Finding more needles by building bigger haystacks

Size and specificity in big data sociolinguistics

Jacob Eisenstein
@jacobeisenstein

Georgia Institute of Technology

September 22, 2017

Sociolinguistics: Big questions, small data

- How does language vary across geographical and social groups?
- How does language change over time?
- How are language differences socially evaluated?

Labov's department store study

- **Linguistic variable**: (r) in fourth floor
- **Social variable**: class (department stores Klein's, Macy's, and Sak's)
- **Situational variable**: feigned misunderstanding

(Labov, 1972)

Labov's department store study

Use of local New York "r-less" variable decreases with

- socioeconomic status
- emphasis.

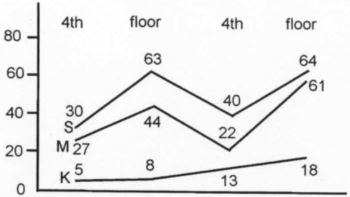


Figure 13.2: Percentage of all (r – 1) by store for four positions (S = Saks, M = Macy's, K = Kleins)

Geographical, socioeconomic, and stylistic variation are all linked!

Why it worked

- The department store study hinged on finding a needle in a multidimensional haystack:
  - (r) variable is ubiquitous
  - ...yet highly differentiated.
- This was only possible because of Labov's intuitive understanding of language and culture in New York.
- What sociolinguistic phenomena do we miss if our research methodology relies so heavily on experimenter intuition?

Outline

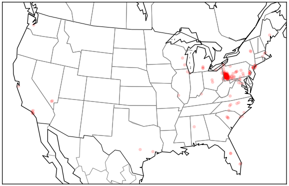
- Rare events and variable discovery
  - Linguistic variables
  - Intersectional social analysis
  - Change and influence
- More data $\rightarrow$ more intractable annotation problems?
  - Meaningful markers of language change
  - Hate speech

Outline

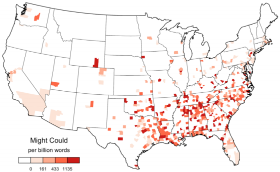
- **Rare events and variable discovery**
  - Linguistic variables
  - Intersectional social analysis
  - Change and influence
- More data $\rightarrow$ more intractable annotation problems?
  - Meaningful markers of language change
  - Hate speech

Rare linguistic events on Twitter

yinz: 3 per million tweets (Eisenstein et al., 2014)

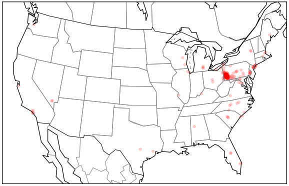


might could: $<$ 10 per million tweets (Grieve et al., 2017)

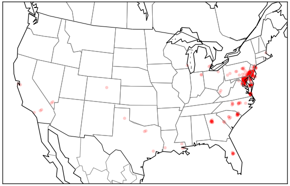


Discovering new linguistic variables

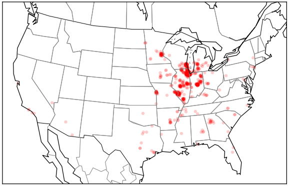
yinz



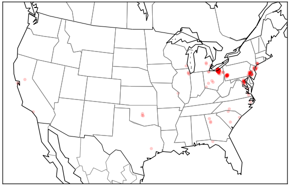
ard ("alright")



lbvs ("laughing but very serious")



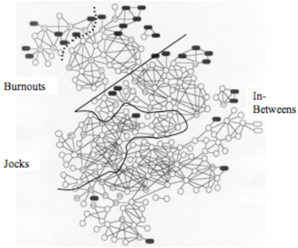
ctfu ("cracking the fuck up")



(Eisenstein et al., 2010; Nguyen & Eisenstein, 2017)

Social variables

- Early studies focused on big "demographic" variables: race, gender, age
- But the real action is:
    - At the **intersections** between social variables (Bucholtz, 2003)
    - In **locally-defined** social categories (Eckert, 2000)



Age	Sex	New York	Dallas
0-17	F	niall, ilysm, hemmings, stalk, ily	fanuary, idol, lmbo, lowkey, jonas
	M	ight, technique, kisses, lesbian, dicks	homies, daniels, oomf, teenager, brah
18-29	F	roses, castle, hmmmm, chem, sinking	socially, coma, hubby, bra, swimming
	M	drunken, manhattan, spoiler, guardians, gonna	harden, watt, astros, rockets, mavs
30-39	F	suite, nyc, colleagues, york, portugal	astros, sophia, recommendations, houston
	M	mets, effectively, cruz, founder, knicks	texans, rockets, embarrassment, tcu, mississippi
40+	F	cultural, affected, encouraged, proverb, unhappy	determine, islam, rejoice, psalm, responsibility
	M	reuters, investors, shares, lawsuit, theaters	mph, wazers, houston, tx, harris

