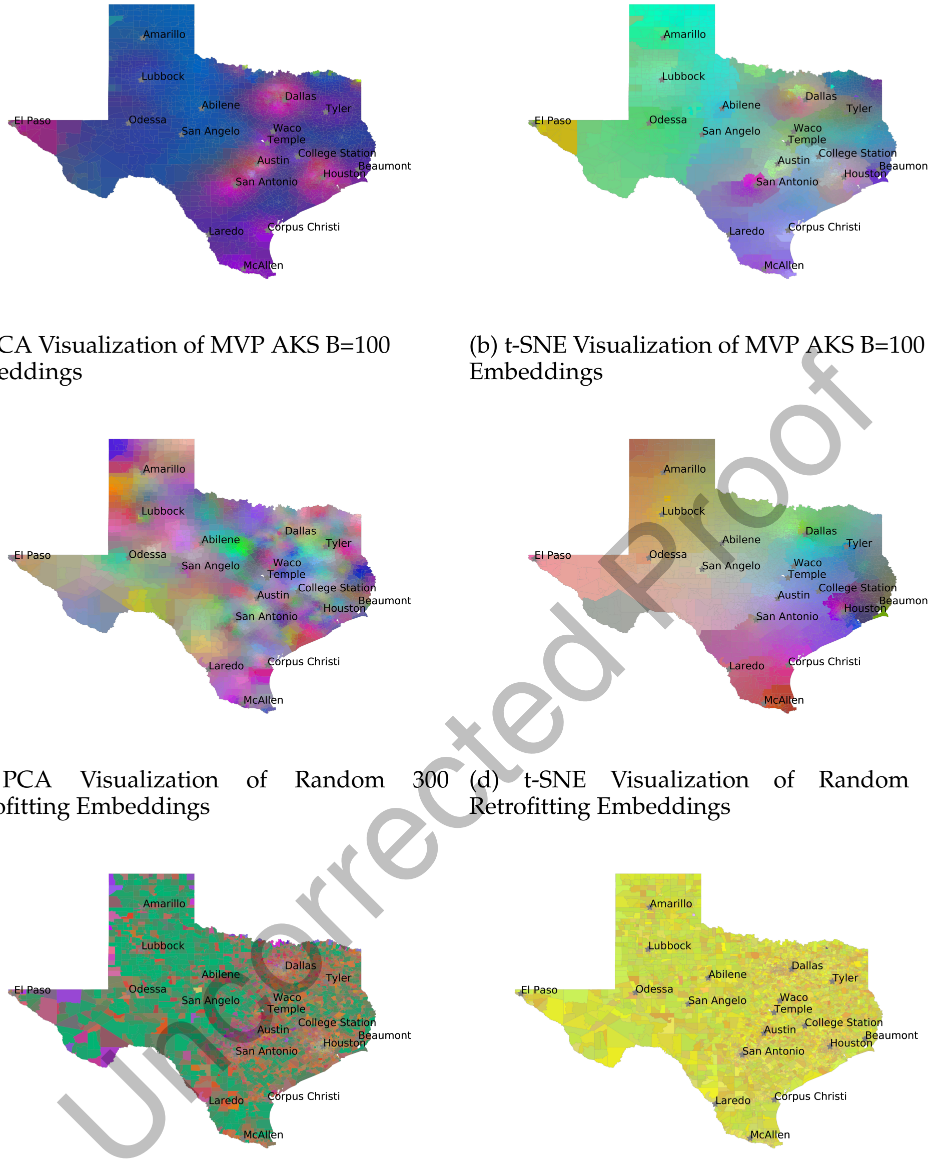
Table 9
Ranking of lexical variation pairs when using extractions from embeddings (Gene) versus using the sociological variable directly (SV). The ranking is done by percentage increase in R2/percentage decrease in AIC from the original embedding to the extraction/sociological variable. AP is the average precision. Bold pairs are pairs that previous research has identified to being relevant to the sociological variable.

Dataset: Liu et al. (2011)

Sociological Variable: Population Density (log scaled)

Rank	Gene AIC	SV AIC	Gene R2	SV R2
1	wheres-whereas	wheres-whereas	homies-homes	mo-more
2	quiero-query	quiero-query	cali-california	th-the
3	max-maximum	max-maximum	mo-more	hr-hour
4	tv-television	tv-television	re-regarding	ft-feet
5	homies-homes	bbq-barbeque	fa-for	wut-what
6	bbq-barbeque	homies-homes	dis-this	**fuckin-fucking**
7	re-regarding	cali-california	**trippin-tripping**	**lookin-looking**
8	cali-california	**trippin-tripping**	th-the	bby-baby
9	convo-conversation	convo-conversation	convo-conversation	dis-this
10	**trippin-tripping**	**freakin-freaking**	mi-my	fa-for
11	**freakin-freaking**	gf-girlfriend	ft-feet	yess-yes
12	mines-mine	mines-mine	hrs-hours	mi-my
13	gf-girlfriend	**sayin-saying**	hr-hour	nun-nothing
14	**sayin-saying**	txt-text	mins-minutes	em-them
15	yess-yes	**chillin-chilling**	yu-you	**talkin-talking**
16	**chillin-chilling**	yess-yes	fav-favorite	naw-no
17	txt-text	cutie-cute	hwy-highway	bout-about
18	cutie-cute	nun-nothing	fb-facebook	cus-because
19	cus-because	lawd-lord	**comin-coming**	prolly-probably
20	nun-nothing	wut-what	fml-family	yo-you
21	lawd-lord	cus-because	tha-the	fml-family
22	**playin-playing**	ohh-oh	tho-though	fam-family
23	ohh-oh	bs-bullshit	wit-with	**freakin-freaking**
24	wut-what	prolly-probably	**playin-playing**	fr-for
25	prolly-probably	pics-pictures	fr-for	quiero-query
26	bs-bullshit	**talkin-talking**	**lookin-looking**	til-till
27	nada-nothing	sis-sister	nada-nothing	**goin-going**
28	wen-when	bby-baby	bro-brother	lil-little
29	**feelin-feeling**	wen-when	cus-because	hrs-hours
30	sis-sister	**feelin-feeling**	yea-yeah	bs-bullshit
AP	0.197	0.196	0.119	0.151

The smoothing approaches generally highlight the cities (possibly with coloring the cities differently) and leave the countryside a uniform color. In other words, using PCA to produce an isogloss map, we only see the urban–rural divide and do not see larger region divides. The reason that is that the urban–rural divide appears to be the biggest



