Challenges of studying and processing dialects in social media

Anna Katrine Jørgensen, Dirk Hovy, and Anders Søgaard
University of Copenhagen
Njalsgade 140
DK-2300 Copenhagen S
soegaard@hum.ku.dk

Abstract

Dialect features typically do not make it into formal writing, but flourish in social media. This enables large-scale variational studies. We focus on three phonological features of African American Vernacular English and their manifestation as spelling variations on Twitter. We discuss to what extent our data can be used to falsify eight sociolinguistic hypotheses. To go beyond the spelling level, we require automatic analysis such as POS tagging, but social media language still challenges language technologies. We show how both newswire- and Twitter-adapted state-of-the-art POS taggers perform significantly worse on AAVE tweets, suggesting that large-scale dialect studies of language variation beyond the surface level are not feasible with out-of-the-box NLP tools.

1 Introduction

Dialectal and sociolinguistic studies are traditionally based on interviews of small sets of speakers of each variety. The *Atlas of North American English* (Labov et al., 2005) has been the reference point for American dialectology since its completion, but is based on only 762 speakers. Dallas is represented by four subjects, the New York City dialect by six, etc. Data is costly to collect, and, as a consequence, scarce.

Written language was traditionally used for formal purposes, and therefore differed in style from colloquial, spoken language. However, with the rise of social media platforms and the vast production of user generated content, differences between written and spoken language diminish. A number of recent papers have explored social media with respect to sociolinguistic and dialectological questions (Rao et al., 2010; Eisenstein, 2013; Volkova et al., 2013; Doyle, 2014; Hovy et al., 2015; Volkova et al., 2015; Johannsen et al., 2015; Hovy and Søgaard, 2015; Eisenstein, to appear). Emails, chats and social media posts serve purposes similar to those of spoken language, and consequently, features of spoken language, such as interjections, ellipses, and phonological variation, have found their way into this type of written language. Our work differs from most previous approaches by investigating several phonological spelling correlates of a specific language variety.

The 284 million active users on Twitter post more than half a billion tweets every day, and some fraction of these tweets are geo-located. Eisenstein (2013) and Doyle (2014) studied the effect of phonological variation across the US on spelling in Twitter posts, and both found some evidence that dialectal phonological variation has a direct impact on spelling

Proceedings of the ACL 2015 Workshop on Noisy User-generated Text, pages 9–18,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics

on Twitter. Both authors note various methodological problems using Twitter as a source of evidence for dialectal and sociolinguistic studies, including what we refer to as USER POPULATION BIAS and TOPIC BIAS below.

In this paper, we collect Twitter data to test eight (8) research hypotheses originating in sociolinguistic studies of African-American Vernacular English (AAVE). The hypotheses relate to three phonological features of AAVE, namely derhotacization, interdental fricative mutation, and backing in /str/. Some of our findings shed an interesting light on existing hypotheses, but our main focus in this paper is to identify the methodological challenges in using social media for testing sociolinguistic hypotheses.

Almost all previous large-scale variational studies using social media have focused on *spelling variation* and *lexical markers* of dialect. Ours is no exception. However, dialectal variation also manifests itself at the morpho-syntactic level. To investigate this variation, we also annotate some data with part-of-speech (POS) tags, using two NLP systems. This approach reveals a severe methodological challenge: sentences containing AAVE features are associated with significant drops in tagger performance.

This result challenges large-scale variational studies on social media that require automated analyses. The observed drops in performance are prohibitive for studying syntactic and semantic variation, and we believe the NLP community should make an effort to provide better and more robust dialect-adapted models to researchers and industry interested in processing social media. The findings also raise the question of whether NLP technology systematically disadvantages groups of non-standard language users.

1.1 Contributions

- We identify eight (8) research hypotheses from the sociolinguistic literature. We test them in a study of the distribution of three phonological features typically associated with AAVE in Twitter data. We test the features' correlations with various demographic variables. Our results falsify the hypothesis that AAVE is male-dominated (but see §3.1).

- We identify five (5) methodological problems common to variational studies in social media and discuss to what extent they compromise the validity of results.

- Further, we show that state-of-the-art newswire and Twitter POS taggers perform much worse on tweets containing AAVE features. This suggests an additional limitation to large-scale sociolinguistic research using social media data, namely that it is hard to analyze variation beyond the lexical level with current tools.

1.2 Sociolinguistic hypotheses

AAVE is, in contrast to other North American dialects, not geographically restricted. Although variation in AAVE does exist, AAVE in urban settings has been established as a uniform system with suprasegmental norms (Ash and Myhill, 1986; Labov et al., 2005; Labov, 2006; Wolfram, 2004). This paper considers the following eight (8) hypotheses from the sociolinguistic literature about AAVE as a ethnolect:

H1: AAVE is an *urban* ethnolect (Rickford, 1999; Wolfram, 2004).

