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

User Profile

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

% of internet users in each group who use Twitter

All Internet Users	8%
Gender	
Men	7
Women	10
Age	
18-29	14
30-49	7
50-64	6
65+	4
Race/Ethnicity	
White, non-Hispanic	5
Black, non-Hispanic	13
Hispanic	18
Household Income	
Less than $30,000	10
$30,000-$49,999	6
$50,000-$74,999	10
$75,000+	6
Education level	
Less than High School	n/a
High School Diploma	5
Some College	9
College+	9

Social Patterns on Twitter

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

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

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

Toronto, ON	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.
Birmingham, AL	@mhirsh32 Would probably be opening a can of soda/ bottle of water, drinking a sip or two, then never touching it again. Still thinking.....
MIC CITY, TX	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!
Washington, DC	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
Secane, PA	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.
Dallas, TX	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.

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

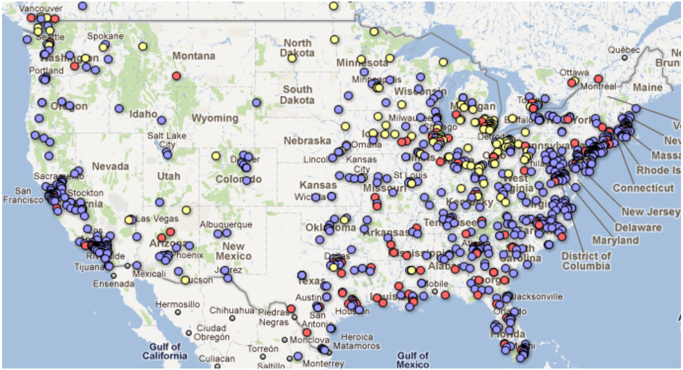
Mapped using Google Fusion Tables software

- 'soda'/'pop'/'coke'
- 'hella'/'very'
- 'needs X-ed'

Map Comparison: Soda vs. pop vs. coke

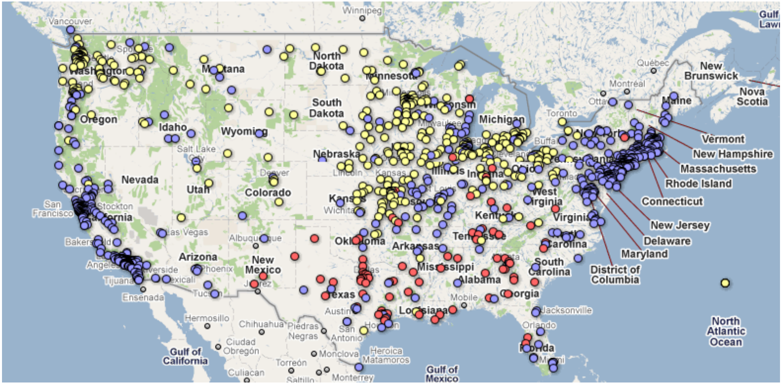
- Account for over 90% of soft drink variation (Vaux 2003)
  - 'Pop' predominant in Midwest to Pacific Northwest
  - 'Coke' predominant in the South (South Carolina to Texas)
  - 'Soda' used everywhere, but used exclusively in New England and Southwest

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



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"

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)

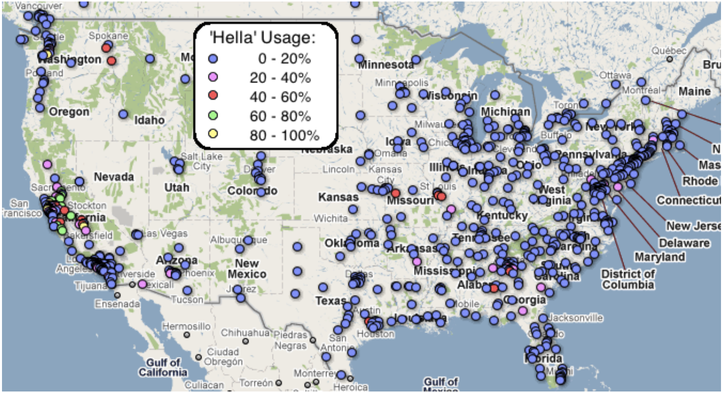
New Research: ‘hella’/‘very’

- ‘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}’)

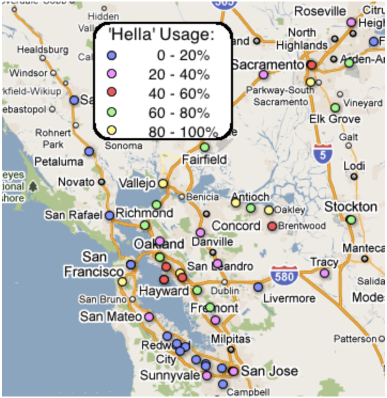
Over 300,000 data points:

(yellow = ‘very’; red = ‘hella’)

5-binned map


'Hella' Usage:
0 - 20%
20 - 40%
40 - 60%
60 - 80%
80 - 100%
Oregon
Idaho
Nevada
Utah
Kansas
Minnesota
Michigan
Kentucky
Texas
Maine
Rhode
Connecticut
Delaware
Maryland
District of Columbia
Gulf of California
Jacksonville
Houston
Tucson
Ensenada
Tijuana
Hermosillo

Silicon Valley speakers are hella standard



Silicon Valley speakers are hella standard

City	'very'	'hella'	% 'hella'
Mountain View, CA	317	3	**0.9%**
Santa Clara, CA	111	19	**14.6%**
San Jose, CA	768	367	**22.3%**
Sacramento, CA	1115	1262	**53.1%**
Oakland, CA	695	1307	**62.6%**
Vallejo, CA	70	374	**84.2%**
Columbus, OH	1483	105	6.6%

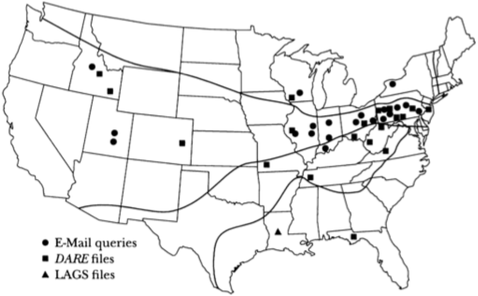
Comparison of *very*/*hella* usage in Northern California cities

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'

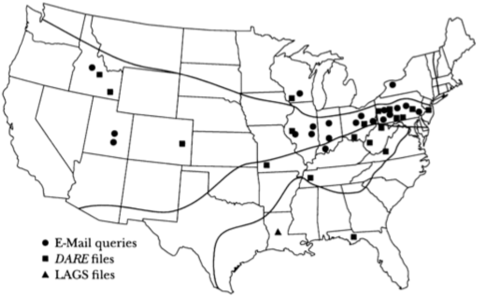
FIGURE 1
Informal Attestation of *need* + V-*en*



(from Murray et al. 1996)

Prior attestation of 'needs X-ed'

FIGURE 1
Informal Attestation of *need* + V-*en*



(from Murray et al. 1996)

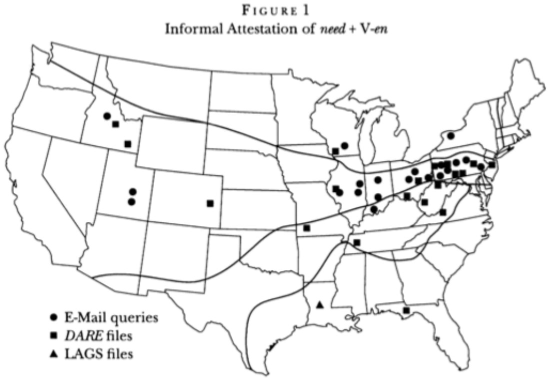
Selected verbs: 'done', 'fixed', 'fired', 'washed', 'filled'

6,406 data points, 1,884 locations

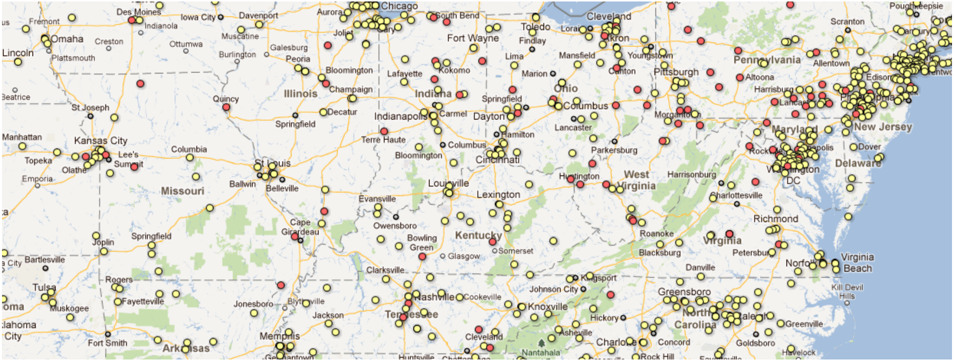
The 'needs' of the many...

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)

Conclusions

- 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

Conclusions

- 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

Conclusions

- 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!

References

Asur, S. and Huberman, B. (2010) Predicting the Future With Social Media. *Proceedings of the ACM International Conference on Web Intelligence*.

Bamman, D. (2011) lexicalist. Retrieved from http://www.lexicalist.com/.

