Examining Regional Variation Through Online Geotagged Corpora

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree

Master of Arts

in the Graduate School of The Ohio State University

By

Robert Brice Russ

Graduate Program in Linguistics

The Ohio State University

2013

Thesis Committee:
Professor Kathryn Campbell-Kibler, Advisor

Professor Brian Joseph
Abstract

Large-scale dialect surveys have long been a fundamental component of sociolinguistics and variation studies, but they have traditionally required significant investments of time and resources to collect relatively small amounts of data. In this study, I examine whether textual corpora collected from the Internet, particularly the social-networking website Twitter, can be used to conduct such surveys more quickly with less effort. I discuss the utility of Twitter as a linguistic data source, explain the computational and linguistic methods necessary to collect and process worthwhile data, and use corpora from Twitter to plot the distribution of three regional variables in American English: soft drink terminology, the use of ‘hella’ as an intensifier, and the morphosyntactic construction ‘needs X-ed’. I find that these corpora can replicate the findings of previous surveys (as seen with soft drink terms), reveal the existence of previously undocumented regional divides (as seen with ‘hella’), and highlight possible examples of ongoing linguistic diffusion (as seen with ‘needs X-ed’). I conclude that, although this method has issues with tracking certain types of sociolinguistic variables, it is still a promising source of data for conducting rapid studies in many areas of regional linguistic variation.
Acknowledgements

This study is the result of innumerable contributions. I wish to thank first and foremost my advisor, Kathryn Campbell-Kibler, and my advising committee of Brian Joseph and Chris Brew. The OSU Changelings and Clippers discussion groups, the attendees of GURT 2011, Walt Wolfram and Penelope Eckert have all provided valuable suggestions, and Bert Vaux was very gracious in providing the raw data from the Harvard Survey of North American Dialects. I also wish to thank Pete Warden and Jacob Eisenstein for their methodological insights.
Vita

May 2004 .......................... East Forsyth High School
June 2008 .......................... B.A. Linguistics, University of North Carolina at Chapel Hill
September 2009 - August 2012 ...... Graduate Teaching Associate, Department of Linguistics, The Ohio State University

Fields of Study

Major Field: Linguistics
# Table of Contents

Abstract ................................................................................................................................. ii

Acknowledgements ................................................................................................................ iii

Vita ........................................................................................................................................ iv

List of Figures ........................................................................................................................ vi

List of Tables .......................................................................................................................... viii

Chapter 1 Introduction .......................................................................................................... 1

Chapter 2 Background and Prior Research ......................................................................... 4
  2.0.1 Twitter As Social Mirror .............................................................................................. 9
  2.0.2 Variable Selection ....................................................................................................... 10

Chapter 3 Methods ................................................................................................................ 15
  3.0.3 Data Collection ......................................................................................................... 15
  3.0.4 Data Processing and Disambiguation ....................................................................... 17

Chapter 4 Results .................................................................................................................. 22
  4.0.5 Replicating Survey Findings: Soft Drink Terms ......................................................... 22
  4.0.6 Plotting New Variables: *hella* vs. *very* ................................................................. 25
  4.0.7 Morphosyntactic Variation: *need(s) + past participle* ............................................ 28

Chapter 5 Discussion .............................................................................................................. 32

Chapter 6 Conclusion .......................................................................................................... 35

Bibliography .......................................................................................................................... 37
List of Figures

2.1 A typical Twitter user profile, showing their most recent tweets. Note the use of @-replies in the third tweet and, especially, the location given at the top of the profile. ................................................................. 6

2.2 Chart demonstrating Twitter usage patterns among Internet users as a whole. (Smith and Rainie 2010) ................................................................. 7

2.3 General distribution patterns of *pop*, *soda* and *coke* as found by popvssoda.com (McConchie 2003). ................................................................. 12

2.4 Informal attestation of *needs X-ed* across United States from Murray, Frazer and Simon (1996). Dialect boundary lines in original figure adapted from McDavid (1958). ................................................................. 14

4.1 Map plotting the distribution of *pop* vs. *soda* vs. *coke* in the Twitter corpus. Yellow dots correspond to cities where *pop* predominates, red dots to *coke*, and blue dots to *soda*. ................................................................. 22

4.2 Map plotting the distribution of *pop* vs. *soda* vs. *coke* in the Harvard Survey of North American Dialects. ................................................................. 23

4.3 Comparison of variant usage rates between the HDS and Twitter corpora in states with large sample sizes ................................................................. 25

4.4 Map plotting the distribution of *hella* vs. *very* in the Twitter corpus. ........ 26

4.5 Map plotting the distribution of *hella* vs. *very* in the Twitter corpus, colored by usage rate. ................................................................. 26

4.6 Map plotting the distribution of *hella* vs. *very* in the Twitter corpus, colored by usage rate and focusing on the Bay Area. ................................................................. 27
4.7 Map plotting the distribution of *needs X-ed* vs. *needs X-ing* and *needs to be X-ed* in the Twitter corpus.  
4.8 Focus of previous figure on 'Midland' area of United States.
List of Tables

4.1 Comparison of variant counts for soft drinks between the HDS and Twitter corpus ......................................................... 24
4.2 Comparison of very/hella usage in Northern California cities .... 28
Chapter 1

Introduction

Though the study of sociolinguistic variation is vast, exploring linguistic diversity across a plethora of categories and contexts, the field of dialectology continues to play a key role in our understanding of language and its function in society. As one of the most salient forms of linguistic variation to many individuals, for example, dialects often play a key role in the construction and manifestation of individual and regional identities. This salience also makes dialectology a useful pedagogical tool; the discussion and study of dialects and dialect maps can spark further interest in students and laypeople about sociolinguistics and linguistics as a whole.

