Computers, Environment and Urban Systems 59 (2016) 244–255

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Understanding U.S. regional linguistic variation with Twitter data analysis

Yuan Huang [a], Diansheng Guo [a,*], Alice Kasakoff [a], Jack Grieve [b]

[a] Department of Geography, University of South Carolina, United States
[b] School of Languages and Social Sciences, Aston University, United Kingdom

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form 9 December 2015
Accepted 15 December 2015
Available online 31 December 2015

Keywords:
Social media
Linguistic
Twitter
American dialects
Regionalization
US regions
Spatial data mining

ABSTRACT

We analyze a Big Data set of geo-tagged tweets for a year (Oct. 2013–Oct. 2014) to understand the regional linguistic variation in the U.S. Prior work on regional linguistic variations usually took a long time to collect data and focused on either rural or urban areas. Geo-tagged Twitter data offers an unprecedented database with rich linguistic representation of fine spatiotemporal resolution and continuity. From the one-year Twitter corpus, we extract lexical characteristics for twitter users by summarizing the frequencies of a set of lexical alternations that each user has used. We spatially aggregate and smooth each lexical characteristic to derive county-based linguistic variables, from which orthogonal dimensions are extracted using the principal component analysis (PCA). Finally a regionalization method is used to discover hierarchical dialect regions using the PCA components. The regionalization results reveal interesting linguistic regional variations in the U.S. The discovered regions not only confirm past research findings in the literature but also provide new insights and a more detailed understanding of very recent linguistic patterns in the U.S.

1. Introduction

Dialects are forms or varieties of language that belong to a specific region or social group (Chambers & Trudgill, 1998). Research in dialectology not only seeks to understand language differences, language innovations and language variations through time and space, but also helps reveal patterns of information diffusion and cultural interpenetration (Di Nunzio, 2013). Most research on dialects relies on surveys and interviews, which may not contain enough information to identify regional linguistic variations objectively due to the small sample size and lack of computational statistical methods (Grieve, 2009). For example, the recent nationwide linguistic research, described in the *Atlas of North American English*, only contains 762 surveys (individuals) for 297 urban areas (Labov, Ash, & Boberg, 2006). Grieve (2009) introduced quantitative spatial autocorrelation statistics as well as using corpora of natural language data to dialectology. Grieve et al. (2011, 2013) also analyzed regional linguistic variation in American English based on a 26-million-word corpus of letters to editors and the data from Labov et al. (2006); however, neither data set captures linguistic variation in rural areas.

In this research, we use geo-tagged Twitter data as an alternative linguistic database, which can offer spatial and temporal continuity, granularity and up-to-date dynamics for linguistic studies. We present a linguistic study using a one-year dataset of geo-tagged tweets in the continental U.S. (48 states and Washington D.C.), from Oct. 7, 2013 to Oct. 6, 2014, which contains 6.6 million unique Twitter users, 924 million geo-tagged tweets, and 7.8 billion words.

Dialect variations can be examined by differences in lexicon, phonology, grammar, and pragmatics (Wolfram & Schilling-Estes, 2005). However, it is infeasible to attempt to study all linguistic variables that characterize dialects. Therefore, dialect studies often use representative sets of linguistic variables, which may include lexical (Grieve et al., 2011; Kurath, 1949), phonetic and phonological (Labov, Ash, & Boberg, 2006; O'Cain, 1979), and grammatical variation (Atwood, 1953). For this study, we use *lexical alternations* to examine linguistic variations and use counties in the U.S. as the unit for spatial analysis of regional linguistic variations.

In this research, we address two important questions: How do linguistic characteristics vary from place to place based on geo-tagged Twitter data and what are the linguistic regions and sub-regions in the U.S.? Twitter data not only offers spatial–temporal continuity but also allows close examination of a language in its *casual* expressions. Our data has 7.8 billion words and 6.6 million Twitter users, which is much larger than those being used in previous studies. We try to answer the above two questions based on the regional patterns generated by

* Corresponding author at: Department of Geography, University of South Carolina, 709 Bull Street, Room 127 , Columbia, SC, 29208, United States.
E-mail address: guod@mailbox.sc.edu (D. Guo).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.12.003
0198-9715/© 2015 Elsevier Ltd. All rights reserved.

each single variable, as well as the aggregated regional patterns. Adaptive kernel smoothing is used to estimate unknown values and to reduce noise. A hierarchical regionalization method is used to discover dialect regions with the top PCA components of linguistic variables extracted from tweets. The regionalization results reveal interesting linguistic regional variations in the U.S. and each region can also have sub-regions of local linguistic characteristics.

2. Background

The traditional way to collect dialect variation was to send out fieldworkers to collect linguistic related transcriptions from selected communities and representative speakers (McDavid, McDavid, Kretzschmar, Lerud, & Ratliff, 1986). One representative survey was conducted by Hans Kurath (1949) who proposed a plan for a Linguistic Atlas of the United States and Canada, which set the foundation of the project *Linguistic Atlas of Middle and South Atlantic States* (*LAMSAS*) (Kretzschmar, 1988). *LAMSAS* included 1162 interviewed subjects and the data collection period was from 1933 to 1974 (Nerbonne & Kleiweg, 2003). Then Kretzschmar (1993) spent several years making the data in LAMSAS accessible for reanalysis. Another work that has had a profound influence on North America English dialect research is the Atlas of *North American English* (*ANAE*) (Labov et al., 2006). It indicated that dialect diversity is increasing and several dialect regions display homogeneity across great distances (Labov, 2011). However, the interviewed subjects in both *LAMSAS* and *ANAE* are rather few people compared to the population and it took a long time to collect the data. *ANAE* even does not include rural areas. Grieve (2009) put forward a corpus-based regional dialect survey based on letters to editors and presented a statistical analysis of lexical variations in American English (Grieve et al., 2011). Their approach includes three steps: (1) identify significant regional variation patterns with spatial autocorrelation measures; (2) apply factor analysis to identify common dialect patterns; and (3) conduct cluster analysis to identify dialect regions. However, the data set focuses on formal written English.