(Pavalanathan & Eisenstein, 2015)

Age	Sex	New York	Dallas
0-17	F	niall, ilysm, hemmings, stalk, ily	fanuary, idol, lmbo, lowkey, jonas
	M	ight, technique, kisses, lesbian, dicks	homies, daniels, oomf, teenager, brah
18-29	F	roses, castle, hmmmm, chem, sinking	socially, coma, hubby, bra, swimming
	M	drunken, manhattan, spoiler, guardians, gonna	harden, watt, astros, rockets, mavs
30-39	F	suite, nyc, colleagues, york, portugal	astros, sophia, recommendations, houston
	M	mets, effectively, cruz, founder, knicks	texans, rockets, embarrassment, tcu, mississippi
40+	F	cultural, affected, encouraged, proverb, unhappy	determine, islam, rejoice, psalm, responsibility
	M	reuters, investors, shares, lawsuit, theaters	mph, wazers, houston, tx, harris

(Pavalanathan & Eisenstein, 2015)

Age	Sex	New York	Dallas
0-17	F	niall, ilysm, hemmings, stalk, ily	fanuary, idol, lmbo, lowkey, jonas
	M	ight, technique, kisses, lesbian, dicks	homies, daniels, oomf, teenager, brah
18-29	F	roses, castle, hmmmm, chem, sinking	socially, coma, hubby, bra, swimming
	M	drunken, manhattan, spoiler, guardians, gonna	harden, watt, astros, rockets, mavs
30-39	F	suite, nyc, colleagues, york, portugal	astros, sophia, recommendations, houston
	M	mets, effectively, cruz, founder, knicks	texans, rockets, embarrassment, tcu, mississippi
40+	F	cultural, affected, encouraged, proverb, unhappy	determine, islam, rejoice, psalm, responsibility
	M	reuters, investors, shares, lawsuit, theaters	mph, wazers, houston, tx, harris

(Pavalanathan & Eisenstein, 2015)

Age	Sex	New York	Dallas
0-17	F	niall, ilysm, hemmings, stalk, ily	fanuary, idol, lmbo, lowkey, jonas
	M	ight, technique, kisses, lesbian, dicks	homies, daniels, oomf, teenager, brah
18-29	F	roses, castle, hmmmm, chem, sinking	socially, coma, hubby, bra, swimming
	M	drunken, manhattan, spoiler, guardians, gonna	harden, watt, astros, rockets, mavs
30-39	F	suite, nyc, colleagues, york, portugal	astros, sophia, recommendations, houston
	M	mets, effectively, cruz, founder, knicks	texans, rockets, embarrassment, tcu, mississippi
40+	F	cultural, affected, encouraged, proverb, unhappy	determine, islam, rejoice, psalm, responsibility
	M	reuters, investors, shares, lawsuit, theaters	mph, wazers, houston, tx, harris

(Pavalanathan & Eisenstein, 2015)

Discovering social variables

Discovering social variables

K-means clustering on Twitter timelines

	% Women
hubs blogged recipe fabric	90%
kidd hubs xo =]	80%
wyd #oomf lmbo shyt	60%
n_ggas wyd finna shyt	26%
#nhl #bruins #mlb knicks	11%

- Do women use more standard language?
- Is men's writing more "informational"?

(Bamman et al., 2014)

Discovering social variables

K-means clustering on Twitter timelines

	% Women
hubs blogged recipe fabric	90%
kidd hubs xo =]	80%
wyd #oomf lmbo shyt	60%
n_ggas wyd finna shyt	26%
#nhl #bruins #mlb knicks	11%

- **Do women use more standard language?**
- Is men's writing more "informational"?

(Bamman et al., 2014)

Discovering social variables

K-means clustering on Twitter timelines

	% Women
hubs blogged recipe fabric	90%
kidd hubs xo =]	80%
wyd #oomf lmbo shyt	60%
n_ggas wyd finna shyt	26%
#nhl #bruins #mlb knicks	11%

- **Do women use more standard language?**
- Is men's writing more "informational"?