The study of linguistic geography, particularly when conducted and mapped on a large scale as seen in dialectal surveys and atlases, can serve as a bridge between other fields in the social sciences and humanities such as history and sociology (Wolfram and Schilling-Estes, 1998:20), and broad dialectal surveys can reveal interesting variational findings and possible communities of interest for other variationist studies. When these factors are considered alongside the intrinsic research value of dialects, it is clear that the collection and mapping of large-scale dialect atlases is still an important component of modern sociolinguistic research.

Even the methods and findings of early dialectological surveys have helped to guide efforts in other areas of linguistic variation (Milroy and Gordon, 2003:11). In the US, such surveys have included Hans Kurath’s Linguistic Atlas of New England (or LANE)
(Kurath 1934) and the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS),
which surveyed dialectal forms from New York to northeastern Florida (Kretzschmar 2005).
More recently, projects such as Labov, Ash and Boberg’s Telsur have attempted to doc-
ument regional variation across the United States via telephone surveys of a carefully-
sampled population (Labov 1999).

Unfortunately, such surveys as those mentioned above are extremely time- and
resource-intensive, often requiring years’ or decades’ worth of work by a team of several
researchers to elicit, collect and process data for thousands, or perhaps merely hundreds,
of speakers. LAMSAS, for example, collected 1,162 records over the course of 41 years
(1933 - 1974), and even the telephone-based Telsur required approximately 10 years (1991
- 2000) to collect data from 700 subjects. Both of these surveys required a staff of several
researchers to transcribe and analyze the collected data, making the process even longer
and more difficult.

A method of conducting dialect surveys with greater speed and ease would have
a variety of benefits. Not only would this allow for the study of more variables and more
speakers, each requiring fewer researchers, but a shortened timeframe for data collection
would allow researchers to study newly emerging terms and variables without waiting
years for results. A sufficiently rapid method of data collection may even allow for the
tracking of variables in quasi-real time, mapping their spread and diffusion on a monthly–
or weekly–basis.

Might the Internet provide the answer? On the World Wide Web, hundreds of mil-
lions of people communicate each day, often through recorded and archived textual media
which are typically much easier to analyze than their auditory counterparts.¹ In recent years, large textual corpora of Internet-based communications have been created and used to linguistic effect (e.g. Michel et al. 2010, Russ 2008). The collection and processing of corpora in which linguistic utterances are linked to the location of their utterance (or ‘geotagged’) may well provide a method of conducting large-scale dialect geography surveys with less time and effort.

In this paper, I propose that online geotagged corpora provide a means of conducting rapid dialect surveys, and examine the possibility of using one such source of corpora, the social-networking service Twitter, to conduct large-scale surveys of regional variation. I begin in Section 2 by providing an introduction to Twitter, discussing how its structure helps it accurately reflect offline social patterns and phenomena, review various (linguistic and non-linguistic) studies which have made use of this reflection, and describe three variables of American English which are well-suited for testing the validity of using Twitter to map dialectal variation. In Section 3, I explain the methods used to collect, process, analyze and map data for these variables from Twitter, including the usage of collocations to disambiguate word-senses of target variables. I outline the results of these variable surveys in Section 4, and use each variable to illustrate various advantages and disadvantages of this system in detail in Section 5. I conclude in Section 6 by summarizing the key points of this research, outlining possible avenues for further research, and mentioning ongoing plans to make this process easily available to all interested in conducting similar studies.

¹Using textual media also comes with its share of disadvantages, as well; phonetic and phonological variation, for example, may be difficult to impossible to study in a textual source. This issue is discussed further in Sections 2 and 6.
Chapter 2

Background and Prior Research

Twitter is a social networking service which allows users to post ‘tweets’, messages of 140 textual characters or less, for public\(^1\) viewing. Tweets can be sent via SMS (or ‘texting’), the Twitter website (http://www.twitter.com), specialized applications on cell phones, and other similar media. Created in 2006 (Carlson 2011a), Twitter (as of 2011) has over 300 million registered accounts (Carlson 2011b), and approximately 50 million users log in to Twitter each day (Twitter 2011e). Tweets are linked to individual user profiles, which contain fields for the user to input their name, basic biographical information, and their location.

Not only is Twitter a large-scale source of messages which are associable with the location of their speaker, these messages are produced in a variety of contexts and styles and created by a userbase that exhibits far more diversity than one might stereotypically associate with an Internet-based social networking service. Twitter can be used for broadcasting information to a wide audience; Twitter often brands itself as a medium for receiving and sharing information among many entities (Twitter 2011a), and an early study of Twitter found that one of its primary usages was for “daily chatter”.\(^2\) (Java et al. 2006)

\(^1\)Twitter also allows users to create ‘private’ profiles, in which their tweets are only shared with a self-selected set of users, but this option is not commonly used among the Twitter user base.

\(^2\)In fact, such chatter taken to an absurdly mundane level is often considered a stereotypical feature of Twitter. The archetypal tweet of this sort is along the lines of “Just ate a ham sandwich” (see, e.g. Porterfield and Carnes 2011).
Yet at the same time, Twitter is a common source of dialogic communication. Honeycutt and Herring (2009) examined the usage of “@-replies”, tweets which begin with an @ and the username of their intended recipient\(^3\), noting that Twitter is becoming used “for informal collaborative purposes” (9). Indeed, messages on Twitter often exhibit personally-oriented discourse and a generally informal style; Twitter uses the phrase “What’s happening?” at the top of its home page as a prompt for tweets, and Twitter is commonly considered a ‘microblog’, a condensed version of informal, personal weblogs (see Honeycutt and Herring).

