Netspeak: Dialect, Genre, Register?

Jacob Eisenstein
@jacobbeisenstein

Georgia Institute of Technology

April 17, 2015
Variation by speaker and by medium

Dialect variation
geography, ethnicity, class, gender, ...
Variation by speaker and by medium

<table>
<thead>
<tr>
<th>Standard writing</th>
<th>Variation by Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netspeak</td>
<td></td>
</tr>
</tbody>
</table>

Variation by speaker and by medium
Variation by speaker and by medium

Dialect variation
geography, ethnicity, class, gender, ...

Netspeak

Standard writing

Variation by Medium
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- How can we use large digital corpora to automatically induce dialect difference?
- Why would “netspeak” features vary with geography?
A landscape of digital communication

Instant messaging
Tagliamonte and Denis 2008

Email
Baron 1998

Text messages
Ling 2005
Anis 2007

Chatrooms
Paolillo 1999

Twitter
Eisenstein et al 2010
Zappavigna 2012
Doyle 2014

Blogs, Forums,
Wikipedia
Herring and Paolillo 2006
Androutsopoulos 2007
Scherrer and Rambow 2010

more private

more public
Twitter

- 140-character messages
- Each user has a custom **timeline** of people they’ve chosen to **follow**.
- Most data is publicly accessible, and social network and geographical metadata is available.
Who are these people?

(Pew Research Center)

<table>
<thead>
<tr>
<th>Category</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>All internet users</td>
<td>18%</td>
<td>23%</td>
</tr>
<tr>
<td>Men</td>
<td>17</td>
<td>24*</td>
</tr>
<tr>
<td>Women</td>
<td>18</td>
<td>21*</td>
</tr>
<tr>
<td>White, Non-Hispanic</td>
<td>16</td>
<td>21*</td>
</tr>
<tr>
<td>Black, Non-Hispanic</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Hispanic</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>18-29</td>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>30-49</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>50-64</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>65+</td>
<td>5</td>
<td>10*</td>
</tr>
<tr>
<td>High school grad or less</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Some college</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>College+ (n=685)</td>
<td>18</td>
<td>30*</td>
</tr>
<tr>
<td>Less than $30,000/yr</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>$30,000-$49,999</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>$50,000-$74,999</td>
<td>15</td>
<td>27*</td>
</tr>
<tr>
<td>$75,000+</td>
<td>19</td>
<td>27*</td>
</tr>
<tr>
<td>Urban</td>
<td>18</td>
<td>25*</td>
</tr>
<tr>
<td>Suburban</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>Rural</td>
<td>11</td>
<td>17</td>
</tr>
</tbody>
</table>

▶ % of online adults who use Twitter; per-message statistics will differ.

▶ Representativeness concerns are real, but there are potential solutions.

▶ Social media has important representativeness advantages too.
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- How can we use large digital corpora to automatically induce dialect difference?
- Why would “netspeak” features vary with geography?
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- How can we use large digital corpora to automatically induce dialect difference?
- Why would “netspeak” features vary with geography?
## Variation in digital writing

<table>
<thead>
<tr>
<th>Linguistic variables</th>
<th>Social variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexical items from speech</td>
<td>geography</td>
</tr>
</tbody>
</table>
Yinz

- 2nd-person pronoun
- Western Pennsylvania
- Very rare: appears in 535 of $10^8$ messages
Yall

- 2nd-person pronoun
- Southeast, African-American English
- Once per 250 messages
Hella

- Intensifier, e.g.
  i got hella nervous
- Northern California
- Once per 1000 messages
Jawn

- Noun, diffuse semantics
- Philadelphia, hiphop\(^2\)
- Once per 1000 messages

- @user ok u have heard this jawn right
- i did wear that jawn but it was kinda warm this week
Summary of spoken dialect terms

