

# Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media

Vivek Kulkarni and Bryan Perozzi and Steven Skiena

Stony Brook University

Department of Computer Science, USA

{vvkulkarni,bperozzi,skiena@cs.stonybrook.edu}

## Abstract

In this paper we present a new computational technique to detect and analyze statistically significant geographic variation in language. While previous approaches have primarily focused on lexical variation between regions, our method identifies words that demonstrate semantic and syntactic variation as well.

We extend recently developed techniques for neural language models to learn word representations which capture differing semantics across geographical regions. In order to quantify this variation and ensure robust detection of true regional differences, we formulate a null model to determine whether observed changes are statistically significant. Our method is the first such approach to explicitly account for random variation due to chance while detecting regional variation in word meaning.

To validate our model, we study and analyze two different massive online data sets: millions of tweets from Twitter as well as millions of phrases contained in the Google Book Ngrams. Our analysis reveals interesting facets of language change across countries.

## 1 Introduction

Detecting and analyzing regional variation in language is central to the field of socio-variational linguistics and dialectology (Tagliamonte 2006; Labov 1980; Milroy 1992). Since online content is an agglomeration of material originating from all over the world, language on the Internet demonstrates geographic variation. The abundance of geo-tagged online text enables a study of geographic linguistic variation at scales that are unattainable using classical methods like surveys and questionnaires.

Characterizing and detecting such variation is challenging since it takes different forms: lexical, syntactic and semantic. Most existing work has focused on detecting lexical variation prevalent in geographic regions (Bamman, Eisenstein, and Schnoebelen 2014; Doyle 2014; Eisenstein et al. 2010; 2014). However, regional linguistic variation is not limited to lexical variation.

In this paper we address this gap. Our method, GEODIST, is the first computational approach for tracking and detecting

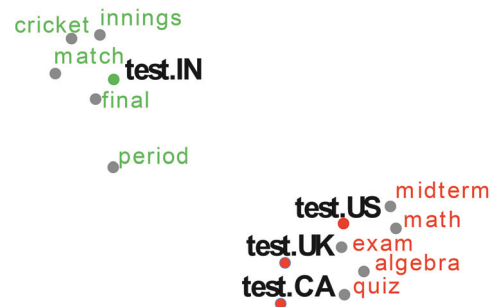


Figure 1: The latent semantic space captured by our method (GEODIST) reveals geographic variation between language speakers. In the majority of the English speaking world (e.g. US, UK, and Canada) a `test` is primarily used to refer to an exam, while in India a `test` indicates a cricket match played over five days.

statistically significant linguistic shifts of words across geographical regions. GEODIST detects syntactic and semantic variation in word usage across regions, in addition to purely lexical differences. GEODIST builds on recently introduced neural language models that learn *word embeddings*, extending them to capture region-specific semantics (see Figure 1 for a visualization of the semantic variation captured by GEODIST). Since observed regional variation could be due to chance, GEODIST *explicitly* introduces a null model to ensure detection of only statistically significant differences between regions.

One might argue that simple baseline methods like (analyzing part of speech or frequency) might be sufficient to identify regional variation. However these methods capture different modalities, and therefore detect different types of changes (restricted to lexical or syntactic changes).

We use our method to investigate linguistic variation across Twitter between four English speaking countries and investigate regional variation in the Google Books Ngram Corpus data.<sup>1</sup> Our methods detect a variety of changes including regional dialectal variations, region specific usages, words incorporated due to code mixing and differing semantics.

<sup>1</sup>An extended analysis and full set of results are discussed in our preprint (Kulkarni, Perozzi, and Skiena 2015)

## 2 Method: GEODIST

As we remarked in the previous section, linguistic variation is not restricted only to lexical or syntactic variation. In order to detect subtle semantic changes, we need to infer cues based on the contextual usage of a word. To do so, we use distributional methods which learn a latent semantic space that maps each word  $w \in \mathcal{V}$  to a continuous vector space  $\mathbb{R}^d$ .

We differentiate ourselves from the closest related work to our method (Bamman and others 2014), by *explicitly* accounting for random variation between regions, and proposing a method to detect statistically significant changes.