(a) PCA Visualization of MVP AKS B=100 Embeddings

(b) t-SNE Visualization of MVP AKS B=100 Embeddings

(c) PCA Visualization of Random 300 Retrofitting Embeddings

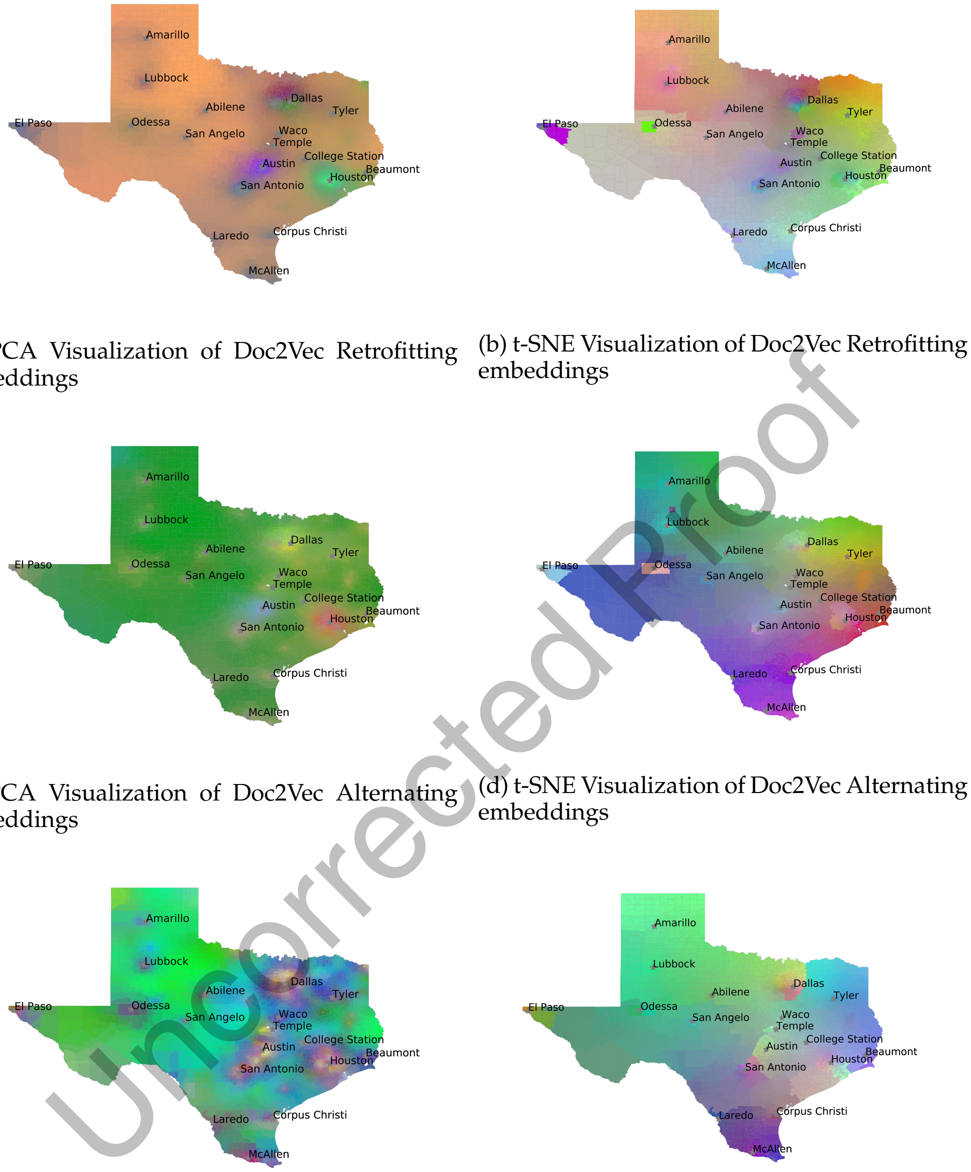
(d) t-SNE Visualization of Random 300 Retrofitting Embeddings

(e) PCA Visualization of Doc2Vec None embeddings

(f) t-SNE Visualization of Doc2Vec None embeddings

Figure 17
Visualization of voting precinct embeddings using PCA (left) and t-SNE (right).

source of variation in the data and PCA is designed to extract the biggest sources of variation. However, by attaching itself to the strongest signal, PCA is unable to find key regional differences in language use. Thus, while PCA approaches are useful for analyzing the information contained in embeddings, it has limited ability to produce isogloss boundaries.



(a) PCA Visualization of Doc2Vec Retrofitting embeddings

(b) t-SNE Visualization of Doc2Vec Retrofitting embeddings

(c) PCA Visualization of Doc2Vec Alternating embeddings

(d) t-SNE Visualization of Doc2Vec Alternating embeddings

(e) PCA Visualization of BERTLEF Alternating embeddings

(f) t-SNE Visualization of BERTLEF Alternating embeddings

Figure 18
Visualization of voting precinct embeddings using PCA (left) and t-SNE (right).

6.2 t-Distributed Stochastic Neighbor Embedding

To fix the above issue, we explore a different dimensionality reduction approach, t-SNE (Van der Maaten and Hinton 2008). Unlike PCA, which tries to find the strongest signals overall, t-SNE instead tries to make sure that points that are similar in the original space are similar in the reduced space. As retrofitting enforces places that are geographically close to have similar embeddings, t-SNE may be much more capable of capturing regions.

The right images in Figures 17 and 18 use t-SNE to visualize embeddings. We see that there are largely three blocks: one block to the East, one block to the Southwest, and one block to the Northwest. This indicates that t-SNE may be better at identifying isoglosses than PCA.

By comparing to the dialect areas in our DAREDS analysis (Section 5.1), we see that the block to the East overlaps nicely with the predicted "Gulf States" dialect region. Similarly, we see that the Southwest block overlaps nicely with the West and Southwest blocks. Finally, the Northwest region seems distinct from the other regions. This indicates that we may have a region that is not accounted for by the Dictionary of American Regional English (Cassidy, Hall, and Von Schneidemesser 1985). It may be because in the nearly 40 years since publication, Texas may have experienced a great linguistic shift. Alternatively, the region may be understudied and thus may reflect a dialect we know little about. In either case, the t-SNE graphs may have shown a particular region of Texas that warrants further investigation.

