Examining Large-Scale Regional Variation Through Online Geotagged Corpora

Brice Russ

Department of Linguistics
The Ohio State University
http://www.ling.osu.edu/~rbruss

2012 ADS Annual Meeting
Research Question

- Are textual corpora, collected from the Internet and tagged for location, feasible sources for creating dialect maps and studying regional variation?
- (e.g. Twitter)
Motivating Implications

- Online corpora provide more data more quickly
- Language observed in conversational settings, rather than elicited
Motivating Implications

- Online corpora provide more data more quickly
- Language observed in conversational settings, rather than elicited
  - Allows for collection of more variables, more speakers with less supervision
Motivating Implications

- Online corpora provide more data more quickly
- Language observed in conversational settings, rather than elicited
  - Allows for collection of more variables, more speakers with less supervision
  - Can track the spread of linguistic variables in (quasi-)real time
Outline

- Why (and how) Twitter can be used to study dialect variation
- Distribution of three variables:
  - Soft drink terminology (‘soda’/‘pop’/‘coke’)
  - Intensifier ‘hella’ (vs. ‘very’)
  - The ‘needs X-ed’ construction
- Findings and conclusions
Introduction To Twitter

- Microblogging service available via WWW, SMS
- Send publicly available messages of $\leq 140$ characters
Twitter as Data Source

- Very prolific source of textual linguistic data
  - 200 million tweets/day as of August 2011
Twitter as Data Source

- Very prolific source of textual linguistic data
  - 200 million tweets/day as of August 2011
- Used for conversational and informal purposes
  (Honeycutt and Herring 2009, Smith 2011)
Twitter as Data Source

- Very prolific source of textual linguistic data
  - 200 million tweets/day as of August 2011
- Used for conversational and informal purposes
  (Honeycutt and Herring 2009, Smith 2011)
- Exhibits diversity in age, gender, social class
  (Smith and Rainie 2010)
### Diversity Patterns on Twitter (Smith and Rainie)

#### Twitter use by demographic group

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Twitter Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Internet Users</td>
<td>8%</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>7</td>
</tr>
<tr>
<td>Women</td>
<td>10</td>
</tr>
<tr>
<td>Age</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>Race/Ethnicity</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>Household Income</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>Education level</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>
</tbody>
</table>
Social Patterns on Twitter

Twitter used to conduct public-opinion polling (O’Connor et al. 2010), predict box-office revenues (Asur and Huberman, 2010)
Twitter used to conduct public-opinion polling (O’Connor et al. 2010), predict box-office revenues (Asur and Huberman, 2010)

Eisenstein et al. (2010) and Bamman (2010) have studied textual/lexical variation on the macro-level
Social Patterns on Twitter

- Twitter used to conduct public-opinion polling (O’Connor et al. 2010), predict box-office revenues (Asur and Huberman, 2010)
- Eisenstein et al. (2010) and Bamman (2010) have studied textual/lexical variation on the macro-level
  - Eisenstein et al. use topic models to predict user location
  - Topics include both regional variables (‘hella’) and cultural markers (food, sports teams)
  - Demonstrates general existence of regional variation on Twitter
Data Collection

- Collected tweets using Python script calling Streaming API (Paul 2010), given a set of keywords predetermined by user
  - Non-spoken data
  - Difficult to examine phonetic/phonological variation

- Data collected in spring and summer of 2011 (primarily June - August)

- Script collects tweet and location of the tweeting user
  - Cities represent current location of speakers, not origin
Data Collection

- Collected tweets using Python script calling Streaming API (Paul 2010), given a set of keywords predetermined by user
  - Non-spoken data
  - Difficult to examine phonetic/phonological variation
- Data collected in spring and summer of 2011 (primarily June - August)
- Script collects tweet and location of the tweeting user
  - Cities represent current location of speakers, not origin
- Regular expression used to filter out ‘non-locations’
- ‘Re-tweets’ (forwarded posts) are excluded
## Sample Data

<table>
<thead>
<tr>
<th>Location</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toronto, ON</td>
<td>I remember when people would try and pear pressure me to Drink pop and they’d say no one will no. Wrong, I’ll know.</td>
</tr>
<tr>
<td>Birmingham, AL</td>
<td>@mhirsh32 Would probably be opening a can of soda/ bottle of water, drinking a sip or two, then never touching it again. Still thinking.....</td>
</tr>
<tr>
<td>MIC CITY, TX</td>
<td>To stop drinking soda, I imagine the same yucky feeling I get when I see ppl lifting cigarettes to their lips...so far, it’s working!</td>
</tr>
<tr>
<td>Washington, DC</td>
<td>Eric Weaver gives honest view that his org is doing what they do as a subsidized service. Not everyone ”needs” 2 be profit driven #mfusa2011</td>
</tr>
<tr>
<td>Secane, PA</td>
<td>Drinking diet soda doesn’t do shit when you’ve got a familt sized bag of nacho cheese combos and a twix bar in front of you too.</td>
</tr>
<tr>
<td>Dallas, TX</td>
<td>Fired up my Crock Pot for this first time this morning. Picked recipe that needs to cook for 10 hours so it should be ready when I get home.</td>
</tr>
</tbody>
</table>
Homographic Ambiguity