Finally, Twitter (despite popular misconception) exhibits a fair amount of social and cultural diversity, at least for users in the United States. A survey conducted by the Pew Research Center’s Internet and American Life Project (Smith and Rainie 2010) examined Twitter usage between different sexes, age groups, ethnicities, rural/urban locations, and income and education levels. Although significant differences were found in some classifications–Internet users under the age of 29, for example, were found to more commonly be users of Twitter, as were users who self-identified as Hispanic–between 4% and 18% of all classifications surveyed were users of Twitter. No significant differences in usage were found between income levels, and only minor differences were found in sex and education level. Ultimately, Smith and Rainie found that 8% of Internet users as a whole have used Twitter or a similar microblogging service.

In short, Twitter is a promising source for many types of linguistic research which require data from across the United States–such as dialectology and regional variation.

\(^3\)This feature was created *ad hoc* by early users of Twitter, but was later officially incorporated into the service’s architecture.
Figure 2.1: A typical Twitter user profile, showing their most recent tweets. Note the use of @-replies in the third tweet and, especially, the location given at the top of the profile.

Twitter also has several technical advantages which improve its utility to prospective researchers. First, as has been alluded to previously, Twitter produces a truly copious amount of data. As of March 2011, over 140 million tweets were being sent per day (Twitter 2011b). For the sake of comparison, there are more tweets posted to Twitter on a daily basis than there are words in the British National Corpus.
### Twitter use by demographic group

% of internet users in each group who use Twitter

<table>
<thead>
<tr>
<th>All Internet Users</th>
<th>8%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>7</td>
</tr>
<tr>
<td>Women</td>
<td>10</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>14</td>
</tr>
<tr>
<td>30-49</td>
<td>7</td>
</tr>
<tr>
<td>50-64</td>
<td>6</td>
</tr>
<tr>
<td>65+</td>
<td>4</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td>5</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>13</td>
</tr>
<tr>
<td>Hispanic</td>
<td>18</td>
</tr>
<tr>
<td><strong>Household Income</strong></td>
<td></td>
</tr>
<tr>
<td>Less than $30,000</td>
<td>10</td>
</tr>
<tr>
<td>$30,000-$49,999</td>
<td>6</td>
</tr>
<tr>
<td>$50,000-$74,999</td>
<td>10</td>
</tr>
<tr>
<td>$75,000+</td>
<td>6</td>
</tr>
<tr>
<td><strong>Education level</strong></td>
<td></td>
</tr>
<tr>
<td>Less than High School</td>
<td>n/a</td>
</tr>
<tr>
<td>High School Diploma</td>
<td>5</td>
</tr>
<tr>
<td>Some College</td>
<td>9</td>
</tr>
<tr>
<td>College+</td>
<td>9</td>
</tr>
<tr>
<td><strong>Geography</strong></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>11</td>
</tr>
<tr>
<td>Suburban</td>
<td>8</td>
</tr>
<tr>
<td>Rural</td>
<td>5</td>
</tr>
</tbody>
</table>

Source: The Pew Research Center’s Internet & American Life Project, November 3-24, 2010 Post-Election Tracking Survey. n=2,257 adult internet users ages 18 and older, including 755 cell phone interviews. Interviews were conducted in English and Spanish.

Figure 2.2: Chart demonstrating Twitter usage patterns among Internet users as a whole. (Smith and Rainie 2010)
made public an application programming interface (or API) which allows developers and researchers to automatically collect and store publicly posted tweets\(^5\) via a number of programs or programming languages.

These data are automatically stored as text, a modality with mixed benefits and disadvantages in linguistic analysis. Purely textual data are often significantly easier to analyze than their auditory counterparts; the relative standardization of graphemes (especially in computerized formats) and the lack of graphemic variation in comparison to phonetic/phonological variation make the grouping, counting and processing of similar words or tokens a less complex process. On the other hand, textual data have historically been considered inferior sources in many areas of linguistic research\(^6\) because of their ‘derivative’ nature as a mere reflection of speech; one linguistic textbook summarizes this view when stating outright that “writing is not language” (Mihalicek and Wilson 2011:592).

Yet this attitude has softened somewhat in recent years, as both the nature of textual linguistic studies and the textual medium itself have evolved. The field of corpus linguistics itself is largely reliant on textual sources, and the work of Douglas Biber and others has produced findings of sociolinguistic interest in areas ranging from differences in British and American English (Biber 1987) to component features of stylistic factor groups (Biber 1988).

More recently, researchers of computer-mediated communication (or CMC) has begun to explore conversational elements present in the largely textual media of online interaction. Lynn Cherny’s early studies in online communities find that language use online

---

\(^5\) And associated information, including, most notably for this research, location.

\(^6\) But not all; notably, historical linguistics by its nature is often highly reliant on textual data.
reflects patterns of identity, as is done offline (Cherny 1999), and Crystal (2006) concludes in his survey of linguistic features that CMC\(^7\) shares elements of both speech and writing. Though many areas of CMC display new and fascinating linguistic features and patterns, online textual communication, particularly on Twitter, may well have the potential to provide insight into offline phenomena.

### 2.0.1 Twitter As Social Mirror

Studies have recently emerged in a variety of fields which make use of Twitter as such a reflection of offline social and linguistic phenomena. O’Connor et al. (2010), for example, used Twitter to conduct measures of public opinion, primarily of Barack Obama during 2008 and 2009, and discovered that their data correlated with consumer confidence and job approval polls by as much as 80%. Asur and Huberman (2010) used data from Twitter to predict the box-office revenues of recently-released movies, outperforming traditional predictors such as the Hollywood Stock Exchange.