7. Summary

We demonstrated that it is possible to embed areas as small as voting precincts and that doing so can lead to higher resolution analyses of sociolinguistic phenomena. To make this feasible, we proposed a novel embedding approach that alternates training with smoothing. We showed that both training and smoothing have negative effects when it comes to embedding voting precincts and that smoothing in particular can cause numerical issues. In contrast, we found that alternating training and smoothing mitigates these issues.

We also proposed new evaluations that reflect how voting precinct embeddings can be used directly by sociolinguists. The first explores how well different models are able to predict the location of a dialect given terms specific to that dialect. The second explores how well different models are able to capture preferences in lexical variants, such as the preference between *pop* and *soda*. We then propose a methodology where we identify portions of the embeddings that correspond to sociological variables and use these portions to find novel linguistic insights, thereby connecting sociological variables with linguistic expression. Finally, we explored approaches for using the embeddings to identify isoglosses and showed that PCA overly focuses on the urban–rural divide while t-SNE produces distinct regions.

7.1 Future Work

Finally, we present some directions for future work:

- Although we can produce embeddings that reflect language use in an area, further research is needed to produce more interpretable representations (while retaining accuracy and ease of construction) and more informative uses of regional embeddings. We do propose a method of connecting linguistic phenomena to lexical variation using regional embeddings, but much more work is needed to devise methods that directly address linguists' needs.

- Currently, there is a divide between traditional linguistic approaches to analyzing variation and computational linguistic approaches to analyzing variation. Given access to a wide variety of social media data, one goal may be to close the gap between these approaches and develop definitions of variation that can represent linguistic insights as well as are rigorous and scalable. There is work that uses linguistic features to define regional embeddings (Bohmann 2020), but this still operates under traditional linguistic metrics and region-insensitive methodology (embeddings). Future work could build on our results to produce a flexible definition of variation that could directly leverage Twitter data.
- Finally, a future direction could be to connect the regional embedding work with temporal embedding work (e.g., Hamilton, Leskovec, and Jurafsky 2016; Rosenfeld and Erk 2018) to have a unified spacio–temporal exploration of Twitter data. There is quite a bit of work that does do spacio–temporal work with Twitter data (e.g., Goel et al. 2016; Eisenstein et al. 2014), but this work makes limited use of embedding models. Future work could better explain movement of language patterns with greater accuracy and resolution.

Appendix A. Grieve and Asnaghi (2013) Lexical Variation Pairs

In Table A1, we provide the list of alternates used in our count-based models.

Table A1: Lexical variants from Grieve and Asnaghi (2013) using in our count-based models. "Main" is the variant with the largest frequency. "Alternates" is the list of other variants. "Num VP" are the number of voting precincts that include use of at least one variant. "Main total" is the total frequency of the "Main" variant. "Alt total" is the total frequency of the alternative variants. "P-Value" is the p-value from Moran's I. Gray lines are variant sets that were removed for having a p-value below 0.001 or appear in less than 1000 precincts.

Main	Alternates	Num VP	Main Total	Alt Total	P-Value
before	afore	4416	16267	33	0.000
lane	alley	2684	14615	2939	0.000
car	automobile	6425	309589	162	0.000
baby	infant	5117	21176	187	0.000
bag	sack	2026	4217	381	0.000
ban	prohibit, forbid	4297	29532	235	0.000
beg	plead	2261	5268	138	0.000
best	greatest	5750	32971	1408	0.000
bet	wager	5750	36660	29	0.000
big	large	4979	24258	1326	0.000
bought	purchased	1630	2289	147	0.000
butte	mesa	1342	2250	872	0.000
cab	taxi	1664	3736	288	0.000
center	middle	3314	24299	3878	0.000
clothes	clothing	1733	2342	1254	0.000
understand	comprehend	2761	4937	50	0.000
creek	stream	1332	5075	1179	0.000
dad	father	4705	16457	2344	0.000
dinner	supper	2490	7873	275	0.000
sleepy	drowsy	1894	2898	37	0.000
each other	one another	1552	2164	170	0.000
hug	embrace	2947	8201	326	0.000
loyal	faithful	1336	1410	644	0.000
real	genuine	6559	67748	307	0.000
sneakers	gym shoes, running shoes, tennis shoes	216	256	85	0.000
honest	truthful	2675	4724	51	0.000
rush	hurry	2874	4753	1867	0.000
ill	sick	7266	223879	5173	0.000
wrong	incorrect	3364	7136	62	0.000

little	small	5227	24025	3846	0.000
maybe	perhaps	3296	6423	178	0.000
mom	mother	5727	27826	5489	0.000
needed	required	2007	4526	445	0.000
prairie	plains	540	3896	476	0.000
student	pupil	1383	5573	34	0.000
fast	quick, rapid	4325	11958	7274	0.000
sad	unhappy	5000	23613	192	0.000
stomach	belly, tummy	1778	2110	1419	0.000
trash	garbage, rubbish	1248	1726	248	0.000
while	whilst	3950	12434	48	0.000
smart	intelligent	1521	2453	225	0.000
holiday	vacation	1542	1850	1339	0.000
island	isle	881	2261	1091	0.000
slim	slender	492	916	11	0.000
especially	particularly	1269	1816	38	0.000
obviously	clearly	1357	1141	777	0.000
rude	impolite	1262	1860	2	0.000
grandma	grandmother, granny, nana	2259	1739	2339	0.000
bathroom	restroom, washroom	1005	1151	443	0.000
garage sale	rummage sale, tag sale, yard sale	182	218	94	0.000
icing	frosting	579	899	62	0.000
grandpa	grandfather	860	1024	140	0.000
rare	scarce	691	1063	12	0.000
anywhere	anyplace	737	979	8	0.000
ping pong	table tennis	101	184	2	0.000
pharmacy	drug store	392	3243	5	0.000
sunset	sundown	941	7725	115	0.000
dawn	daybreak	340	523	92	0.000
bucket	pail	666	974	32	0.000
brag	boast	370	403	43	0.000
madness	insanity	612	780	185	0.000
false	untrue	336	512	12	0.000
expensive	costly	459	520	22	0.000
global	worldwide	460	1007	329	0.000
couch	sofa	810	891	400	0.000
spine	backbone	186	191	93	0.000
fridge	refrigerator	333	324	73	0.000
porch	veranda	340	526	36	0.000