Biber, D., Conrad, C. & Reppen, R. (1998) Corpus linguistics: investigating language structure and use. Cambridge, UK: Cambridge University Press.

Bucholtz, M,. Bermudez, N., Fung, V., Edwards, L., and Vargas, R. (2007). Hella Nor Cal or Totally So Cal? The Perceptual Dialectology of California. Journal of English Linguistics, 35(4), 325-352.

Bucholtz, M. (2007) Word Up: Social Meanings of Slang in California Youth Culture. In Monaghan, L., and Goodman, J. (Eds.), A cultural approach to interpersonal communication: essential readings (243 - 267). Malden, MA: Blackwell.

Eisenstein, J., O'Connor, B., Smith, N., & Xing, E. (2010) A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: 1277-1287.

Honeycutt, C., & Herring, S. (2009) *Proceedings of the Forty-Second Hawaii International Conference on System Sciences (HICSS-42)*. Los Alamitos, CA: IEEE Press.

References

Murray, T., Frazer, T., and Simon, B. (1996) Need + Past Participle in American English. American Speech, 71(3), 255-271.

O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, N. (2010) From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Paul, R. (2010, April 21) Tutorial: consuming Twitter's real-time stream API in Python. Retrieved from http://arstechnica.com/open-source/guides/2010/04/tutorial-use-twitters-new-real-time-stream-api-in-python.ars

Smith, A. (2011, November 15) Why Americans use social media. Retrieved from http://pewinternet.org/Reports/2011/Why-Americans-Use-Social-Media/Main-report.aspx.

Smith, A. & Rainie, L. (2010, December 8) Overview: The people who use Twitter. Retrieved from http://pewinternet.org/Reports/2010/Twitter-Update-2010/Findings/Overview.aspx.

Ulrey, K.S. (2009). Dinner Needs Cooked, Groceries Need Bought, Diapers Need Changed, Kids Need Bathed: Tracking The Progress Of Need + Past Participle Across The United States. (Unpublished masters' thesis.) Ball State University: Muncie, IN.

Vaux, B. (2003) Harvard Survey of North American Dialects. Retrieved from http://www4.uwm.edu/FLL/linguistics/dialect/index.html.

Contact

rbruss@ling.osu.edu
Twitter: @kilroywashere
Maps and script available at:
http://www.briceruss.com/ADStalk

Is Coke It?

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)

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

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")

Disambiguation Through Collocation Groups

soda		pop		coke	
a soda	770	to pop	3397	diet coke	1482
diet soda	637	pop up	2961	and coke	1030
soda and	576	a pop	1748	a coke	872
of soda	401	pop in	1362	coke and	700
orange soda	363	pop culture	1254	of coke	577
and soda	332	pop music	1240	the coke	332
baking soda	319	pop out	1042	coke in	250
drink soda	293	and pop	820	coke is	219
soda is	284	the pop	787	& coke	214
the soda	256	pop a	781	cherry coke	211
soda on	224	of pop	749	coke zero	182
cream soda	219	pop off	649	coke bottle	160
drinking soda	219	pop star	509	on coke	147
...	...	pop the	501	coke with	145
...	...	pop it	479	my coke	133

Disambiguation Through Collocation Groups

soda		pop		coke	
a soda	770	to pop	3397	diet coke	1482
diet soda	637	pop up	2961	and coke	1030
soda and	576	*a pop*	1748	*a coke*	872
of soda	401	pop in	1362	coke and	700
orange soda	363	pop culture	1254	of coke	577
and soda	332	pop music	1240	the coke	332
baking soda	319	pop out	1042	coke in	250
drink soda	293	and pop	820	coke is	219
soda is	284	the pop	787	& coke	214
the soda	256	pop a	781	cherry coke	211
soda on	224	of pop	749	coke zero	182
cream soda	219	pop off	649	coke bottle	160
drinking soda	219	pop star	509	on coke	147
soda in	212	pop the	501	coke with	145
grape soda	211	pop it	479	my coke	133

Disambiguation Through Collocation Groups

soda		pop		coke	
a soda	770	to pop	3397	diet coke	1482
diet soda	637	pop up	2961	and coke	1030
soda and	576	a pop	1748	a coke	872
of soda	401	pop in	1362	coke and	700
orange soda	363	pop culture	1254	of coke	577
and soda	332	pop music	1240	the coke	332
baking soda	319	pop out	1042	coke in	250
drink soda	293	and pop	820	coke is	219
soda is	284	the pop	787	& coke	214
the soda	256	pop a	781	cherry coke	211
soda on	224	of pop	749	coke zero	182
cream soda	219	pop off	649	coke bottle	160
drinking soda	219	pop star	509	on coke	147
soda in	212	pop the	501	coke with	145
grape soda	211	pop it	479	my coke	133