More recently, researchers have begun to make use of this tool to answer more linguistically-oriented questions. Burger et al. (2011), for example, attempted to predict the gender of Twitter users via their tweets, username and Twitter profile, achieving a 92% success rate. Other research has even started to explore the possibility of using Twitter as a source for studying regional variation. In the Lexicalist project (Bamman 2010), individual words or tokens are plotted on a state-by-state level based on their usage rates in Twitter updates, blog posts, and other sources of “chatter”. Rather than focus specifically on regional variation, Lexicalist continues the theme of online data as an overall social mirror, also tracking approximate usage among various demographics (primarily age and gender)\(^7\) Which Crystal refers to as ‘Netspeak’.

---

\(^7\)Which Crystal refers to as ‘Netspeak’.
and promoting “trending” tokens, such as those related to current events or the entertainment industry.

In a study which focuses more closely on geographic differences in language use, Eisenstein et al. (2010) use topic-modeling processes to attempt to predict the location of a Twitter user solely from the content of their tweets. Some topics contain components that may reflect dialectal differences; the models determine, for example, that a user of certain Spanish-language terms such as papi (‘father’) is more likely to be from Los Angeles or the American Southwest.

However, many of Eisenstein et al.’s topics again focus on purely cultural features. A sports fan may use the word “Celtics” in many of their tweets, indicating they are from the Boston area, whereas a frequent tweeter of “Kings”\(^8\) would more likely be from Northern California. Such terms are not unique to the idiolect of a given region’s speakers (i.e. any NBA fan could discuss the Celtics or Kings) and do not act as different terms for the same referent (as would generally be necessary to consider them variants of the same linguistic variable), making them potentially less interesting from a linguistic viewpoint than from, perhaps, a sociological or geographical one. Nevertheless, Eisenstein et al. provide very strong evidence for the general existence of regional textual differences on Twitter, demonstrating that it can be easily witnessed on the level of individual words or tokens.

2.0.2 Variable Selection

Building on the general patterns of regional variation found above, my research focuses on examining the possibility of using Twitter to study variables likely to be of sociolinguistic and dialectological, rather than purely cultural, interest. Three variables from

\(^8\)Referencing the NBA’s Sacramento Kings.
American English were selected to test the potential utility of linguistic data from Twitter across as wide a range as possible. The first, soft drink terminology (e.g. ‘soda’ vs. ‘pop’ vs. ‘coke’), tests the ability of this method to find dialectal patterns similar to those in previously well-studied linguistic variables. The second, the use of ‘hella’ as a pre-adjectival intensifier, tests the ability of this method to provide insight into dialectal features that have not yet been mapped on a large scale. The third and final variable, variation between ‘needs X-ed’ or ‘needs + (past participle)’ (as seen in constructions such as The car needs washed) and ‘needs to be X-ed’, tests the utility of this method in morphosyntactic variation, in addition to the lexical variation examined in the first two variables.

**Soft Drink Terminology**

Regional variation in terms for soft drinks or carbonated beverages is one of the most well-known American English sociolinguistic variables among both linguists and laypeople. Recognized as a common dialectal variable since at least the mid-20th century (McDavid 1948), soft drink terminology has been surveyed on a national scale by linguistic projects such as the Harvard Survey of North American Dialects (Vaux 2005) and more informal works such as popvssoda.com (McConchie 2003).

Many terms have been attested for soft drink terminology. McDavid notes the common usage of *tonic*, and the Harvard Dialect Survey found usage for *cocola* and (of course) *soft drink*. Nevertheless, *soda*, *pop*, and *coke* are by far the most frequently used in contemporary American English—in the HDS, they account for over 90% of the 10,669 responses—and as such will be the terms of focus for in the study of this variable. In general, *soda* is most common in the Northeast, Southwest, and the area surrounding St. Louis, *coke* is predominant in the South from the Carolinas to Texas and New Mexico, and *pop* is generally used
from the Midwest to Pacific Northwest.

Figure 2.3: General distribution patterns of *pop*, *soda* and *coke* as found by popvssoda.com (McConchie 2003).

Intensifying ‘hella’

*Hella* is an intensifier, likely a clipping of the phrase “hell of a” (Bucholtz 2007) and generally used in alternation with terms such as *very*, as seen in sentences such as:

- Man this lab class is hella boring...
- its a very boring bible belt city unless you work for a bank

---

9Other intensifiers do, of course, exist, such as *incredibly* and *extremely*. However, *very* is by far the most common of these, and is primarily being used here to demonstrate the relative distribution of *hella*. *Very*, *incredibly*, etc., have not yet been found to exhibit any meaningful regional variation between themselves, above and beyond their general opposition respective to *hella*.

12
Hella has been in use since at least the mid-1990s (if not earlier) and is strongly associated with speakers from Northern California. In particular, hella is associated with the San Francisco Bay Area, both perceptually by Californians in general (Bucholtz et al. 2007) and students in California’s Central Coast (Smith 2011), as well as via qualitative observation of Bay Area students (Bucholtz 2007, Shankar 2008). However, this intensifier appears to have never been geographically surveyed.

‘Needs X-ed’

In the “need(s) X-ed” or “need(s) + past participle” variable, constructions such as needs done and needs fixed alternate with need(s) + present participle and need(s) to be + past participle (e.g. needs fixing, needs to be fixed). First attested in Scotch and Irish English as early as 1841 (‘need’ 2011), need(s) + past participle appears to have spread to dialects of American English in the early 20th century. Murray, Frazer and Simon (1996), in the most comprehensive study of need(s) + past participle in American English to date, note that this construction has been attested since at least the 1950s, but has not yet been rigorously examined. Their study, though not quite a national documentation of usage itself, compiles previous research and survey responses to create a map (see Figure 2.4) where need(s) + part participle has been “informally attested”.