hot tub	jacuzzi	159	154	40	0.000
sudden	abrupt	525	590	14	0.000
wallet	billfold	337	465	1	0.000
instantly	instantaneously	157	170	2	0.000
hallway	corridor	313	313	161	0.000
disappear	vanish	324	340	44	0.000
explode	blow up	358	218	181	0.000
bleach	clorox	209	241	6	0.000
bookstore	bookshop	90	153	14	0.000
polite	courteous	97	101	10	0.000
fatal	deadly, lethal	286	431	348	0.000
on accident	by accident	160	107	71	0.000
accomplishment	achievement	249	186	185	0.000
brave	courageous	356	480	68	0.000
except for	aside from	299	285	52	0.000
eggplant	aubergine	46	56	2	0.000
cut the grass	mow the grass, mow the lawn	28	18	10	0.000
out loud	aloud	278	284	55	0.000
cellar	basement	147	259	148	0.000
cinema	movie theater	397	1221	174	0.000
similar to	akin to	70	68	12	0.001
shant	shall not	120	82	60	0.001
quilt	comforter	94	181	33	0.001
inappropriate	improper	133	130	40	0.001
sunrise	sun up	485	3486	14	0.003
cemetery	graveyard	191	318	120	0.004
sufficient	adequate	81	56	33	0.008
inquire	enquire	28	49	2	0.028
jeep	suv	524	873	199	0.050
casket	coffin	92	70	60	0.058
thrive	flourish	131	224	57	0.067
fierce	ferocious	181	250	19	0.067
unbearable	insufferable	45	42	4	0.079
unexplainable	inexplicable	24	18	8	0.105
endurance	stamina	80	90	28	0.114
defy	disobey	50	48	9	0.166
dampen	moisten	8	8	1	0.183
passionate	impassioned	159	205	1	0.208
saggy	droopy	49	38	14	0.263
furthest	farthest	62	40	25	0.294
agree to	consent to	90	93	3	0.361

food processor	cuisinart	3	3	2	0.439
somewhere else	elsewhere	197	147	62	0.443
skillet	frying pan	65	93	6	0.493
mailman	postman	23	22	6	0.566
afire	ablaze, aflame	31	29	19	0.575
inadequate	insufficient	22	11	11	0.612
enclose	inclose	9	10	1	0.656
husk	shuck	253	330	129	0.662
ski doo	snowmobile	2	1	1	0.671
slow cooker	crock pot	19	16	8	0.745
flammable	inflammable	5	8	4	0.754
murderous	homicidal	11	6	5	0.760
entrust	intrust	19	14	9	0.799
unarm	disarm	33	47	3	0.857
shoelace	shoestring	21	16	8	0.884
water fountain	drinking fountain	22	23	4	0.890
incarcerate	imprison	17	9	8	0.908
leaned in	leaned forward	4	4	1	0.909

Appendix B. DAREDS Dialect-Specific Terms

In Table A2, we provide the list of dialect-specific terms used in our dialect prediction evaluation.

Table A2: Dialect specific terms from DAREDS used in our analysis. “Num VP” is the number of voting precincts the term appears in. “Total Freq” is the total frequency of the term.

DAREDS Dialect	Term	Num VP	Total Freq
Gulf States	aguardiente	1	1
Gulf States	bogue	1	1
Gulf States	cavalla	1	1
Gulf States	chinaberry	1	3
Gulf States	cooter	12	23
Gulf States	curd	17	18
Gulf States	doodlebug	1	1
Gulf States	jambalaya	27	27
Gulf States	loggerhead	1	3
Gulf States	maguey	4	5
Gulf States	nibbling	3	3
Gulf States	nig	72	76

Gulf States	pollywog	1	1
Gulf States	redfish	14	20
Gulf States	sardine	4	4
Gulf States	scratcher	8	8
Gulf States	shinny	3	4
Gulf States	squinch	1	1
Gulf States	whoop	488	588
Southwest	acequia	2	5
Southwest	agarita	1	1
Southwest	agave	38	72
Southwest	aguardiente	1	1
Southwest	alacran	1	1
Southwest	alberca	12	12
Southwest	albondigas	3	3
Southwest	alcalde	5	6
Southwest	alegria	20	21
Southwest	armas	8	16
Southwest	arriero	1	1
Southwest	arroba	1	1
Southwest	arrowwood	2	5
Southwest	atajo	1	1
Southwest	atole	7	7
Southwest	ayuntamiento	1	3
Southwest	azote	1	1
Southwest	baile	41	54
Southwest	bajada	1	30
Southwest	baldhead	2	2
Southwest	barranca	3	3
Southwest	basto	5	5
Southwest	beaner	31	32
Southwest	blinky	3	4
Southwest	booger	47	49
Southwest	burro	17	44
Southwest	caballo	12	13
Southwest	caliche	1	1
Southwest	camisa	16	16
Southwest	carcel	2	2
Southwest	carga	7	39

Southwest	cargador	8	9
Southwest	carreta	5	6
Southwest	cenizo	2	2
Southwest	chalupa	17	17
Southwest	chaparreras	1	1
Southwest	chapo	47	67
Southwest	chaqueta	2	2
Southwest	charco	7	8
Southwest	charro	27	39
Southwest	chicalote	1	1
Southwest	chicharron	4	4
Southwest	chiquito	20	25
Southwest	cholo	39	40
Southwest	cienaga	1	1
Southwest	cocinero	1	1
Southwest	colear	1	1
Southwest	comadre	11	12
Southwest	comal	31	124
Southwest	compadre	37	97
Southwest	concha	15	18
Southwest	conducta	4	4
Southwest	cowhand	2	2
Southwest	cuidado	25	29
Southwest	cuna	4	5
Southwest	dinero	75	84
Southwest	dueno	2	2
Southwest	enchilada	39	47
Southwest	encinal	4	9
Southwest	estufa	1	1
Southwest	fierro	16	77
Southwest	freno	5	5
Southwest	frijole	2	2
Southwest	garbanzo	5	9
Southwest	goober	26	29
Southwest	gotch	6	6
Southwest	greaser	3	3
Southwest	grulla	5	8
Southwest	jacal	2	3