- Variables exhibit lexical ambiguity
- Example: ‘pop’
Homographic Ambiguity

- Variables exhibit lexical ambiguity
- Example: ‘pop’
  - “im startin to feel like its bad to drink pop haha”
Homographic Ambiguity

- Variables exhibit lexical ambiguity
- Example: ‘pop’
  - “im startin to feel like its bad to drink pop haha”
  - “he would give us a pop quiz at 8 in the morning”
Homographic Ambiguity

- Variables exhibit lexical ambiguity
- Example: ‘pop’
  - “im startin to feel like its bad to drink pop haha”
  - “he would give us a pop quiz at 8 in the morning”
  - “I have this thing for Pop Tarts.”
Homographic Ambiguity

- Variables exhibit lexical ambiguity
- Example: ‘pop’
  - “im startin to feel like its bad to drink pop haha”
  - “he would give us a pop quiz at 8 in the morning”
  - “I have this thing for Pop Tarts.”
- Must distinguish the appropriate sense from homographs
Collocations

- Categorize variants by co-occurring words/phrases
- Common sense-disambiguators in corpus linguistics (e.g. Biber et al. 1998)
Collocations

- Categorize variants by co-occurring words/phrases
- Common sense-disambiguators in corpus linguistics (e.g. Biber et al. 1998)
- Manually select most frequent collocations with the desired sense/meaning
- Example: ‘pop’
  - pop \{out, up, under\}
  - pop \{music, artist, album\}
Collocations

- Categorize variants by co-occurring words/phrases
- Common sense-disambiguators in corpus linguistics (e.g. Biber et al. 1998)
- Manually select most frequent collocations with the desired sense/meaning
- Example: ‘pop’
  - pop \{out, up, under\}
  - pop \{music, artist, album\}
  - \{drink, drinking\} pop
Variables

Mapped using Google Fusion Tables software

- ‘soda’/‘pop’/‘coke’
- ‘hella’/‘very’
- ‘needs X-ed’
Map Comparison: Soda vs. pop vs. coke

- Account for over 90% of soft drink variation (Vaux 2003)
  - ‘Pop’ predominant in Midwest to Pacific Northwest
  - ‘Coke’ predominant in the South (South Carolina to Texas)
  - ‘Soda’ used everywhere, but used exclusively in New England and Southwest
Dialect map plotted from Twitter corpus

(yellow = ‘pop’; red = ‘coke’; blue = ‘soda’)
2,952 tweets, 1,118 locations
Dialect map plotted from Harvard Dialect Survey

Variable #1: ‘soda’/‘pop’/‘coke’
Variable #2: ‘hella’/‘very’
Variable #3: ‘needs X-ed’
New Research: ‘hella’/‘very’

- ‘Hella’ as an intensifier (in similar environment to ‘very’)
New Research: ‘hella’/‘very’

- ‘Hella’ as an intensifier (in similar environment to ‘very’)
  - “Man this lab class is **hella** boring...”
  - “its a **very** boring bible belt city unless you work for a bank”
“Hella’ as an intensifier (in similar environment to ‘very’)
  “Man this lab class is hella boring…”
  “its a very boring bible belt city unless you work for a bank”

Associated perceptually with ‘Northern California’ (Bucholtz et al. 2007), but usage has only been examined anecdotally (Bucholtz 2007)
New Research: ‘hella’/‘very’

- ‘Hella’ as an intensifier (in similar environment to ‘very’)
  - “Man this lab class is hella boring...”
  - “its a very boring bible belt city unless you work for a bank”

- Associated perceptually with ‘Northern California’ (Bucholtz et al. 2007), but usage has only been examined anecdotally (Bucholtz 2007)

- Collocates used here to remove non-similar environments (‘hella {people, ppl, followers, money}’)

Brice Russ  Dialect Geography on Twitter
Over 300,000 data points:

(yellow = ‘very’; red = ‘hella’)

Brice Russ
Dialect Geography on Twitter
5-binned map
Silicon Valley speakers are hella standard
Silicon Valley speakers are hella standard

<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>

Comparison of *very/*hella usage in Northern California cities
Morphosyntax: ‘needs X-ed’

- ‘need + (past participle)’ common in Midwest (Murray et al. 1996)
- Varies with ‘needs X-ing’ and ‘needs to be X-ed’
Prior attestation of ‘needs X-ed’

(from Murray et al. 1996)
Prior attestation of ‘needs X-ed’

(from Murray et al. 1996)


6,406 data points, 1,884 locations
The ‘needs’ of the many...