This map indicates that needs X-ed is most commonly found between Pennsylvania and Illinois, which aligns with traditional opinions⁹ about the variable’s spread. However, no full-scale survey of this variable (especially in casual, non-elicited speech) has yet been conducted.

⁹Murray, Frazer and Simon, for example, note that virtually all prior attestations occurred “in the Midland area.” (1996:259)
Figure 2.4: Informal attestation of *needs X-ed* across United States from Murray, Frazer and Simon (1996). Dialect boundary lines in original figure adapted from McDavid (1958).
In this study, individual corpora were created for these three variables, each corpus containing a number of tweets which potentially contained an instance of a given variant. These corpora were collected, cleaned, processed and mapped to examine regional usage patterns using the processes described below. Notably, collocations were used to disambiguate word-senses and ensure that only tweets which represented a usage of a given variable were included in the final corpus.

3.0.3 Data Collection

Geotagged corpora were created using a hand-coded script\(^1\) which made calls to the Twitter Streaming API. The Streaming API (Twitter 2011c), a subset of the API mentioned in Section 2, allows users to automatically collect and process data posted to Twitter in real-time, constantly returning tweets as long as the connection is left open.

The Streaming API is highly configurable in its output, capable of returning not only the tweets posted to Twitter, but also the tweeter’s username, time of posting, and self-identified ‘real name’ and location, among other data. (It should be noted that this location most commonly refers to the Twitter user’s current city of residence, rather than their hometown or city of birth (Honeycutt and Herring); as such, this method primarily examines synchronic dialect usage contemporarily, rather than plotting usage by measures.

\(^1\) An early version of this script was adapted from a guide publicly available on the tech website *Ars Technica*. (Paul 2010)
This script makes use of the ‘statuses/filter’ method of the Streaming API, which returns tweets containing any of several keywords selected by the user. For example, if the API were called with the keywords *y’all* and *you*, it would only return tweets which contained the words *y’all* or *you*.

The Streaming API notably does not return every single tweet on Twitter which matches the requested keyword. First, all tweets from ‘private’ accounts (see Section 2) are excluded. Twitter then removes spam-like tweets for ‘result quality’; these include tweets which are posted repeatedly over a short period of time, tweets posted simultaneously over several accounts, and tweets which appear to be automated responses to certain keywords (Twitter 2011d). Twitter then identifies the relevant tweets (e.g. those containing the selected keyword(s)), and returns a rate-limited subset of those tweets\(^2\). In short, although the tweets returned by the API do not contain every single relevant tweet, the set of returned tweets are a random subset of the data on Twitter most likely to represent human communication.

The corpora for this study were collected periodically between January and September of 2011, during all times of day and all days of the week. The script was coded to return only tweets and their respective locations for a given keyword; no personally identifying data were collected.

\(^2\)Twitter’s exact rate limiting standards for the Streaming API have not been made public, but empirical tests indicate that the rate allows at least one tweet per second, if not more.
3.0.4 Data Processing and Disambiguation

As with any raw corpus, the collected data needed several steps of post-processing to most accurately reflect the variables in question. Not all self-identified user locations, for example, transparently reflect the location of the user. Some locations may be uninformative attempts at humor (“in front of a computer”), too general (“America”), or otherwise ambiguous (“East Side”, “Springfield”). As a result, only tweets with a clearly identified city and state (e.g. “San Antonio, TX”, “Carrboro, NC”) were retained for analysis. Furthermore, many posts to Twitter are “retweets”, essentially forwards of another person’s tweet; these were also excluded, as they reflect the linguistic features of the original tweeter, not the user forwarding it.

Finally, the Twitter API returns matches for multi-token keywords (e.g. “you all”) when the tokens appear at any point in the tweet. For example, matches for “you all” may include “What are you all doing?”, but they may also include tweets such as “All of you guys need to get a life.” Only tweets which include the multi-token keyword in the correct, adjacent order should be included in the final corpus; consequently, any tweets which did not follow this pattern were excluded from the corpus.

Even after these basic cleaning procedures, though, significant issues remain which may affect the accuracy of the corpus. Twitter returns all tweets matching a given keyword, regardless of their linguistic usage. If pop is selected as a keyword in a search for soft drink terms, Twitter will return tweets such as the following:

- Went today with Flg to talk to a school about helping get 1000 pounds of POP can tabs.
- TiVoing 30 for 30 while I go play responsible student. I’m sneaking a pop into the auditorium.
However, Twitter will also return tweets such as:

- Pop told me it would be days like this!
- he would give us a pop quiz at 8 in the morning
- I have this thing for Pop Tarts.

in which pop is clearly not describing a soft drink. In order to ensure that each corpus contains tweets in which the keyword is representing the desired linguistic variable, some process is necessary to reduce the homographic ambiguity seen above.

This study uses the process of collocates to achieve these aims. Collocates, or collocations, are essentially “the words that tend to co-occur with each target word” (Biber et al. 1998:6). Collocates are often used in corpus linguistics to disambiguate senses when the data set is too large to identify different word-senses by manually examining the data or its concordances.

When a word appears with a given collocate several times in the data, in many cases the sense of the word will be approximately constant across its various usages\(^3\). Furthermore, collocations can be grouped together as representative of the same general sense. For example, were the word pop to appear in one of the following collocations:

- pop \{out, up, under\}

it would generally be in use as a verb, meaning “to spring”, “to jump”, or something similar. Were it to appear, on the other hand, in collocations such as:

- pop \{music, artist, album\}

\(^3\)This is not always the case, of course; if a word has multiple noun-based senses, the collocation “a [WORD]\(^3\)”, for example, will generally not disambiguate these senses.
it would generally be regarded as a clipping of ‘popular’, specifically referring to a certain
genre of music.