(Bamman et al., 2014)

Discovering social variables

K-means clustering on Twitter timelines

	% Women
hubs blogged recipe fabric	90%
kidd hubs xo =]	80%
wyd #oomf lmbo shyt	60%
n_ggas wyd finna shyt	26%
#nhl #bruins #mlb knicks	11%

- Do women use more standard language?
- **Is men's writing more "informational"?**

(Bamman et al., 2014)

Discovering social variables

K-means clustering on Twitter timelines

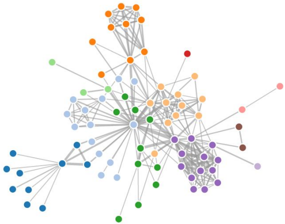
	% Women
hubs **blogged recipe fabric**	90%
kidd hubs xo =]	80%
wyd #oomf lmbo shyt	60%
n_ggas wyd finna shyt	26%
#nhl #bruins #mlb knicks	11%

- Do women use more standard language?
- **Is men's writing more "informational"?**

(Bamman et al., 2014)

How does language change?

- **Exposure**: To use a new word, you must be exposed to it.
- **Influence**: The *choice* of whether to use a new word is socially motivated.



The trajectory of language change is thus shaped by social network structures and social evaluation (Goel et al., 2016).

Whodunnit?

- Person i first uses a new linguistic feature at time t. **Who is responsible?**

Whodunnit?

- Person i first uses a new linguistic feature at time t. **Who is responsible?**
- Suspect j should have the following properties:
  - Used the same feature at a time $t' < t$
  - Likely to be observed by i

Whodunnit?

- Person i first uses a new linguistic feature at time t. **Who is responsible?**
- Suspect j should have the following properties:
  - Used the same feature at a time $t' < t$
  - Likely to be observed by i
- A random sample does not suffice!
  - Number of influence events decrease with the **square** of the sampling rate.
  - Might miss j and mistakenly blame k, who used the feature at time $t'' < t'$

Language change as a networked cascade

Social network

Bart	Lisa
Bart	Milhouse
Lisa	Homer
Homer	Barney
...	...

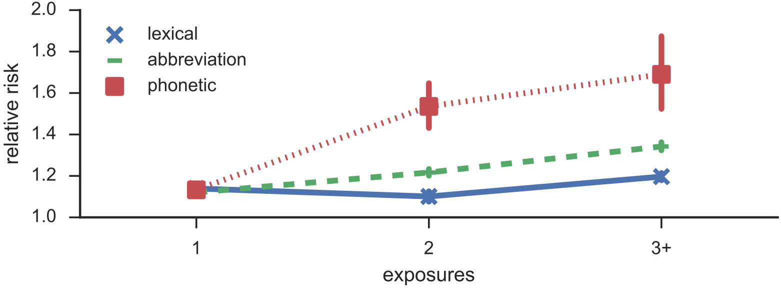
Locations

Bart	Los Angeles
Milhouse	Los Angeles
Lisa	Atlanta
Homer	Chicago
...	...

Language

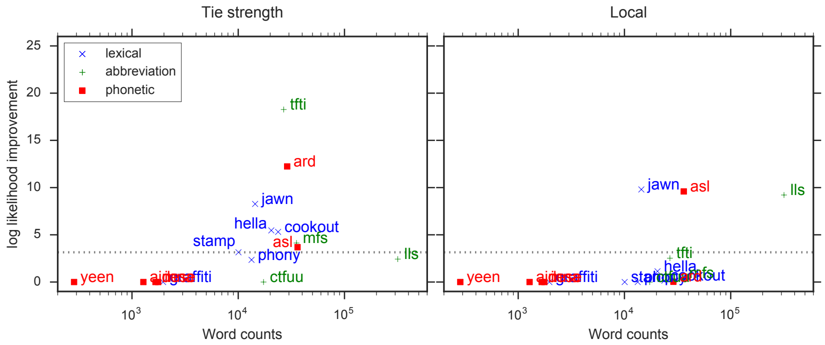
Bart	jawn	Feb 1, 2013, 13:45
Milhouse	jawn	Feb 1, 2013, 13:50
Homer	hella	Feb 1, 2013, 18:15
Bart	lls	Feb 2, 2013, 07:30
Milhouse	lls	Feb 2, 2013, 07:40
...	...	...

Language change as epidemic



- Relative risk: likelihood of infection given exposure, normalized against rate in randomly-rewired network.
- Rel. risk > 1: evidence of non-random contagion.
- For phonetic variables, risk increases with multiple exposures, a characteristic of complex contagion.

The role of tie strength



- We estimate a Hawkes process model of the spread of new words over time.
- Modeling **tie strength** improves fit for many words, suggesting that language change spreads over strong ties.