Dark areas (Northeast, etc.) represent overlap of data points
Range from Murray et al.: Illinois to New Jersey
Focus on ‘Midwest’ region

Diffusion southward since Murray et al? (cf. Ulrey 2009)
Twitter is a very promising source for studying regional variation.
Conclusions

- Twitter is a very promising source for studying regional variation
- Data can be collected easily and effectively without interviews, supervision
Twitter is a very promising source for studying regional variation

Data can be collected easily and effectively without interviews, supervision

Most effective with common lexical variables
Twitter is a very promising source for studying regional variation
Data can be collected easily and effectively without interviews, supervision
Most effective with common lexical variables
Collocations can prove useful in defining variable contexts
Future Research Goals

- Improve data collection, mapping processes
- Present version of program for public use
  - Python script available; standalone application forthcoming
  - Tools for corpora collection, collocation, mapping
- Explore larger corpora
  - Library of Congress Twitter Corpus in development
Thank you!

Thanks also to:
- Kathryn Campbell-Kibler
- Chris Brew
- Brian Joseph
- Changelings, Clippers, and the attendees of GURT 2011
- Bert Vaux
- Jacob Eisenstein
- Pete Warden
- Walt Wolfram
- ...and many others!
*Proceedings of the ACM International Conference on Web Intelligence.*


Contact

rbruss@ling.osu.edu
Twitter: @kilroywashere
Maps and script available at:
http://www.briceruss.com/ADStalk
Corpus #1 does not include tweets using:

- Coca-Cola
- Diet Coke, Cherry Coke, etc.
- Capitalized ‘Coke’
- ‘drinking a coke’

*Can* Coke(brand) and Coke(drink) be fully disambiguated?
Data Collection Procedure

1. Script sends keyword requests (‘soda’/‘pop’/‘coke’) for Twitter live public (> 90%) stream
Data Collection Procedure

1. Script sends keyword requests (‘soda’/‘pop’/‘coke’) for Twitter live public (> 90%) stream
2. Twitter removes spam-like tweets from stream
Data Collection Procedure

1. Script sends keyword requests (‘soda’/‘pop’/‘coke’) for Twitter live public (> 90%) stream
2. Twitter removes spam-like tweets from stream
3. Twitter sets access level (10 tweets out of every 100)
1. Script sends keyword requests (‘soda’/‘pop’/‘coke’) for Twitter live public (> 90%) stream
2. Twitter removes spam-like tweets from stream
3. Twitter sets access level (10 tweets out of every 100)
4. Twitter returns all tweets matching keyword, rate-limited
Data Collection Procedure

1. Script sends keyword requests (‘soda’/‘pop’/‘coke’) for Twitter live public (> 90%) stream
2. Twitter removes spam-like tweets from stream
3. Twitter sets access level (10 tweets out of every 100)
4. Twitter returns all tweets matching keyword, rate-limited
Spam Cleaning Process

Twitter removes accounts or tweets from the stream which:

- Repeatedly post duplicate tweets or links
- Post the same message over multiple accounts
- Aggressively follow and unfollow accounts
- Abuse ‘trending topics’ or hashtags
  - (e.g. “Get a loan from Unscrupulous Bank! #justinbieber #chicagobulls #twowordanswers”)

Brice Russ
Dialect Geography on Twitter
## Disambiguation Through Collocation Groups