<table>
<thead>
<tr>
<th></th>
<th>rate</th>
<th>region</th>
</tr>
</thead>
<tbody>
<tr>
<td>yinz</td>
<td>200,000</td>
<td>mainly used in Western PA</td>
</tr>
<tr>
<td>yall</td>
<td>250</td>
<td>ubiquitous</td>
</tr>
<tr>
<td>hella</td>
<td>1000</td>
<td>ubiquitous, but more frequent in Northern California</td>
</tr>
<tr>
<td>jawn</td>
<td>1000</td>
<td>mainly used in Philadelphia</td>
</tr>
</tbody>
</table>

- Overall: mixed evidence for spoken language dialect variation in Twitter.
- But are these the right words?
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- How can we use large digital corpora to automatically induce dialect difference?
- Why would “netspeak” features vary with geography?
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- **How can we use large digital corpora to automatically induce dialect difference?**
- Why would “netspeak” features vary with geography?
Measuring regional specificity

Per region $r$,

- **Difference** in frequencies, $f_{i,r} - f_i$

<table>
<thead>
<tr>
<th>word</th>
<th>$c_{SF}$</th>
<th>$c_{USA}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>11914</td>
<td>907185</td>
</tr>
<tr>
<td>hella</td>
<td>398</td>
<td>3332</td>
</tr>
<tr>
<td>!</td>
<td>3677</td>
<td>276604</td>
</tr>
<tr>
<td>,</td>
<td>4898</td>
<td>382834</td>
</tr>
<tr>
<td>san</td>
<td>172</td>
<td>1654</td>
</tr>
<tr>
<td>?</td>
<td>2435</td>
<td>185482</td>
</tr>
<tr>
<td>for</td>
<td>2868</td>
<td>221142</td>
</tr>
<tr>
<td>sf</td>
<td>133</td>
<td>364</td>
</tr>
<tr>
<td>on</td>
<td>2523</td>
<td>194203</td>
</tr>
<tr>
<td>the</td>
<td>7901</td>
<td>630319</td>
</tr>
</tbody>
</table>
Measuring regional specificity

Per region $r$,

- **Difference** in frequencies, $f_{i,r} - f_i$
- **Log-ratio** in frequencies, $\log f_{i,r} - \log f_i = \log \frac{f_{i,r}}{f_i}$

<table>
<thead>
<tr>
<th>word</th>
<th>$CSF$</th>
<th>$C_{USA}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#mattomil</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>#bart</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>#davidlyons</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>cost=</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>#io14</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>#know14</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>haight</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>#gdc2014</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>#prejudices</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>muni</td>
<td>16</td>
<td>17</td>
</tr>
</tbody>
</table>
Measuring regional specificity

Per region $r$,

- **Difference** in frequencies, $f_{i,r} - f_i$
- **Log-ratio** in frequencies, $\log f_{i,r} - \log f_i = \log \frac{f_{i,r}}{f_i}$
- **Regularized**
  maximum-likelihood estimate

$$\hat{\eta}_r = \arg \max_\eta \log P(\text{counts} \mid \eta; f) - \lambda |\eta|$$

<table>
<thead>
<tr>
<th>word</th>
<th>$c_{SF}$</th>
<th>$c_{USA}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bart</td>
<td>52</td>
<td>98</td>
</tr>
<tr>
<td>#sfgiants</td>
<td>56</td>
<td>138</td>
</tr>
<tr>
<td>francisco</td>
<td>91</td>
<td>235</td>
</tr>
<tr>
<td>sf</td>
<td>133</td>
<td>364</td>
</tr>
<tr>
<td>oakland</td>
<td>51</td>
<td>219</td>
</tr>
<tr>
<td>giants</td>
<td>51</td>
<td>334</td>
</tr>
<tr>
<td>warriors</td>
<td>46</td>
<td>368</td>
</tr>
<tr>
<td>bay</td>
<td>95</td>
<td>788</td>
</tr>
<tr>
<td>hella</td>
<td>398</td>
<td>3332</td>
</tr>
<tr>
<td>fasho</td>
<td>38</td>
<td>344</td>
</tr>
</tbody>
</table>
Discovering words