Previous linguistic studies that use Twitter data have mainly focused on natural language processing and parts-of-speech tagging. Hong, Convertino, and Chi (2011) conducted a systematic analysis on the cross-language differences in tweets. Petrovic, Osborne, and Lavrenko (2010) built a Twitter corpus to help researchers work on natural language processing. Gimpel et al. (2011) used Twitter data to address the problem of part-of-speech tagging. Recently, more research has begun to use Twitter to study linguistic variations. Gonçalves and Sánchez (2014) used two years of Twitter data to study Spanish varieties at a global scale. Eisenstein, O'Connor, Smith, and Xing (2014) applied a latent vector autoregressive model on 107 million Twitter messages to study the diffusion of linguistic change over the United States. Criticisms of using Twitter data are mainly based on the uncertainty of its data quality and its socio-demographic representativeness (Crampton et al., 2013). Longley, Adnan, and Lansley (2015) attempted to profile Twitter users in terms of age, gender, and ethnicity based on user names. They point out that Twitter data may have an over representation of males and young adults. Goodchild (2013) argued that although big data may lack a normal process for quality control and rigorous sampling, big data can still be of high quality with its detailed, timely and original information (Kitchin, 2013).

Traditional dialectology research is generally qualitative. Séguy (1971) was the first to introduce statistical analysis of aggregated regional linguistic variation, an approach to dialectology known as *dialectometry*, which has been expanded on by various researchers who use multivariate and spatial methods to identify common patterns of regional linguistic variation (Goebl, 2006; Grieve et al., 2011; Heeringa, 2004; Kretzschmar, 1996; Lee & Kretzschmar, 1993; Nerbonne, 2006, 2009; Nerbonne & Kretzschmar, 2003; Nerbonne et al., 1996; Szmrecsanyi, 2013; Wieling & Nerbonne, 2011). Multivariate analysis usually involves examination of the joint relationship of variables and dimension reduction (James & McCulloch, 1990). Nerbonne (2006) introduced factor analysis to aggregate linguistic analysis. Thill, Kretzschmar, Casas, and Yao (2008) adopted Kohonen's (2001) self-organizing map to analyze the variations of word usage and pronunciation using the *LAMSAS* dataset. Principal component analysis (PCA) is another popular method used for multivariate analysis, which reduces variable dimensions with fewer measurements while retaining data variability in the original data (Rao, 1964). In spatial analysis, regionalization is the process of constructing homogeneous regions, e.g., climate zones or dialect regions, by optimizing a homogeneity function during the partition of space (Goodchild, 1979; Guo, 2008; Haining, Wise, & Blake, 1994, Handcock & Csillag, 2004; Masser & Scheurwater, 1980; Spence, 1968). Guo (2008) proposed a family of regionalization methods for constrained hierarchical clustering and partitioning (REDCAP) with multivariate information and a homogeneity measure, which has been applied in different domains such as forestry (Kupfer, Gao, & Guo, 2012) and health studies (Wang, Guo, & McLafferty, 2012). In this research, we use PCA to extract variables for describing linguistic characteristics and use REDCAP to discover dialect regions with the top PCA components.

Table 1
Content word lexical alternations.

Alternation		Alternation		Alternation	
Variant A	Other variant(s)	Variant A	Other variant(s)	Variant A	Other variant(s)
Bag	Sack	Mom	Mother	Absurd	Ridiculous
Clearly	Obviously	Whilst	While	Chuckle	Laugh
Grandfather	Grandpa	Center	Middle	Disturb	Bother
Couch	Sofa	Clothing	Clothes	Humiliating	Embarrassing
Automobile	Car	Best	Greatest	Job	Employment
Pupil	Student	Loyal	Faithful	Joy	Pleasure
Maybe	Perhaps	Real	Genuine	Likely	Probable
Especially	Particularly	Sad	Unhappy	Normal	Usual
Alley	Lane	Smart	Intelligent	Starting	Beginning
Holiday	Vacation	Baby	Infant	Start	Begin
Big	Large	Bet	Wager	Stupid	Dumb
Little	Small	Bought	Purchased	Unclothed	Naked
Supper	Dinner	Careful	Cautious	Bathroom	Restroom/washroom
Wrong	Incorrect	Comprehend	Understand	Envious	Jealous/covetous
Anywhere	Anyplace	Rude	Impolite	Quick	Fast/rapid
Required	Needed	Drowsy	Sleepy	Stomach	Tummy/belly
Each other	One another	Honest	Truthful	Trash	Garbage/rubbish
Afore	Before	Hug	Embrace	Grandma	Grandmother/granny/nana
Dad	Father	Hurry	Rush	All you	Y'all/you all/you guys
Ill	Sick	Band	Aid		