In each of this study’s three variables, collocations are used in a unique way to per-
form word-sense disambiguation.

To create the corpus of soft drink terms, the keywords soda, pop, and coke were
used. In order to confirm that the resulting tweets were using the variants in the appro-
priate senses, the most frequent collocates of each variant were manually examined for
collocations which were both common among all three variants and which uniquely re-
ferred to the appropriate sense. drink X (e.g. drink pop) and drinking X were found to be
the two collocations best-suited for this purpose; there was no apparent non-target sense
of any of the tokens involved which would substantially overlap with the target sense in
these collocations4, and a manual review of a sample of tweets which contained such col-
locates indicated that the variables here were being used in the soft drink sense. All tweets
which used these collocations were therefore retained for analysis.

There is the case of possible ambiguity in the collocates of drink coke and drinking
coke, where coke could be a reference to the Coca-Cola brand name rather than a generic
soda. This collocation automatically removes all tweets which name Coca-Cola explicitly
(including Diet Coke, Cherry Coke, and other brand names); the collocation program as
coded is case-sensitive, removing all mentions of capitalized ‘Coke’ as well. Some tweets
may remain which are referring to Coca-Cola by the name coke, but these are impression-
istically observed not to be exceedingly common.

---

4For example, drink X would not take a word with a verbal sense (such as ‘pop (up)’), nor would it generally
take a clipping of ‘popular’ (cf. *drink pop music).
To create a corpus analyzing the regional distribution of *hella*, the keywords *hella* and *very* were used. Though this method could be used to simply track raw counts of *hella* in various cities, contrasting its usage with a similar intensifier such as *very* allows us to better track how often it is used in opposition to other intensifiers. *Hella* primarily appears before adjectives (as seen above), but can occasionally be seen before nouns (e.g. “Man all my homies & homegirls got hella followers on twitter”), meaning “many” or “a lot of”.

As a result, collocations were used here to remove instances of *hella* where it was being used prenominally, to ensure that the two variants were being examined in similar environments. The most frequent right-collocates of *hella* were manually examined for nouns; all tweets which contained these collocates were removed from the corpus. *Hella people, hella money*\(^5\), *hella ppl* (a shortening of ‘people’), and *hella followers* were the most common collocates. For consistency, the corpus was also cleaned of any collocates where *very* was used in a similar context, though few to no such collocates are likely to exist.

Because searching for the exact *needs X-ed* construction directly is impossible with Twitter’s current keyword structure, a trial corpus was constructed using the word *needs*, and the 200 most common right-collocates were manually searched for examples which used the *needs X-ed* construction—i.e., the most common verbs which are found in this construction. Five verbs were found to be particularly common in this format: *done, fixed, fired, washed,* and *filled*. The relevant constructions which used these verbs\(^6\) were used as keywords to construct a new corpus, which was then collected and processed. (There does not appear to be any evidence that certain verbs are more likely to appear in *needs X-ed* constructions—though Murray, Frazer and Simon found the construction to be acceptable across a variety of verbs. If there were, however, this method could potentially cause over-

\(^5\)This may also include some tweets which use *money* as a slang adjective meaning ‘of high quality’.

\(^6\)In full: *Needs done, needs to be done, needs doing, needs fixed, needs to be fixed, needs fixing, needs fired, needs to be fired, needs firing, needs washed, needs to be washed, needs washing, needs filled, needs to be filled, and needs filling.*
sampling of *needs X-ed* with respect to other variants.)

After the corpora were cleaned and processed, another script tallied the usage of each variant of a given variable by city, calculating the proportional usage of each variant on a city-by-city basis. These proportions were then geocoded and mapped automatically using Google’s online Fusion Tables software.
4.0.5 Replicating Survey Findings: Soft Drink Terms

The resulting corpus for soft drink terminology contained 2,952 instances of one of these variants in 1,118 different locations, which were then geocoded and plotted. In all, there were 1,958 instances of *soda*, 547 instances of *coke*, and 447 instances of *pop*. In Figure 4.1, a yellow dot indicates that a plurality of the tweets from that city used *pop*. A red dot indicates that a plurality from that city used *coke*, and likewise for blue and *soda*. 

![Figure 4.1: Map plotting the distribution of pop vs. soda vs. coke in the Twitter corpus. Yellow dots correspond to cities where pop predominates, red dots to coke, and blue dots to soda.](image)

22
In this map, *pop* is generally seen in the Midwest and Upper Inland South from western Pennsylvania to the Dakotas, *coke* is predominant in the South from North Carolina west to Texas, and *soda*, although witnessed throughout the United States, is most common in the Northeast and West Coast. This distribution is indeed highly consistent with previous large-scale surveys of this variable. When the Harvard Dialect Survey data, for example, are mapped using the same guidelines as the above data (see Figure 4.2), the similarity is immediately apparent.

![Figure 4.2: Map plotting the distribution of *pop* vs. *soda* vs. *coke* in the Harvard Survey of North American Dialects.](image)

In the HDS map, *pop* again is common in the Midwest, *coke* predominates in the South, and *soda* is seen nationally, but particularly in the Northeast and Southwest. Although *soda* is not as graphically dominant in the Harvard Dialect Survey map as it is in Figure 4.1, both surveys find it to be the majority variant (as seen in Table 4.1).