H2: AAVE features are more present in the Gulf states than in the rest of the United States (Rastogi et al., 2011).

H3: The likelihood of speaking AAVE correlates negatively with income and educational level, and AAVE is more frequently appropriated by men (Rickford, 1999; Rickford, 2010).

H4: Derhotacization is more frequent in African Americans than in European Americans (Labov et al., 2005; Rickford, 1999).

H5: Derhotacization is negatively correlated with income and educational level (Rickford, 1999).

H6: Interdental fricative mutation is more frequent in AAVE than in European American speech (Pollock et al., 1998; Thomas, 2007).

H7: Interdental fricative mutation is predominantly found in the Gulf states (Rastogi et al., 2011).

H8: Backing in /str/ (to /skr/) is unique to AAVE (Rickford, 1999; Thomas, 2007; Labov, 2006).

Hypotheses 1–8 are investigated by correlating the distribution of phonological variants in geo-located tweets with demographic information.

Our method is similar to those proposed by Eisenstein (2013) and Doyle (2014), lending statistical power to sociolinguistic analyses, and circumventing traditional issues with data collection such as the Observer's Paradox (Labov, 1972b; Meyerhof, 2006). Our work differs from previous work by studying phonological *rules* associated with specific dialects, as well as considering a wide range of actual sociolinguistic research hypotheses, but our main focus is the methodological problems doing this kind of work, as well as assessing the limitations of such work.

1.3 Methodological problems

One obvious challenge relating social media data to sociolinguistic studies is that there is generally not a one-to-one relationship between phonological variation and spelling variation. People, in other words, do not spell the way they pronounce. Eisenstein (2013) discusses this challenge ((1) WRITING BIAS), but shows that effects of the phonological environment carry over to social media, which he interprets as evidence that there is at least some causal link between pronunciation and spelling variation.

A related problem is that non-speakers of AAVE may cite known features of AAVE with specific purposes in mind. They may use it in citations, for example:

(1) My 5 year old sister texted me on my mums phone saying "why did you take a picher in da bafroom" lool okay b (Twitter, Feb 21 2015)

or in meta-linguistic discussions:

(2) Whenever I hear a black person inquire about the location of the "bafroom"... (Twitter, Jan 20 2015)

We refer to these phenomena as (2) META-USE BIAS. This bias is important with rare phenomena. With "bafroom", it seems that about 1 in 20 occurrences on Twitter are meta-uses. Meta-uses may also serve social functions. AAVE features are used as cultural markers by Latinos in North Carolina (Carter, 2013), for example.

Some of the research hypotheses considered (**H3** and **H5**) relate to demographic variables such as income and educational levels. While we do not have socio-economic information about the individual Twitter user, we can use the geo-located tweets to study the correlation between socio-economic variables and linguistic features at the level of cities or ZIP codes.[1]

Eisenstein et al. (2011) note that this level of abstraction introduces some noise. Since Twitter users do not form representative samples of the population, the mean income for a city or ZIP code is not necessarily the mean income for the Twitter users in that area. We refer to this problem as the (3) USER POPULATION BIAS.

Another serious methodological problem known as (4) GALTON'S PROBLEM (Naroll, 1961; Roberts and Winters, 2013), is the observation that cross-cultural associations are

[1] Unlike many others, we rely on physical locations rather than user-entered profile locations. See Graham et al. (2014) for discussion.

often explained by geographical diffusion. In other words, it is the problem of discriminating historical from functional associations in cross-cultural surveys. Briefly put, when we sample tweets and income-levels from US cities, there is little independence between the city data points. Linguistic features diffuse geographically and do not change at random, and we can therefore expect to see more spurious correlations than usual. Like with the famous example of chocolate and Nobel Prize winners, our positive findings may be explained by hidden background variables. A positive correlation between income-level and a phonological pattern may also have cultural, religious or geographical explanations.

Reasons to be less worried about GALTON'S PROBLEM in our case, include that a) we only consider standard hypotheses from the sociolinguistics literature and not a huge set of previously unexplored, automatically generated hypotheses, b) we sample data points at random from all across the US, giving us a very sparse distribution compared to country-level data, but more notably, c) location is an important, explicit variable in our study. GALTON'S PROBLEM is typically identified by clustering tests based on location (Naroll, 1961). Obviously, the phonological features considered here cluster geographically, as evidenced by our geographic correlations in Table 2, but since our studies explicitly test the influence of location, it is not the case for most of the hypotheses considered here that geographic diffusion is the underlying explanation for something else.

In §3, we discuss whether these four methodological problems compromise the validity of our findings. One other methodological problems that may be relevant for other studies of dialect in social media, is almost completely irrelevant for our study: It is often important to control for topic in dialectal and sociolinguistic studies (Bamman et al., 2014), e.g., when studying the lexical preferences of speakers of urban ethnolects. We call this problem (5) TOPIC BIAS. Using word pairs with equivalent meanings for our studies, we implicitly control for topic (but see §3.1).