- **New York**: flatbush, baii, brib, bx, staten, mta, odee, soho, deadass, werd
- **Los Angeles**: pasadena, venice, anaheim, dodger, disneyland, angeles, compton, ucla, dodgers, melrose
- **Chicago**: #chicago, lbvs, chicago, blackhawks, #bears, #bulls, mfs, cubs, burbs, bogus
- **Philadelphia**: jawn, ard, #phillies, sixers, phils, wawa, philadelphia, delaware, philly, phillies

place names  entities  words
ard

alternative spelling for alright

► @name ard let me kno
► lol u’ll be ard
laughing but very serious

- i wanna rent a hotel room just to swim lbvs
- tell ur momma 2 buy me a car lbvs
odee

intensifier, related to overdose or overdone

- i’m odee sleepy
- she said she odee miss me
- its rainin odee :(
emoticon indicating mild annoyance

- ___ -

flight delayed -___- just what i need
Variation in digital writing

Linguistic variables

▶ lexical items from speech

Social variables

▶ geography

Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors. The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.
Variation in digital writing

Linguistic variables

- lexical items from speech
- novel orthographies

Social variables

- geography

Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors. The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.
Phonologically-motivated variables

-t,-d deletion jús, ol
th-stopping dis, doe
r-lessness togetha, neva, lawd, yaself, shawty
vowels tha (the), mayne (man), bruh, brah (bro)
relaxed pronunciations prolly, aight
“allegro spellings”⁵ gonna, finna, fitna, bouta, tryna, iono
In speech, “g” is deleted more often from verbs. Does this syntactic conditioning transfer to writing?
G-deletion

- In speech, “g” is deleted more often from verbs. **Does this syntactic conditioning transfer to writing?**
- Corpus: 120K tokens of top 200 unambiguous -ing words (ex. king, thing, sing)
- Part-of-speech tags from CMU Twitter tagger.\(^6\)
G-deletion: type-level analysis

(Colored by most common POS tag)
## G-deletion: logistic regression

<table>
<thead>
<tr>
<th>Category</th>
<th>Log odds</th>
<th>%</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>.227</td>
<td>.200</td>
<td>89,173</td>
</tr>
<tr>
<td>Noun</td>
<td>-.013</td>
<td>.083</td>
<td>18,756</td>
</tr>
<tr>
<td>Adjective</td>
<td>-.213</td>
<td>.149</td>
<td>4,964</td>
</tr>
<tr>
<td>monosyllable</td>
<td>-2.57</td>
<td>.001</td>
<td>108,804</td>
</tr>
<tr>
<td>Total</td>
<td>.178</td>
<td></td>
<td>112,893</td>
</tr>
</tbody>
</table>
G-deletion: logistic regression

<table>
<thead>
<tr>
<th></th>
<th>Log odds</th>
<th>%</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>.227</td>
<td>.200</td>
<td>89,173</td>
</tr>
<tr>
<td>Noun</td>
<td>-.013</td>
<td>.083</td>
<td>18,756</td>
</tr>
<tr>
<td>Adjective</td>
<td>-.213</td>
<td>.149</td>
<td>4,964</td>
</tr>
<tr>
<td>monosyllable</td>
<td>-2.57</td>
<td>.001</td>
<td>108,804</td>
</tr>
<tr>
<td>High Euro-Am county</td>
<td>-.194</td>
<td>.117</td>
<td>28,017</td>
</tr>
<tr>
<td>High Afro-Am county</td>
<td>.145</td>
<td>.241</td>
<td>27,022</td>
</tr>
<tr>
<td>High pop density county</td>
<td>.055</td>
<td>.228</td>
<td>27,773</td>
</tr>
<tr>
<td>Low pop density county</td>
<td>-.017</td>
<td>.144</td>
<td>28,228</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>.178</strong></td>
<td></td>
<td><strong>112,893</strong></td>
</tr>
</tbody>
</table>
Variation in digital writing