Outline

- **Rare events and variable discovery**
  - Linguistic variables
  - Intersectional social analysis
  - Change and influence
- More data $\rightarrow$ more intractable annotation problems?
  - Meaningful markers of language change
  - Hate speech

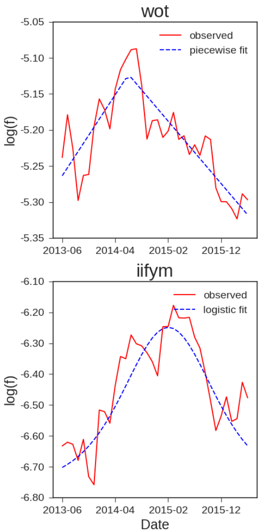
Outline

- Rare events and variable discovery
  - Linguistic variables
  - Intersectional social analysis
  - Change and influence
- **More data $\rightarrow$ more intractable annotation problems?**
  - Meaningful markers of language change
  - Hate speech

Which innovations succeed?

Most innovations fail.

- What are the social characteristics of successful innovations (Altmann et al., 2011)?
- How does the linguistic system constrain the field of possibilities for successful innovation (Stewart & Eisenstein, 2017)?



Finding (attempted) innovations

1. Design characteristic models of innovation success and failure.
   - continuous growth
   - piecewise linear growth and decline
   - logistic distribution

Finding (attempted) innovations

1. Design characteristic models of innovation success and failure.
   - continuous growth
   - piecewise linear growth and decline
   - logistic distribution
2. Find the top 5% words that fit each model.
   - Some of these terms are linguistic innovations
   - ... but others are names, dates, topics (2017, killary, drumpf, berniebot)

Finding (attempted) innovations

1. Design characteristic models of innovation success and failure.
   - continuous growth
   - piecewise linear growth and decline
   - logistic distribution
2. Find the top 5% words that fit each model.
   - Some of these terms are linguistic innovations
   - ... but others are names, dates, topics (2017, killary, drumpf, berniebot)
3. Manually remove terms that are topical rather than linguistic innovations.
   - We can do this because type-level annotation is cheap!

Results

Successful innovations are:

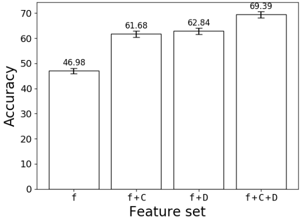
- Widely disseminated across authors, threads, and forums
- Usable in a wide range of linguistic contexts

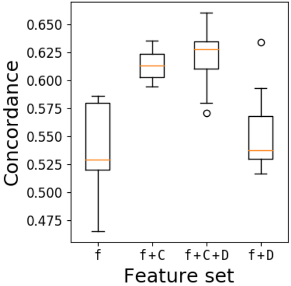
Results

Successful innovations are:

- Widely disseminated across authors, threads, and forums
- Usable in a wide range of linguistic contexts

These features make it possible to predict **which** words will succeed, and **when** the others will fail.





Hate speech on Reddit

- What is the effect of eliminating forums for hate speech?
  - Do forum participants export hate speech elsewhere?
  - Or does the elimination of the “echo chamber” reduce hate speech overall?
- In 2015, Reddit closed several forums for violations of its anti-harassment policy, including `r/CoonTown` and `r/FatPeopleHate`, putting this question to the test.

(Chandrasekharan et al., 2017)

Some examples

Some examples

- It would be so much easier if this [n-word] was taken outside and shot. Then rasslle up his eight or nine [kids] and shoot them so we can terminate that line of genes.
- You fucking fatass, you made the decision to be a fat fuck after you decided to stuff your fat fucking face instead of acting like a normal human being.

A day after the paper came out

46.8k

Reddit's bans of r/coontown and r/fatpeoplehate worked--many accounts of frequent posters on those subs were abandoned, and those who stayed reduced their use of hate speech ▸ comp.social.gatech.edu

5 days ago by asbruckman Professor | Interactive Computing x2

6874 comments share save hide give gold report

[–] Hey-Grandan2 349 points 5 days ago

What excacly qualifies for hate speech?

permalink embed save report give gold reply

[–] eegilbert **Author of Article** 652 points 5 days ago

One of the authors here. There was an unsupervised computational process used, documented on pages 6 and 7, and then a supervised human annotation step. Both lexicons are used throughout the rest of work.

permalink embed save parent report give gold reply

[+] ***Comment removed*** *5 days ago* (58 children)*

[–] Laminar_flo 92 points 5 days ago

Ok, adding to that, how did you ensure that the manual filtering process was ideological neutral and not just a reflection of the political sensitivities of the person filtering?

permalink save parent report give gold reply

[–] qwenjwenfljnanq 11 points 5 days ago

But then how did you differentiate between hate speech and people talking *about* hate speech?

permalink save parent report give gold reply

[–] Mode1961 -14 points 5 days ago

> number of words that indicate hate speech

Who choose those words.