**Learning region specific word embeddings** Given a corpus  $\mathcal{C}$  with  $R$  regions, we seek to learn a region specific word embedding  $\phi_r : \mathcal{V}, \mathcal{C}_r \mapsto \mathbb{R}^d$  using a neural language model. For each word  $w \in \mathcal{V}$  the neural language model learns:

1. A global embedding  $\delta_{\text{MAIN}}(w)$  for the word ignoring all region specific cues.
2. A differential embedding  $\delta_r(w)$  that encodes differences from the global embedding specific to region  $r$ .

The region specific embedding  $\phi_r(w)$  is computed as:  $\phi_r(w) = \delta_{\text{MAIN}}(w) + \delta_r(w)$ . Before training, the global word embeddings are randomly initialized while the differential word embeddings are initialized to 0. During each training step, the model is presented with a set of words  $w$  and the region  $r$  they are drawn from. Given a word  $w_i$ , the context words are the words appearing to the left or right of  $w_i$  within a window of size  $m$ . We define the set of active regions  $\mathcal{A} = \{r, \text{MAIN}\}$  where MAIN is a placeholder location corresponding to the global embedding and is always included in the set of active regions. The training objective then is to maximize the probability of words appearing in the context of word  $w_i$  conditioned on the active set of regions  $\mathcal{A}$ . Specifically, we model the probability of a context word  $w_j$  given  $w_i$  as:

$$\Pr(w_j | w_i) = \frac{\exp(\mathbf{w}_j^T \mathbf{w}_i)}{\sum_{w_k \in \mathcal{V}} \exp(\mathbf{w}_k^T \mathbf{w}_i)} \quad (1)$$

where  $\mathbf{w}_i$  is defined as  $\mathbf{w}_i = \sum_{a \in \mathcal{A}} \delta_a(w_i)$ .

During training, we iterate over each word occurrence in  $\mathcal{C}$  to minimize the negative log-likelihood of the context words. Our objective function  $J$  is thus given by:

$$J = \sum_{w_i \in \mathcal{C}} \sum_{\substack{j=i-m \\ j! = i}}^{i+m} -\log \Pr(w_j | \mathbf{w}_i) \quad (2)$$

We optimize the model parameters using stochastic gradient descent, as  $\phi_t(w_i) = \phi_t(w_i) - \alpha \times \frac{\partial J}{\partial \phi_t(w_i)}$  where  $\alpha$  is the learning rate. We compute the derivatives using the back-propagation algorithm. We set  $\alpha = 0.025$ , context window size  $m$  to 10 and size of the word embedding  $d$  to be 200 unless stated otherwise.

**Distance Computation between regional embeddings** After learning word embeddings for each word  $w \in \mathcal{V}$ , we then compute the distance

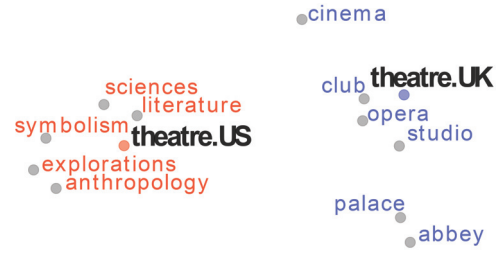


Figure 2: Semantic field of `theatre` as captured by GEODIST method between the UK and US. `theatre` is a field of study in the US while in the UK it primarily associated with opera or a club.

of a word between any two regions  $(r_i, r_j)$  as  $\text{SCORE}(w) = \text{COSINEDISTANCE}(\phi_{r_i}(w), \phi_{r_j}(w))$  where  $\text{COSINEDISTANCE}(u, v)$  is defined by  $1 - \frac{u^T v}{\|u\|_2 \|v\|_2}$ .

Figure 2 illustrates the information captured by our GEODIST method as a two dimensional projection of the latent semantic space learned, for the word `theatre`. In the US, the British spelling `theatre` is typically used only to refer to the performing arts. Observe how the word `theatre` in the US is close to other subjects of study: `sciences`, `literature`, `anthropology`, but `theatre` as used in UK is close to places showcasing performances (like `opera`, `studio`, etc). We emphasize that these regional differences detected by GEODIST are inherently *semantic*, the result of a level of language understanding unattainable by methods which focus solely on lexical variation (Eisenstein, Smith, and Xing 2011).