The Twitter corpus map also reflects the unusual spread of *soda* into St. Louis and
Table 4.1: Comparison of variant counts for soft drinks between the HDS and Twitter corpus

<table>
<thead>
<tr>
<th></th>
<th>Twitter</th>
<th>HDS</th>
<th>T (%)</th>
<th>H (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>soda</td>
<td>1958</td>
<td>5651</td>
<td>66.3</td>
<td>58.5</td>
</tr>
<tr>
<td>pop</td>
<td>447</td>
<td>2675</td>
<td>15.1</td>
<td>27.7</td>
</tr>
<tr>
<td>coke</td>
<td>547</td>
<td>1320</td>
<td>18.5</td>
<td>13.7</td>
</tr>
</tbody>
</table>

eastern Missouri, as seen in the HDS map (Figure 4.2) and, particularly, McConchie’s survey map (Figure 2.3).

The data from these two corpora were also compared more quantitatively by examining the usage of the three variants in each state with at least 120 tokens in the Twitter soft drink corpus; the results of this, plotted by percentage, can be seen in Figure 4.3. Though there are some occasional discrepancies, such as the undercounting of coke for Georgia and the slight undercounting of pop for Illinois in the Twitter corpus, this chart finds the ratios of each variant to be surprisingly consistent between corpora in most cases.

This does not mean, though, that there are no apparent differences in the general distributions of these two maps. As alluded to in Table 4.1, soda does appear more diffuse in the Twitter corpus map than the HDS. Because soda is used very frequently nationally (being the majority variant in both the HDS and Twitter corpora), it is feasible to expect it to be used in areas that are pop or coke-dominant; if such a usage is the only collected instance of a variant from a given city, it would then appear on the map—a possibly disproportionate number of times, in this case. The Twitter corpus also lacks significant data at this point for the sparsely populated Montana-Idaho-Wyoming area.
4.0.6 Plotting New Variables: hella vs. very

The resulting corpus for *hella* contained 360,860 instances of one of these variants in 13,428 different locations, which were then geocoded and plotted. Of these, 29,219 were instances of *hella*, with the remainder being instances of *very*. In Figure 4.4 below, a red dot indicates that a majority of tweets from that city used *hella*; likewise for yellow and *very*. Due to the particularly large corpus for this variable, locations on this map represent cities from which at least 40 tweets originated.

As this map shows, predominant usage of *hella* is largely restricted to the Bay Area. This generally holds true even if finer distinctions are made; Figure 4.5 below plots city-level *hella* usage into one of five categories.
Figure 4.4: Map plotting the distribution of *hella* vs. *very* in the Twitter corpus.

Figure 4.5: Map plotting the distribution of *hella* vs. *very* in the Twitter corpus, colored by usage rate.
This map does make clear, though, that *hella* is commonly found in some non-Bay Area locations, most notably Washington and Missouri. The predominance of *hella* in some parts of Washington, as well as marginal *hella* usage in Arizona, may be the result of Californian emigration to all parts of the West Coast—a process so well-known to have been given the disparaging name of ‘Californication’ (Burton 1972). The usage of *hella* elsewhere is more opaque, however; Bucholtz notes that *hella* has been “anecdotally reported to be in circulation around the country [as a] marked, trendy term” (2007:8), which may happen to be gaining particular prominence in these areas. The patterns behind such usages of *hella* on a national scale are an interesting topic for future research.

Figure 4.6: Map plotting the distribution of *hella* vs. *very* in the Twitter corpus, colored by usage rate and focusing on the Bay Area.
An equally intriguing, and more visibly striking, pattern emerges when the Bay Area itself is selected for examination. As seen in Figure 4.6, *hella* is common to predominant along the east side of the San Francisco Bay, particularly between Oakland and Fremont, with common usage extending inland as far northeast as Sacramento. On the other side of the bay, however, the situation is completely different; usage of *hella* is low to none in San Jose, Mountain View, and the rest of the Silicon Valley. This pattern can be seen quantitatively in the table below, with three sample ‘Silicon Valley’ cities compared against cities on the bay’s east side (with a non-Californian city included for comparison).

<table>
<thead>
<tr>
<th>City</th>
<th>‘very’</th>
<th>‘hella’</th>
<th>% ‘hella’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mountain View, CA</td>
<td>317</td>
<td>3</td>
<td>0.9%</td>
</tr>
<tr>
<td>Santa Clara, CA</td>
<td>111</td>
<td>19</td>
<td>14.6%</td>
</tr>
<tr>
<td>San Jose, CA</td>
<td>768</td>
<td>367</td>
<td>22.3%</td>
</tr>
<tr>
<td>Sacramento, CA</td>
<td>1115</td>
<td>1262</td>
<td>53.1%</td>
</tr>
<tr>
<td>Oakland, CA</td>
<td>695</td>
<td>1307</td>
<td>62.6%</td>
</tr>
<tr>
<td>Vallejo, CA</td>
<td>70</td>
<td>374</td>
<td>84.2%</td>
</tr>
<tr>
<td>Columbus, OH</td>
<td>1483</td>
<td>105</td>
<td>6.6%</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of *very/hella* usage in Northern California cities

Though this survey by itself cannot pinpoint the reasons behind this variation, some possible explanations will be proposed in Section 5.

4.0.7 Morphosyntactic Variation: *need(s) + past participle*

The resulting corpus for *needs X-ed* contained 6,406 instances of one of the relevant variants in 1,884 different locations, which were then geocoded and plotted. Of these, 341 instances were of *needs X-ed*, with the remainder being *needs X-ing* or *needs to be X-ed*. In Figure 4.7, a red dot indicates that a plurality of tweets from that city used the *needs X-ed* construction, while a yellow dot indicates dominance by other constructions.
Figure 4.7: Map plotting the distribution of needs X-ed vs. needs X-ing and needs to be X-ed in the Twitter corpus.