Southwest	junco	2	3
Southwest	kiva	9	25
Southwest	lechuguilla	1	1
Southwest	loafer	4	4
Southwest	maguey	4	5
Southwest	malpais	1	2
Southwest	menudo	94	107
Southwest	mescal	1	1
Southwest	mestizo	3	8
Southwest	milpa	2	3
Southwest	nogal	4	5
Southwest	nopal	8	9
Southwest	olla	6	9
Southwest	paisano	14	73
Southwest	pasear	7	8
Southwest	pelado	1	1
Southwest	peon	17	17
Southwest	picacho	2	11
Southwest	pinole	2	2
Southwest	plait	2	2
Southwest	potrero	4	4
Southwest	potro	6	12
Southwest	pozo	3	4
Southwest	pulque	2	2
Southwest	quelite	1	1
Southwest	ranchero	14	19
Southwest	reata	6	28
Southwest	runaround	3	3
Southwest	seesaw	3	3
Southwest	serape	6	12
Southwest	shorthorn	1	1
Southwest	slouch	2	2
Southwest	tamale	47	64
Southwest	tinaja	2	2
Southwest	tomatillo	5	21
Southwest	tostada	16	23
Southwest	tule	3	6
Southwest	vaquero	19	37

Southwest	vara	2	2
Southwest	wetback	18	18
Southwest	zaguan	1	3
Texas	agarita	1	1
Texas	banquette	3	3
Texas	blackland	3	4
Texas	bluebell	14	15
Texas	borrego	10	17
Texas	cabrito	5	27
Texas	caliche	1	1
Texas	camote	1	1
Texas	cenizo	2	2
Texas	cerillo	1	1
Texas	chicharra	1	1
Texas	coonass	3	3
Texas	ducking	66	68
Texas	firewheel	19	114
Texas	foxglove	3	3
Texas	goatsbeard	1	2
Texas	granjeno	1	3
Texas	grulla	5	8
Texas	guayacan	2	3
Texas	hardhead	1	1
Texas	huisache	4	7
Texas	icehouse	46	132
Texas	juneteenth	12	16
Texas	kinfolk	88	96
Texas	lechuguilla	1	1
Texas	mayapple	1	1
Texas	mayberry	8	8
Texas	norther	3	3
Texas	piloncillo	1	1
Texas	pinchers	1	1
Texas	piojo	18	20
Texas	praline	14	17
Texas	priss	5	5
Texas	redhorse	1	1
Texas	resaca	5	5

Texas	retama	11	31
Texas	sabino	2	2
Texas	scissortail	1	3
Texas	sendero	9	26
Texas	shallot	1	1
Texas	sharpshooter	3	3
Texas	sook	1	1
Texas	sotol	6	28
Texas	spaniard	2	2
Texas	squinch	1	1
Texas	tecolote	2	6
Texas	trembles	1	1
Texas	tush	4	4
Texas	vamos	392	580
Texas	vaquero	19	37
Texas	vara	2	2
Texas	washateria	16	24
Texas	wetback	18	18
West	arbuckle	8	25
West	barefooted	2	2
West	barf	44	47
West	bawl	10	10
West	biddy	3	6
West	blab	3	3
West	blat	3	3
West	boudin	29	36
West	breezeway	6	10
West	buckaroo	9	10
West	bucking	19	21
West	bunkhouse	4	5
West	caballo	12	13
West	cabeza	70	74
West	cack	4	4
West	calaboose	1	2
West	capper	2	2
West	chapping	1	1
West	chileno	1	1
West	chippy	7	12

West	clabber	1	1
West	clunk	1	1
West	cribbage	1	1
West	cutback	1	1
West	dally	3	3
West	dogger	2	3
West	entryway	7	8
West	freighter	1	1
West	frenchy	4	5
West	gaff	2	7
West	gesundheit	1	1
West	glowworm	1	1
West	goop	5	5
West	grayback	1	2
West	groomsman	1	2
West	hackamore	1	2
West	hardhead	1	1
West	hardtail	2	5
West	headcheese	1	1
West	heave	3	3
West	heinie	1	1
West	highline	4	8
West	hoodoo	1	2
West	husk	1	1
West	irrigate	1	1
West	jibe	4	5
West	jimmies	4	8
West	kaput	1	1
West	kike	15	16
West	latigo	3	4
West	lockup	3	4
West	longear	1	1
West	lunger	1	1
West	maguey	4	5
West	makings	7	30
West	manzanita	5	6
West	mayapple	1	1
West	mochila	4	4

West	nester	1	1
West	nighthawk	6	10
West	paintbrush	19	29
West	partida	5	5
West	peddle	3	3
West	peeler	1	1
West	pincushion	3	6
West	pith	1	1
West	plastered	9	9
West	podunk	2	2
West	pollywog	1	1
West	prat	1	1
West	puncher	5	5
West	riffle	1	1
West	ringy	1	1
West	rustle	1	1
West	rustler	3	4
West	seep	4	4
West	serape	6	12
West	sinker	11	15
West	sizzler	5	5
West	snoozer	1	1
West	snuffy	2	2
West	sprangletop	1	1
West	sunfish	1	1
West	superhighway	1	1
West	swamper	2	4
West	tallboy	2	2
West	tamarack	2	3
West	tenderfoot	2	4
West	tennie	1	1
West	tumbleweed	11	37
West	vamos	392	580
West	waddy	2	2
West	waken	9	9
West	washateria	16	24
West	weedy	1	1
West	wienie	4	4

West	wrangle	4	5
West	zori	1	1

Appendix C. Han and Baldwin (2011) Lexical Variants

Table A3: Lexical variants from Han and Baldwin (2011) used in our lexical variant evaluation. “Canonical” is the canonical form as identified by annotators and “Variant” is the non-standard variant. “Var VP” and “Var Freq” are the number of voting precincts that contain the variant and the total frequency. “Can VP” and “Can Freq” are similar for the Canonical form.

Variant	Canonical	Var VP	Var Freq	Can VP	Can Freq	Shared VP
ahh	ah	1009	1319	1162	1800	1839
bb	baby	665	861	4828	17472	4908
bc	because	2808	6220	4802	17280	5276
bday	birthday	1281	2033	4650	19210	4814
bf	boyfriend	974	1194	2172	3398	2653
bro	brother	3735	12036	2747	5263	4535
bs	bullshit	953	1308	1395	1952	2016
btw	between	686	862	1890	6710	2288
chillin	chilling	1174	1653	888	1185	1773
comin	coming	563	681	3612	10765	3737
congrats	congratulations	1542	2945	881	1765	2002
convo	conversation	521	586	960	1259	1336
cus	because	541	675	4802	17280	4876
cuz	because	2288	3959	4802	17280	5162
da	the	2326	5497	7669	598549	7670
dat	that	1648	2900	7134	142061	7145
dawg	dog	806	1240	2356	5337	2750
de	the	3267	21053	7669	598549	7692
def	definitely	617	2575	1832	3224	2141
doin	doing	941	1272	4153	11681	4334
fam	family	2040	3921	3862	12856	4376
fb	facebook	1127	1637	1246	1962	2037
freakin	freaking	554	654	1555	2157	1884
fuckin	fucking	1891	3064	4209	12868	4547
gettin	getting	1380	1992	5066	21187	5226
gf	girlfriend	772	942	1474	2087	1959
goin	going	1446	2089	5881	33556	5949