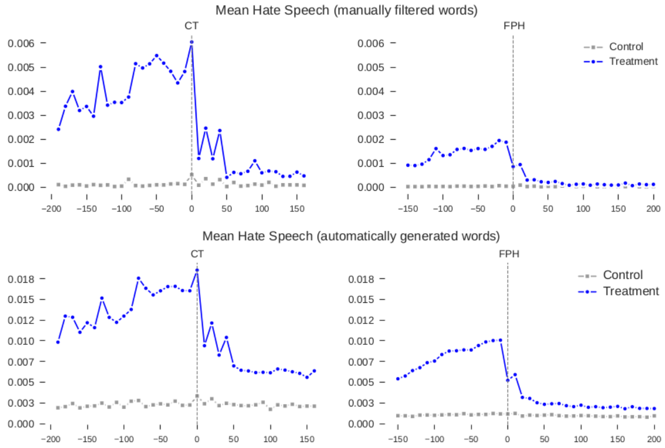
permalink save parent report give gold reply

“What excacly qualifies for hate speech?”

1. For each subreddit, run SAGE to identify words that are unusually frequent (Eisenstein et al., 2011).
2. Examine the top 100, manually remove words that are not intrinsically linked to racist / anti-fat discourse.
   - the forum itself: fph, ct
   - the act of posting offensive content: shitposting, shitlord
   - words frequently used in non-hate speech contexts: IQ, welfare, cellulite

We kept 20% of the original lexicon, $\kappa \approx .88$

Results with and without annotation



Control: forums with high cross-posting with `ct` and `fph`.

Outline

- Rare events and variable discovery
  - Linguistic variables
  - Intersectional social analysis
  - Change and influence
- More data $\rightarrow$ more intractable annotation problems?
  - Meaningful markers of language change
  - Hate speech

Some claims

- Rare events like lexical variables have much to teach us about sociolinguistics, but new methods are required.

Some claims

- Rare events like lexical variables have much to teach us about sociolinguistics, but new methods are required.
- Big data makes it easy to ask lazy questions about big social categories, but it also enables fine-grained intersectional analysis.

Some claims

- Rare events like lexical variables have much to teach us about sociolinguistics, but new methods are required.
- Big data makes it easy to ask lazy questions about big social categories, but it also enables fine-grained intersectional analysis.
- Longitudinal data offers a strikingly direct view of changes in progress, but automated analysis must be paired with manual annotation to ensure that the results are meaningful.

Some questions

- Big data requires automation, and automation implies errors.
- Some errors are more erroneous than others.
  - Missing 50% of hate speech markers at random $\rightarrow$ not so bad?
  - Missing an entire dialect of (((hate speech))) $\rightarrow$ not so good!
- Needed: rigorous methodologies for testing for (and correcting!) bias in automated big data analysis, and for iterating on variable discovery.

Acknowledgments

- The organizers for this event!
- **Students**: Eshwar Chandrasekharan, Rahul Goel, Ioannis Paparrizos, Umashanthi Pavalanathan, Sandeep Soni
- **Collaborators**: Fernando Diaz, Eric Gilbert, Adam Glynn, Hanna Wallach
- **Sponsors**: National Science Foundation, National Institutes for Health, Air Force Office of Scientific Research

References I

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PloS one*, *6*(5), e19009.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160.

Bucholtz, M. (2003). Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics*, *7*(3), 398–416.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The effectiveness of reddit's 2015 ban through the lens of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, *1*.

Eckert, P. (2000). *Linguistic variation as social practice*. Blackwell.

Eisenstein, J., Ahmed, A., & Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, (pp. 1041–1048).

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 1277–1287).

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE*, *9*.

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., & Eisenstein, J. (2016). The social dynamics of language change in online networks. In *The International Conference on Social Informatics (SocInfo)*.

Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in modern american english online. *English Language & Linguistics*, *21*(1), 99–127.

Labov, W. (1972). The social stratification of (r) in new york city department stores. In *Sociolinguistic Patterns* (pp. 43–54). University of Pennsylvania of Press.

Nguyen, D. & Eisenstein, J. (2017). A kernel independence test for geographical language variation. *Computational Linguistics*, *in press*.

Pavalanathan, U. & Eisenstein, J. (2015). Confounds and consequences in geotagged twitter data. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Stewart, I. & Eisenstein, J. (2017). Making "fetch" happen: The influence of social and linguistic context on the success of lexical innovations. *Transactions of the ACL*, *in review*.