This map finds predominance of needs X-ed from the Delaware Valley west to Illinois, in line with findings from Murray, Frazer and Simon. However, this survey also finds frequent needs X-ed usage in Missouri, Kentucky, and Tennessee, as seen in detail in Figure 4.8.
Figure 4.8: Focus of previous figure on ‘Midland’ area of United States.
Though *needs X-ed* can be found sporadically across the United States, the most prominent spread of this variant is clearly to the south (and the South). Such a shift may require reconceptions of the ‘Midland’ dialect and/or its relation to the *needs X-ed* variable—or, if these instances are being produced by transplanted Midland speakers, provide insight into a possible future influence on Upper South/Inland South speakers. This study also confirms Murray, Frazer and Simon’s attestations of *needs X-ed* in eastern Pennsylvania and the Delaware Valley, even in urban areas\(^1\)–a finding which may go against anecdotal conceptions. The Twitter corpus also finds sporadic attestations of *needs X-ed* in the West, particularly in Oregon, which were not discovered in the Murray, Frazer and Simon study.

\(^1\)In Philadelphia, PA, for example, 4 of 79 tweets used *needs X-ed*. 

31
Surveys conducted using online geotagged corpora are clearly capable of both replicating previously surveyed regional variables and revealing interesting findings in previously understudied variables. The constructed map for soft drink terminology, for example, follows the same patterns that are seen in the Harvard Dialect Survey and popvs-soda.com, and collected comparable amounts of data over a matter of months without elicitation, promotion or supervision.

Meanwhile, the first national survey of *hella* usage reveals interesting findings, particularly in the Bay Area. The divide between heavy *hella* usage on the bay’s east side and the Central Valley, and its virtual nonexistence in the Silicon Valley and the Pacific shore, is so sharp as to be suggestive of an isogloss. Is this division being caused by the immigration of non-*hella* speakers from across the country to work in Silicon Valley’s technology sector? Are social differences between the two communities being reflected on a regional level?

Similar questions could be asked of the appearance of *hella* in several cities around the United States. Is this evidence of the ongoing geographic diffusion of *hella*? If so, is the process causing *hella* to appear more frequently in the Pacific Northwest the same that is leading to its introduction to cities such as Atlanta and St. Louis? This survey cannot answer these questions, of course, but it can make researchers aware of their existence and
provide very promising avenues for future study.

The final studied variable, *needs X-ed*, builds on prior research to hint at the possibility of ongoing linguistic change. Murray, Frazer and Simon found little evidence of *needs X-ed* outside of its core area, save for a few attestations in the Mountain West, but the Twitter corpus finds usage in a variety of nearby states. This supports recent research done by Ulrey (2009), who examined a corpus of news articles to find evidence of *needs X-ed* usage in Florida, Arkansas, and a variety of other Southern states.

At the same time, this methodology is not without its issues. Most notably, the solely textual nature of data from Twitter makes it impossible to use this method to study phonetic variation, and highly infeasible at best to study any type of phonological variation. These domains, the most common fields of many dialectologists and variationist sociolinguists, provide more opportunity for variation and the ability to study fine details of variation (as in, e.g., vowel shifts), whereas lexical and morphosyntactic variables are relatively fewer in number.

Even morphosyntactic and lexical variables do not work universally well with this methodology. Twitter has the potential to return hundreds of thousands of data points in a very short period of time (as seen with *very/hella*), but variables which occur relatively rarely in casual speech may return less copious data. The relative paucity of data for *needs X-ed*, where only 341 tokens were collected during the few months that the script was operated, provides one example of this potential drawback.

Because Twitter is a public medium, the usage of some variables may be influenced by their social evaluation. Labov’s (2001) notable distinction between indicators, markers and stereotypes cautions that some linguistic variables may be both perceived and nega-
tively evaluated by communities which come into contact with them. If users of Twitter are aware that a public audience can read their messages, they might possibly use variants which carry negative social judgments less frequently than if they were using them in an offline, more private conversation (cf. the ‘audience design’ theory of Bell 1984).
Chapter 6

Conclusion

In conclusion, online geotagged corpora, and Twitter in particular, are a very promising source for studying large-scale regional linguistic variation. Data can be collected easily and effectively without interviews, or even supervision, and these data can both confirm previous survey findings and allow new insights into previously unstudied variables. Furthermore, collocations can provide utility in defining variable contexts and conducting basic word-sense disambiguation.

Regarding the variables studied to demonstrate this methodology, this survey shows that the most common terms for soft drinks have not changed dramatically since the last large-scale surveys conducted of these variables in the early 2000s. The morphosyntactic variable needs X-ed, conversely, appears to be diffusing to areas further south and west of its ‘Midland’ origins. Finally, usage of hella is still largely confined to its Northern California origins, but demonstrates unusual, highly visible patterns of usage in that area which are worthy of further quantitative and qualitative study.

This method is ripe for further methods of refinement which will allow for even larger amounts of available data. More theoretically advanced methods of word-sense disambiguation may prove useful, as may geocoding programs which allow for the mapping of less unambiguous Twitter locations. Some Twitter users post their location in latitude and longitude coordinates\(^1\); these data may be mappable on the city level as well.

\(^1\)These are usually updated automatically by a GPS program on the user’s phone or laptop computer.
The scripts and programs which have been used in the collection and mapping of these data have been created with an ultimate eye towards releasing them for public use. Such a public program may ultimately make it possible for any interested researcher to quickly and effectively conduct large-scale dialect surveys, from professors and academics to students in sociolinguistics courses. With the impending rumored release of a multi-billion tweet corpus by the Library of Congress (Raymond 2010), which will provide almost unimaginable opportunities for corpus-based research, this project may well mark the beginning of a new era in online variationist studies.
Chapter 7

Bibliography