**Statistical Significance of Changes** In this section, we outline our method to quantify whether an observed change given by  $\text{SCORE}(w)$  is significant.

Since in our method,  $\text{SCORE}(w)$  could vary due random stochastic processes (even possibly pure chance), whether an observed score is significant or not depends on two factors: (a) the magnitude of the observed score (*effect size*) and (b) probability of obtaining a score more extreme than the observed score, even in the absence of a true effect.

First our method explicitly models the scenario when there is no effect, which we term as the *null model*. Next we characterize the distribution of scores under the null model. Our method then compares the observed score with this distribution of scores to ascertain the significance of the observed score. Our method is described succinctly in Algorithm 1. We deem a change observed for  $w$  as statistically significant when (a) The effect size exceeds a threshold  $\beta$  (set to the 95th percentile) which ensures the effect size is large enough and (b) It is rare to observe this effect as a result of pure chance which we capture using the confidence intervals computed.

Figure 3 illustrates this for two words: `hand` and `buffalo`. Observe that for `hand`, the observed score is smaller than the higher confidence interval, indicating that `hand` has not changed significantly. In contrast `buffalo` which is used differently in New York (since `buffalo` refers

---

**Algorithm 1** SCORESIGNIFICANCE ( $C, B, \alpha$ )

---

**Input:**  $C$ : Corpus of text with  $R$  regions,  $B$ : Number of bootstrap samples,  $\alpha$ : Confidence Interval threshold  
**Output:**  $E$ : Computed effect sizes for each word  $w$ , CI: Computed confidence intervals for each word  $w$   
// Estimate the NULL distribution.  
1:  $BS \leftarrow \emptyset$  {Corpora from the NULL Distribution}.  
NULLSCORES( $w$ ) {Store the scores for  $w$  under null model.}  
2: **repeat**  
3: Permute the labels assigned to text of  $C$  uniformly at random to obtain corpus  $C'$   
4:  $BS \leftarrow BS \cup C'$   
5: Learn a model  $N$  using  $C'$  as the text.  
6: **for**  $w \in \mathcal{V}$  **do**  
7: Compute SCORE( $w$ ) using  $N$ .  
8: Append SCORE( $w$ ) to NULLSCORES( $w$ )  
9: **end for**  
10: **until**  $|BS| = B$   
// Estimate the actual observed effect and compute confidence intervals.  
11: Learn a model  $M$  using  $C$  as the text.  
12: **for**  $w \in \mathcal{V}$  **do**  
13: Compute SCORE( $w$ ) using  $M$ .  
14:  $E(w) \leftarrow \text{SCORE}(w)$   
15: Sort the scores in NULLSCORES( $w$ ).  
16:  $\text{HCI}(w) \leftarrow 100\alpha$  percentile in NULLSCORES( $w$ )  
17:  $\text{LCI}(w) \leftarrow 100(1 - \alpha)$  percentile in NULLSCORES( $w$ )  
18:  $\text{CI}(w) \leftarrow (\text{LCI}(w), \text{HCI}(w))$   
19: **end for**  
20: **return**  $E, \text{CI}$

---

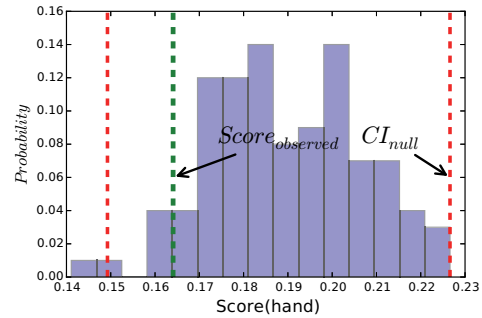
to a place in New York) has a score well above the higher confidence interval under the null model. Incorporating the null model and obtaining confidence estimates enables our method to efficaciously tease out effects arising due to random chance from statistically significant effects.