<table>
<thead>
<tr>
<th>soda</th>
<th>pop</th>
<th>coke</th>
</tr>
</thead>
<tbody>
<tr>
<td>a soda</td>
<td>to pop</td>
<td>diet coke</td>
</tr>
<tr>
<td>diet soda</td>
<td>pop up</td>
<td>and coke</td>
</tr>
<tr>
<td>soda and</td>
<td>a pop</td>
<td>a coke</td>
</tr>
<tr>
<td>of soda</td>
<td>pop in</td>
<td>coke and</td>
</tr>
<tr>
<td>orange soda</td>
<td>pop culture</td>
<td>of coke</td>
</tr>
<tr>
<td>and soda</td>
<td>pop music</td>
<td>the coke</td>
</tr>
<tr>
<td>baking soda</td>
<td>pop out</td>
<td>coke in</td>
</tr>
<tr>
<td>drink soda</td>
<td>and pop</td>
<td>coke is</td>
</tr>
<tr>
<td>soda is</td>
<td>the pop</td>
<td>&amp; coke</td>
</tr>
<tr>
<td>the soda</td>
<td>pop a</td>
<td>cherry coke</td>
</tr>
<tr>
<td>soda on</td>
<td>of pop</td>
<td>coke zero</td>
</tr>
<tr>
<td>cream soda</td>
<td>pop off</td>
<td>coke bottle</td>
</tr>
<tr>
<td>drinking soda</td>
<td>pop star</td>
<td>on coke</td>
</tr>
<tr>
<td>...</td>
<td>pop the</td>
<td>coke with</td>
</tr>
<tr>
<td>...</td>
<td>pop it</td>
<td>my coke</td>
</tr>
</tbody>
</table>

Brice Russ  
Dialect Geography on Twitter
## Disambiguation Through Collocation Groups

<table>
<thead>
<tr>
<th>soda</th>
<th>pop</th>
<th>coke</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>a soda</em></td>
<td>to pop</td>
<td>diet coke</td>
</tr>
<tr>
<td>diet soda</td>
<td>pop up</td>
<td>and coke</td>
</tr>
<tr>
<td>soda and</td>
<td><em>a pop</em></td>
<td><em>a coke</em></td>
</tr>
<tr>
<td>of soda</td>
<td>pop in</td>
<td>coke and</td>
</tr>
<tr>
<td>orange soda</td>
<td>pop culture</td>
<td>of coke</td>
</tr>
<tr>
<td>and soda</td>
<td>pop music</td>
<td>the coke</td>
</tr>
<tr>
<td>baking soda</td>
<td>pop out</td>
<td>coke in</td>
</tr>
<tr>
<td>drink soda</td>
<td>and pop</td>
<td>coke is</td>
</tr>
<tr>
<td>soda is</td>
<td>the pop</td>
<td><em>&amp; coke</em></td>
</tr>
<tr>
<td><em>the soda</em></td>
<td>pop a</td>
<td>cherry coke</td>
</tr>
<tr>
<td>soda on</td>
<td>of pop</td>
<td>coke zero</td>
</tr>
<tr>
<td>cream soda</td>
<td>pop off</td>
<td>coke bottle</td>
</tr>
<tr>
<td>drinking soda</td>
<td>pop star</td>
<td>on coke</td>
</tr>
<tr>
<td>soda in</td>
<td>pop the</td>
<td>coke with</td>
</tr>
<tr>
<td>grape soda</td>
<td>pop it</td>
<td>my coke</td>
</tr>
</tbody>
</table>

Brice Russ | Dialect Geography on Twitter
### Disambiguation Through Collocation Groups

<table>
<thead>
<tr>
<th>soda</th>
<th>pop</th>
<th>coke</th>
</tr>
</thead>
<tbody>
<tr>
<td>a soda</td>
<td>to pop</td>
<td>diet coke</td>
</tr>
<tr>
<td>diet soda</td>
<td>pop up</td>
<td>and coke</td>
</tr>
<tr>
<td>soda and</td>
<td>a pop</td>
<td>a coke</td>
</tr>
<tr>
<td>of soda</td>
<td>pop in</td>
<td>coke and</td>
</tr>
<tr>
<td>orange soda</td>
<td>pop culture</td>
<td>of coke</td>
</tr>
<tr>
<td>and soda</td>
<td>pop music</td>
<td>the coke</td>
</tr>
<tr>
<td>baking soda</td>
<td>pop out</td>
<td>coke in</td>
</tr>
<tr>
<td>drink soda</td>
<td>and pop</td>
<td>coke is</td>
</tr>
<tr>
<td>soda is</td>
<td>the pop</td>
<td>&amp; coke</td>
</tr>
<tr>
<td>the soda</td>
<td>pop a</td>
<td>cherry coke</td>
</tr>
<tr>
<td>soda on</td>
<td>of pop</td>
<td>coke zero</td>
</tr>
<tr>
<td>cream soda</td>
<td>pop off</td>
<td>coke bottle</td>
</tr>
<tr>
<td>drinking soda</td>
<td>pop star</td>
<td>on coke</td>
</tr>
<tr>
<td>soda in</td>
<td>pop the</td>
<td>coke with</td>
</tr>
<tr>
<td>grape soda</td>
<td>pop it</td>
<td>my coke</td>
</tr>
</tbody>
</table>