Linguistic variables
- lexical items from speech
- novel orthographies

Social variables
- geography

Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors. The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.
Variation in digital writing

Linguistic variables
- lexical items from speech
- novel orthographies
- phonetically-motivated spellings

Social variables
- geography
- demographics

Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors. The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.
Two broad categories of variables

1. Imported from speech
   - Lexical variables (jawn, hella)
   - Phonologically-inspired variation (-g and -t,-d deletion)
   - These variables bring traces of their social and linguistic properties from speech.

2. Endogenous to digital writing
   - Abbreviations (ard, lbvs, odee, ctfu, asl, ...)
   - Emoticons (-____-)
   - Why should these vary with geography?
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- **How can we use large digital corpora to automatically induce dialect difference?**
- Why would “netspeak” features vary with geography?
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- How can we use large digital corpora to automatically induce dialect difference?
- Why would “netspeak” features vary with geography?
Language variation in a social network

- Linguistic innovations diffuse to new authors through social networks.
- In Twitter, 97% of strong ties are geographically local.
- Does this explain geographical variation in netspeak?
Broadcast
UNICEF @UNICEF · 2h
#Ebola survivor Isata, 6, was cleared with 43 others yesterday at treatment centre in #SierraLeone. Via @UNICEFsi

Hashtag-initial
Oprah Winfrey @Oprah · Oct 4
@tylerperry please watch #ylanlafixmylife tonight 9 eastern.. Wanna know what you think.

Addressed
Audience size

- #hashtag
- @mention
- broadcast

larger intended audience
Small audience $\rightarrow$ more local language

Logistic regression from context to local netspeak:

![Graph showing logistic regression results for inhibiting and encouraging local variables. The graph includes the following points:
- @-Init: Inhibit local variables -0.25, Encourage local variables 0.25, coefficient 0.3340 (0.018)***
- @-Internal: Inhibit local variables 0.5, Encourage local variables 0.75, coefficient 0.0437 (0.017)***
- #-Init: Inhibit local variables 0, Encourage local variables 0.25, coefficient -0.4037 (0.025)***
- #-Internal: Inhibit local variables 0.25, Encourage local variables 0.5, coefficient -0.2821 (0.009)***]
Local audience → more local language

Logistic regression from context to local netspeak:

Do people know that these words are local, and deliberately use them with local audiences?
Local audience → netspeak language?

Logistic regression from context to non-local netspeak: e.g., lol, im, lmao, ya, haha ...
Local audience → netspeak language?

Logistic regression from context to non-local netspeak: e.g., lol, im, lmao, ya, haha ...

![Graph showing logistic regression results with non-standard words used more often with local audiences, even when the word itself is not local.](image-url)
Local audience $\rightarrow$ netspeak language?

Logistic regression from context to non-local netspeak: e.g., lol, im, lmao, ya, haha ...

Non-standard words are used more often with local audiences, even when the word itself is not local.
Questions for this talk

- Are spoken dialect features transcribed in digitally-mediated writing?
- How can we use large digital corpora to automatically induce dialect difference?
- Why would “netspeak” features vary with geography?
Summary

- Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors.
- The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.

@BarackObama • Feb 12
Speaking of #YOLO: ofa.bo/h2sp
Summary

- Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors.
- The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.

Thanks to my collaborators David Bamman, Umashanthi Pavalanathan, Tyler Schnoebelen, and to support from the National Science Foundation.


Local audience → more local language

Do people know that these words are local, and deliberately use them with local audiences?
Methodological pros and cons

- Orders of magnitude more data
  (enabling the study of rare linguistic phenomena and the \textit{induction} of unknown variables)
- Biased, non-representative population sample
- Informal communication, outside the interview setting
- Metadata on location, time, and social networks