### 3 Results and Analysis

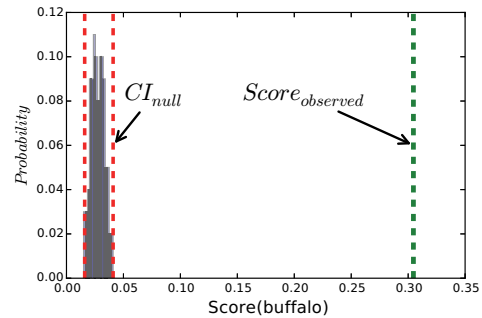
We use a random sample of 30 million ngrams for *American English* and *British English* from the Google Book Ngrams corpus (Michel and others 2011). In Table 1 we show several words identified by our GEODIST method. While *theatre* refers primarily to a building (where events are held) in the UK, in the US *theatre* also refers primarily to the study of the performing arts. The word *extract* is yet another example: *extract* in the US refers to food extracts but is used primarily as a verb in the UK. While the word *store* in English US typically refers to a grocery store or a hardware store, in English UK *store* also refers to a container (for eg. a store of gold). We reiterate here that the GEODIST method picks up on finer distributional cues that baseline methods cannot detect.

### 4 Related Work

A large body of work studies how language varies according to geography and time (Eisenstein et al. 2010; Eisenstein, Smith, and Xing 2011; Bamman, Eisenstein, and Schnoebelen 2014; Bamman and others 2014; Kim et al. 2014;



(a) Observed score for hand



(b) Observed score for buffalo

Figure 3: Observed scores computed by GEODIST (in  $\color{green}\rule{0.4pt}{1cm}$ ) for *buffalo* and *hand* when analyzing regional differences between New York and USA overall. The histogram shows the distribution of scores under the null model. The 98% confidence intervals of the score under null model are shown in  $\color{red}\rule{0.4pt}{1cm}$ . The observed score for *hand* lies well within the confidence interval and hence is not a statistically significant change. In contrast, the score for *buffalo* is outside the confidence interval for the null distribution indicating a statistically significant change.

Kulkarni et al. 2015; Kenter et al. 2015; Gonçalves and Sánchez 2014).

While previous work like (Gulordava and Baroni 2011; Berners-Lee et al. 2001; Kim et al. 2014; Kenter et al. 2015; Brigadir, Greene, and Cunningham 2015) focus on temporal analysis of language variation, our work centers on methods to detect and analyze linguistic variation according to geography. A majority of these works also either restrict themselves to two time periods or do not outline methods to detect when changes are significant. Recently (Kulkarni et al. 2015) proposed methods to detect statistically significant linguistic change over time that hinge on timeseries analysis. Since their methods explicitly model word evolution as a time series, their methods cannot be trivially applied to detect geographical variation.

Several works on geographic variation (Bamman, Eisenstein, and Schnoebelen 2014; Eisenstein et al. 2010; O'Connor and others 2010; Doyle 2014) focus on lexical variation. (Bamman, Eisenstein, and Schnoebelen 2014) study lexical variation in social media like Twitter based on gender identity. (Eisenstein et al. 2010) describe a latent variable

Word	Effect Size	CI(Null)	US Usage	UK Usage
theatre	0.6067	(0.004,0.007)	<i>great love for the theatre</i>	<i>in a large theatre</i>
schedule	0.5153	(0.032,0.050)	<i>back to your regular schedule</i>	<i>a schedule to the agreement</i>
forms	0.595	(0.015, 0.026)	<i>out the application forms</i>	<i>range of literary forms (styles)</i>
extract	0.400	(0.023, 0.045)	<i>vanilla and almond extract</i>	<i>extract from a sermon</i>
leisure	0.535	(0.012, 0.024)	<i>culture and leisure (a topic)</i>	<i>as a leisure activity</i>
extensive	0.487	(0.015, 0.027)	<i>view our extensive catalog list</i>	<i>possessed an extensive knowledge (as in impressive)</i>
store	0.423	(0.02, 0.04)	<i>trips to the grocery store</i>	<i>store of gold (used as a container)</i>
facility	0.378	(0.035, 0.055)	<i>mental health, term care facility</i>	<i>set up a manufacturing facility (a unit)</i>