gon	gonna	1227	1914	5327	22704	5449
hahah	haha	901	1104	4667	15314	4793
hahaha	haha	2597	4730	4667	15314	5097
hahahaha	haha	1201	1595	4667	15314	4821
hrs	hours	739	1393	3043	8568	3284
jus	just	1011	1537	7074	131656	7082
kno	know	929	1377	6425	55510	6453
lawd	lord	510	634	1938	3244	2185
lil	little	2990	7405	4913	21558	5435
lookin	looking	1134	1534	4499	55830	4690
mins	minutes	1583	14602	2352	5244	3164
mis	miss	561	948	5103	19099	5171
nah	no	2882	5869	6526	66786	6604
naw	no	882	1234	6526	66786	6539
nd	and	1972	4823	7449	349628	7455
nothin	nothing	692	839	4074	10591	4213
ohh	oh	736	869	5264	20804	5343
pic	picture	2675	6195	2981	6474	4066
pics	pictures	1521	2483	2123	3707	2881
playin	playing	585	679	3163	7102	3350
pls	please	1107	1635	4164	12972	4388
plz	please	840	1313	4164	12972	4340
ppl	people	2164	3896	5882	34714	6020
prolly	probably	709	847	2968	5624	3242
sayin	saying	626	744	2831	5194	3055
soo	so	1467	2019	7105	123174	7117
talkin	talking	1029	1385	3790	9014	4027
tha	the	1394	2630	7669	598549	7672
tht	that	531	738	7134	142061	7135
thx	thanks	713	1031	4707	19000	4791
til	till	1401	2279	2887	5588	3435
til	until	1401	2279	3842	11761	4301
txt	text	713	886	4102	10789	4229
umm	um	555	625	826	1090	1265
ur	your	2810	5917	6729	83776	6794
wat	what	983	1318	6617	67576	6634
yess	yes	576	665	4924	18365	4997
yr	year	566	809	4530	16848	4614
yu	you	1082	2144	7550	476752	7551

Appendix D. Liu et al. (2011) Lexical Variants

Table A4: Lexical variants from Liu et al. (2011) used in our lexical variant evaluation. “Canonical” is the canonical form as identified by annotators and “Variant” is the non-standard variant. “Var VP” and “Var Freq” are the number of voting precincts that contain the variant and the total frequency. “Can VP” and “Can Freq” are similar for the Canonical form.

Variant	Canonical	Var VP	Var Freq	Can VP	Can Freq	Shared VP
aye	yes	1055	1409	4924	18365	5037
b	be	2915	8312	7081	212570	7108
bae	baby	3001	6203	4828	17472	5312
bb	baby	665	861	4828	17472	4908
bby	baby	814	958	4828	17472	4949
bc	because	2808	6220	4802	17280	5276
bday	birthday	1281	2033	4650	19210	4814
bout	about	3295	8238	6463	94613	6594
bro	brother	3735	12036	2747	5263	4535
bros	brothers	635	1066	1145	1899	1561
bs	bullshit	953	1308	1395	1952	2016
butt	but	1312	1846	6808	86579	6825
c	see	2332	7926	6259	132803	6358
cause	because	4439	13497	4802	17280	5735
chillin	chilling	1174	1653	888	1185	1773
comin	coming	563	681	3612	10765	3737
convo	conversation	521	586	960	1259	1336
cus	because	541	675	4802	17280	4876
cutie	cute	692	880	3951	10397	4073
cuz	because	2288	3959	4802	17280	5162
da	the	2326	5497	7669	598549	7670
dat	that	1648	2900	7134	142061	7145
def	definitely	617	2575	1832	3224	2141
dem	them	556	767	5320	23430	5361
dis	this	891	1269	7247	392504	7249
doin	doing	941	1272	4153	11681	4334
em	them	2585	5577	5320	23430	5578
fa	for	607	942	7429	438864	7431
fam	family	2040	3921	3862	12856	4376
fav	favorite	1422	2199	3531	10655	3920
fb	facebook	1127	1637	1246	1962	2037
feelin	feeling	753	950	3300	7215	3511

fml	family	750	898	3862	12856	4053
fr	for	1059	1672	7429	438864	7436
freakin	freaking	554	654	1555	2157	1884
ft	feet	1273	11113	1303	1916	2173
fuckin	fucking	1891	3064	4209	12868	4547
gettin	getting	1380	1992	5066	21187	5226
gf	girlfriend	772	942	1474	2087	1959
goin	going	1446	2089	5881	33556	5949
gon	going	1227	1914	5881	33556	5936
homie	home	1343	2249	5314	27569	5442
hr	hour	852	2624	2404	5606	2838
hrs	hours	739	1393	3043	8568	3284
ii	i	770	9871	7699	621319	7699
jus	just	1011	1537	7074	131656	7082
k	ok	3145	7414	3940	71563	4824
kno	know	929	1377	6425	55510	6453
lawd	lord	510	634	1938	3244	2185
lil	little	2990	7405	4913	21558	5435
lookin	looking	1134	1534	4499	55830	4690
luv	love	1030	1390	6698	76733	6714
m	am	2507	7994	5176	25099	5507
ma	my	783	1231	7512	309237	7512
mi	my	2204	6510	7512	309237	7551
min	minutes	1203	2314	2352	5244	2941
mines	mine	510	589	2755	5078	2968
mins	minutes	1583	14602	2352	5244	3164
mo	more	585	20581	5669	31459	5706
n	and	3408	17544	7449	349628	7478
nada	nothing	508	712	4074	10591	4187
nah	no	2882	5869	6526	66786	6604
naw	no	882	1234	6526	66786	6539
nd	and	1972	4823	7449	349628	7455
nothin	nothing	692	839	4074	10591	4213
nun	nothing	622	788	4074	10591	4195
ohh	oh	736	869	5264	20804	5343
pic	picture	2675	6195	2981	6474	4066
pics	pictures	1521	2483	2123	3707	2881
playin	playing	585	679	3163	7102	3350
pls	please	1107	1635	4164	12972	4388