Table 1: Examples of statistically significant geographic variation of language detected by our method, GEODIST, between English usage in the United States and English usage in the United Kingdoms in Google Book Ngrams. (CI - the 98% Confidence Intervals under the null model)

model to capture geographic lexical variation. (Eisenstein et al. 2014) outline a model to capture diffusion of lexical variation in social media. Different from these studies, our work seeks to identify semantic changes in word meaning (usage) not limited to lexical variation. The work that is most closely related to ours is that of (Bamman and others 2014). They propose a method to obtain geographically situated word embeddings and evaluate them on a semantic similarity task that typically focuses on named entities, specific to geographic regions. Unlike their work which does not explicitly seek to identify which words vary in semantics across regions, we propose methods to detect and identify which words vary across regions. While our work builds on their work to learn region specific word embeddings, we differentiate our work by proposing a null model, quantifying the change and assessing its significance.

## 5 Conclusions

In this work, we proposed a new method to detect linguistic change across geographic regions. Our method explicitly accounts for random variation, quantifying not only the change but also its significance. This allows for more precise detection than previous methods. We comprehensively evaluate our method on large datasets to analyze linguistic variation between English speaking countries. Our methods are capable of detecting a rich set of changes attributed to word semantics, syntax, and code-mixing.

## Acknowledgments

This research was partially supported by NSF Grants DBI-1355990 and IIS-146113, a Google Faculty Research Award, a Renaissance Technologies Fellowship and the Institute for Computational Science at Stony Brook University. We thank David Bamman for sharing the code for training situated word embeddings. We thank Yingtao Tian for valuable comments.

## References

Bamman, D., et al. 2014. Distributed representations of geographically situated language. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 828–834.

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*.

Berners-Lee, T.; Hendler, J.; Lassila, O.; et al. 2001. The Semantic Web. *Scientific American*.

Brigadir, I.; Greene, D.; and Cunningham, P. 2015. Analyzing discourse communities with distributional semantic models. In *ACM Web Science 2015 Conference*. ACM.

Doyle, G. 2014. Mapping dialectal variation by querying social media. In *EACL*.

Eisenstein, J.; O’Connor, B.; Smith, N. A.; Xing, Eric Eisenstein, J.; O’Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *EMNLP*.

Eisenstein, J.; O’Connor, B.; Smith, N. A.; and Xing, E. P. 2014. Diffusion of lexical change in social media. *PLoS ONE*.

Eisenstein, J.; Smith, N. A.; and Xing, E. P. 2011. Discovering sociolinguistic associations with structured sparsity. In *ACL-HLT*.

Gonçalves, B., and Sánchez, D. 2014. Crowdsourcing dialect characterization through twitter.

Gulordava, K., and Baroni, M. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *GEMS*.

Kenter, T.; Wevers, M.; Huijnen, P.; et al. 2015. Ad hoc monitoring of vocabulary shifts over time. In *CIKM*. ACM.

Kim, Y.; Chiu, Y.-I.; Hanaki, K.; Hegde, D.; and Petrov, S. 2014. Temporal analysis of language through neural language models. In *ACL*.

Kulkarni, V.; Al-Rfou, R.; Perozzi, B.; and Skiena, S. 2015. Statistically significant detection of linguistic change. In *WWW*.

Kulkarni, V.; Perozzi, B.; and Skiena, S. 2015. Freshman or fresher? quantifying the geographic variation of internet language. *CoRR* abs/1510.06786.

Labov, W. 1980. *Locating language in time and space / edited by William Labov*. Academic Press New York.

Michel, J.-B., et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.

Milroy, J. 1992. *Linguistic variation and change: on the historical sociolinguistics of English*. B. Blackwell.

O’Connor, B., et al. 2010. Discovering demographic language variation. In *NIPS Workshop on Machine Learning for Social Computing*.

Tagliamonte, S. A. 2006. *Analysing Sociolinguistic Variation*. Cambridge University Press.