plz	please	840	1313	4164	12972	4340
ppl	people	2164	3896	5882	34714	6020
prolly	probably	709	847	2968	5624	3242
pt	part	570	2138	2647	11220	2823
r	are	2280	5466	6657	76873	6712
rd	road	2123	15149	2022	5075	3220
sayin	saying	626	744	2831	5194	3055
sis	sister	857	1219	2714	5257	3022
soo	so	1467	2019	7105	123174	7117
sum	some	990	1541	6017	42637	6052
talkin	talking	1029	1385	3790	9014	4027
th	the	3238	17089	7669	598549	7672
tha	the	1394	2630	7669	598549	7672
thang	thing	691	876	4434	12995	4550
tho	though	3959	11480	3879	9628	5092
thot	thought	607	791	3690	8510	3844
thru	through	1406	2281	3400	8800	3818
tht	that	531	738	7134	142061	7135
thx	thanks	713	1031	4707	19000	4791
til	till	1401	2279	2887	5588	3435
trippin	tripping	790	975	558	669	1204
turnt	turn	684	836	2918	5943	3161
tx	texas	6275	456640	4983	96986	6869
txt	text	713	886	4102	10789	4229
u	you	5375	34958	7550	476752	7578
ur	your	2810	5917	6729	83776	6794
w	with	4195	28363	7043	146575	7124
wat	what	983	1318	6617	67576	6634
wen	when	524	653	6637	67470	6650
wit	with	1769	3389	7043	146575	7054
wut	what	582	724	6617	67576	6627
y	why	3107	11552	5974	36088	6182
ya	you	4484	15215	7550	476752	7563
yea	yeah	2418	4617	4499	13843	4938
yess	yes	576	665	4924	18365	4997
yo	you	3677	10918	7550	476752	7559
yr	year	566	809	4530	16848	4614
yu	you	1082	2144	7550	476752	7551
yup	yes	1056	1499	4924	18365	5040

Acknowledgments

The authors thank Axel Bohmann, Katrin Erk, John Beavers, Danny Law, Ray Mooney, and Jessy Li for their helpful discussions. The authors also thank the Texas Advanced Computing Center for the computer resources provided.

References

Archive Team. 1996. The twitter stream grab.

Atwood, E. Bagby. 1962. The Regional Vocabulary of Texas. University of Texas Press. https://doi.org/10.7560/733497

Baas, Kevin. n.d. Auto-redistrict. http://autoredistrict.org/.

Bailey, Guy and Margie Dyer. 1992. An approach to sampling in dialectology. *American Speech*, 67(1):3–20. https://doi.org/10.2307/455756

Bailey, Guy and Natalie Maynor. 1985. The present tense of be in southern black folk speech. *American Speech*, 60(3):195–213. https://doi.org/10.2307/454884

Bailey, Guy and Natalie Maynor. 1987. Decreolization? *Language in Society*, 16(4):449–473. https://doi.org/10.1017/S0047404500000324

Bailey, Guy and Natalie Maynor. 1989. The divergence controversy. *American Speech*, 64(1):12–39. https://doi.org/10.2307/455110

Bailey, Guy and Erik Thomas. 2021. Some aspects of african-american vernacular english phonology. In *African-American English*. Routledge, pages 93–118. https://doi.org/10.4324/9781003165330-5

Bailey, Guy, Tom Wikle, and Lori Sand. 1991. The focus of linguistic innovation in Texas. *English World-Wide*, 12(2):195–214. https://doi.org/10.1075/eww.12.2.03bai

Bailey, Guy, Tom Wikle, Jan Tillery, and Lori Sand. 1991. The apparent time construct. *Language Variation and Change*, 3(3):241–264. https://doi.org/10.1017/S0954394500000569

Bayley, Robert. 1994. *Consonant Cluster Reduction in Tejano English*, volume 6. Cambridge University Press. https://doi.org/10.1017/S0954394500001708

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754. https://doi.org/10.18653/v1/S17-2126

Bernstein, Cynthia. 1993. Measuring social causes of phonological variation in Texas. *American Speech*, 68(3):227–240. https://doi.org/10.2307/455631

Bohmann, Axel. 2020. Situating twitter discourse in relation to spoken and written texts: A lectometric analysis. *Zeitschrift für Dialektologie und Linguistik*, 87(2):250–284. https://doi.org/10.25162/zdl-2020-0009

Campbell-Kibler, Kathryn. 2005. *Listener Perceptions of Sociolinguistic Variables: The Case of (ING)*. Ph.D. thesis, Stanford University.

Carver, Craig M. 1987. *American Regional Dialects: A Word Geography*. University of Michigan Press. https://doi.org/10.3998/mpub.12484

Cassidy, Frederic G., Joan Houston Hall, and Luanne Von Schneidemesser. 1985. *Dictionary of American Regional English*, volume 1. Belknap Press of Harvard University.

Cook, Paul, Bo Han, and Timothy Baldwin. 2014. Statistical methods for identifying local dialectal terms from gps-tagged documents. *Dictionaries: Journal of the Dictionary Society of North America*, 35(35):248–271. https://doi.org/10.1353/dic.2014.0020

Di Paolo, Marianna. 1989. Double modals as single lexical items. *American Speech*, 64(3):195–224. https://doi.org/10.2307/455589

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106. https://doi.org/10.3115/v1/E14-1011

Duggan, Maeve. 2015. Mobile Messaging and Social Media 2015. Pew Research Center. https://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PloS ONE*, 9(11):e113114. https://doi.org/10.1371/journal.pone.0113114, PubMed: 25409166

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2012. Mapping the geographical diffusion of new words. In *Proceedings of the NIPS Workshop on Social Network and Social Media Analysis: Methods, Models and Applications*, page 13.

Eisenstein, Jacob, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1365–1374.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615. https://doi.org/10.3115/v1/N15-1184

Firth, David. 1993. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38. https://doi.org/10.1093/biomet/80.1.27

Galindo, D. Letticia. 1988. Towards a description of Chicano English: A sociolinguistic perspective. In *Linguistic Change and Contact (Proceedings of the 16th Annual Conference on New Ways of Analyzing Variation in Language)*, pages 113–23. Department of Linguistics, University of Texas at Austin.

Garcia, Juliet Villarreal. 1976. *The Regional Vocabulary of Brownsville, Texas*. The University of Texas at Austin.

Gillies, Sean, et al. 2007. Shapely: Manipulation and analysis of geometric objects in the cartesian plane. URL: https://pypi.org/project/Shapely/.

Goel, Rahul, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, pages 41–57. https://doi.org/10.1007/978-3-319-47880-7_3

Grier, D. G., Alexander Thompson, A. Kwasniewska, G. J. McGonigle, H. L. Halliday, and T. R. Lappin. 2005. The pathophysiology of HOX genes and their role in cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 205(2):154–171. https://doi.org/10.1002/path.1710, PubMed: 15643670

Grieve, Jack and Costanza Asnaghi. 2013. A lexical dialect survey of American English using site-restricted web searches. In *American Dialect Society Annual Meeting, Boston*, pages 3–5.

Grieve, Jack, Costanza Asnaghi, and Tom Ruette. 2013. Site-restricted web searches for data collection in regional dialectology. *American Speech*, 88(4):413–440. https://doi.org/10.1215/00031283-2691424

Grieve, Jack, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics*, 46(4):293–319. https://doi.org/10.1177/0075424218793191

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(2):193–221. https://doi.org/10.1017/S095439451100007X

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, pages 2116–2121. https://doi.org/10.18653/v1/D16-1229, PubMed: 28580459

Han, Bo and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378.

Heinze, Georg and Michael Schemper. 2002. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419. https://doi.org/10.1002/sim.1047, PubMed: 12210625

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 Tommaso Fornaciari. 2018. Increasing in-class similarity by retrofitting embeddings with demographic information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 671–677. https://doi.org/10.18653/v1/D18-1070

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*, pages 4383–4394. https://doi.org/10.18653/v1/D18-1469

Hovy, Dirk, Afshin Rahimi, Timothy Baldwin, and Julian Brooke. 2020.

Visualizing regional language variation across Europe on twitter. *Handbook of the Changing World Language Map*, pages 3719–3742. https://doi.org/10.1007/978-3-030-02438-3_175

Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding us regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Jones, Taylor. 2015. Toward a description of African American vernacular English dialect regions using "Black twitter". *American Speech*, 90(4):403–440. https://doi.org/10.1215/00031283-3442117

Koops, Christian. 2010. /u/-fronting is not monolithic: Two types of fronted/u/in Houston Anglos. *University of Pennsylvania Working Papers in Linguistics*, 16(2):14.

Koops, Christian, Elizabeth Gentry, and Andrew Pantos. 2008. The effect of perceived speaker age on the perception of pin and pen vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, 14(2):12.

Kosmidis, Ioannis. 2020. brglm2: Bias reduction in generalized linear models. *R package version 0.6*, 2:635.

Kosmidis, Ioannis and David Firth. 2009. Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4):793–804. https://doi.org/10.1093/biomet/asp055

Kulkarni, Vivek, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 615–618. https://doi.org/10.1609/icwsm.v10i1.14798

Labov, William, Sharon Ash, Charles Boberg, et al. 2006. *The Atlas of North American English: Phonetics, Phonology, and Sound Change: a Multimedia Reference Tool*, volume 1. Walter de Gruyter. https://doi.org/10.1515/9783110167467

Lameli, Alfred. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*, volume 54. Walter de Gruyter. https://doi.org/10.1515/9783110331394

Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Liu, Fei, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76.

Mansournia, Mohammad Ali, Angelika Geroldinger, Sander Greenland, and Georg Heinze. 2018. Separation in logistic regression: Causes, consequences, and control. *American Journal of Epidemiology*, 187(4):864–870. https://doi.org/10.1093/aje/kwx299, PubMed: 29020135

McDowell, John and Susan McRae. 1972. Differential response of the class and ethnic components of the austin speech community to marked phonological variables. *Anthropological Linguistics*, pages 228–239.

McFadden, Daniel. 1977. Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. Cowles Foundation Discussion Papers 474, Cowles Foundation for Research in Economics, Yale University.

McFadden, Daniel. 1973. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*. Academic Press, pp. 105–142.

Mencarini, Letizia. 2018. The potential of the computational linguistic analysis of social media for population studies. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 62–68. https://doi.org/10.18653/v1/W18-1109

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Moran, Patrick A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23. https://doi.org/10.1093/biomet/37.1-2.17, PubMed: 15420245

Murray, Ryan and Ben Tengelsen. 2018. Optimal districts. https://github.com/btengels/optimaldistricts.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593. https://doi.org/10.1162/COLI_a_00258

Nguyen, Dong and Jack Grieve. 2020. Do word embeddings capture spelling variation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 870–881. https://doi.org/10.18653/v1/2020.coling-main.75

Pederson, Lee. 1986. *Linguistic Atlas of the Gulf States*, volume 2. University of Georgia Press.

Petyt, Keith Malcolm. 1980. *The Study of Dialect: An Introduction to Dialectology*. Westview Press.

Pröll, Simon. 2013. Detecting structures in linguistic maps—fuzzy clustering for pattern recognition in geostatistical dialectometry. *Literary and Linguistic Computing*, 28(1):108–118. https://doi.org/10.1093/llc/fqs059

Rahimi, Afshin, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 209–216. https://doi.org/10.18653/v1/P17-2033

Řehůřek, Radim and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. http://is.muni.cz/publication/884893/en.

Rosenfeld, Alex and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 474–484. https://doi.org/10.18653/v1/N18-1044

Stone, Mervyn. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x

Tarpley, Fred. 1970. *From Blinky to Blue-John: A Word Atlas of Northeast Texas*. University Press.

Thomas, Erik R. 1997. A rural/metropolitan split in the speech of Texas Anglos. *Language Variation and Change*, 9(3):309–332. https://doi.org/10.1017/S0954394500001940

U.S. Election Assistance Commission. 2017. EAVS deep dive: Poll workers and polling places. https://www.eac.gov/sites/default/files/document_library/files/EAVSDeepDive_pollworkers_pollingplaces_nov17.pdf.

Van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605.

Walsh, Harry and Victor L. Mote. 1974. A Texas dialect feature: Origins and distribution. *American Speech*, 49(1/2):40–53. https://doi.org/10.2307/3087917

Wheatley, Katherine E. and Oma Stanley. 1959. Three generations of East Texas speech. *American Speech*, 34(2):83–94. https://doi.org/10.2307/454372

Widawski, Maciej. 2015. *African American slang: A Linguistic Description*. Cambridge University Press. https://doi.org/10.1017/CBO9781139696562

Xiong, Yijin, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. Fusing label embedding into bert: An efficient improvement for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750. https://doi.org/10.18653/v1/2021.findings-acl.152