Feature	Positive	Negative	Total count
/r/ → /Ø/ or /ə/	brotha	brother	9528
	foreva	forever	3673
	hea	here	4352
	lova	lover	1273
	motha	mother	4668
	ova	over	3441
	sista	sister	5325
	wateva	whatever	2974
	wea	where	5153
	total		40,387
/str/ → /skr/	skreet	street	1226
	skrong	strong	1629
	skrip	strip	1101
	total		3956
/ð/ → /d/ or /v/	brova	brother	3715
	dat	that	2610
	deez	these	4477
	dem	them	3645
	dey	they	2434
	dis	this	2135
	mova	mother	2462
	total		21,478
/θ/ → /t/ or /f/	mouf	mouth	3861
	nuffin	nothing	2861
	souf	south	1102
	teef	teeth	1857
	trough	through	2804
	trow	throw	1090
	total		13,575
All tweets			79,396

Table 1: Word pairs and counts

2 Data and Method

We focus on derhotacization, backing in /str/, and interdental fricative mutation. Specifically, we collect data to study the following four phonological variations (the latter two are both instances of interdental fricative mutation): *a*) derhotacization: /r/ → /Ø/ or /ə/, *b*) /str/ → /skr/, *c*) /ð/ → /d/ or /v/ and, *d*) /θ/ → /t/ or /f/.

In non-rhotic dialects, /r/ is either not pronounced or is approximated as a vocalization in the surface form, when /r/ is in a pre-vocalic position. This can result in an elongation of the preceding vowel or in an off-glide schwa /ə/, e.g., *guard* → /gɑ:d/, *car* → /kɑ:/, *fear* → /fiə/ (Thomas, 2007).

Backing in /skr/ denotes the substitution of /str/ for /skr/ in word-initial positions resulting in pronunciations such as /skrit/ for *street*, /skrɑŋ/ for *strong* and /skrɪp/ for *strip*. Backing in /str/ has been reported to be a unique feature in AAVE, as it is unheard in other North American dialects (Rickford, 1999; Labov, 1972a; Thomas, 2007).

The two interdental fricative mutations relate to substitutions of /ð/ and /θ/ by /d/, /v/ and /t/, /f/ in words such as *that* and *mother* or *nothing* and *with*. It has been reported that mutations of /ð/ and /θ/ are more common among African Americans than among European Americans and that the frequency of the mutations is inversely correlated with socio-economic levels and formality of speaking (Rickford, 1999).

We follow Eisenstein (2013) and Doyle (2014) in assuming that spelling variation may be a result of phonological differences and select 25 word pairs for our study (Tabel 1). For each word pair, we collect positive (e.g., ”skreet”) and negative occurrences (e.g., ”street”), resulting in a total number of 79,396 tweets. The word pairs were chosen based on the unambiguity, frequency and representability of the phonological variations. Uniquely, backing in /str/ is represented by three word pairs of high similarity, which is due to phonological restrictions on the variation of /str/ to /skr/ and to the fact that backing in /str/ is a very rare phenomena.

The Twitter data used in the experiments was gathered from May to August 2014 using TwitterSearch.[2] We only collected tweets with geo-locations in the contiguous United States, from users reporting to tweet in English, and which were also predicted to be in English using *langid.py*.[3] The demographic information was obtained from the 2012 American Community Survey from the United States Census Bureau, as was information about population sizes in US cities. We linked each tweet in our data to demographic information using the geo-coordinates of the tweet and its nearest city in the following way.

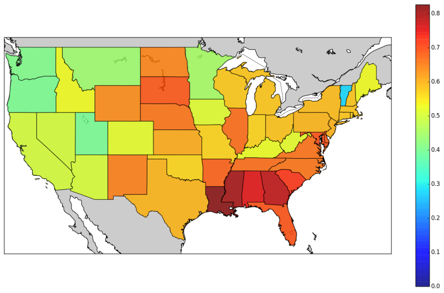

Figure 1: The ratio of AAVE examples over US states

For the 110 US cities of $\geq$ 200,000 inhabitants, we gathered information about: *a*) percentage high school graduates, *b*) percentage below poverty level, *c*) population size, *d*) median household income, *e*) percentage of males, *f*) percentage between 15 and 24 years old, *g*) percentage of African Americans and *h*) unemployment rate.

The overall geographical distribution of our data is shown in Figure 1. The map shows that we see more tweets with AAVE features in the Gulf states, in particular Louisiana, Mississippi and Georgia. This lends preliminary support to **H2**.

3 Results with phonological features

Occurrences of the phonological variations related to AAVE were correlated with the geographic and demographic variables using Spearman’s ρ (Table 2–3), at the level of individual tweets. From the correlation coefficients we see that the distributions of the three chosen AAVE rules are best explained by longitude, the distinction between the Gulf states and the rest of the US, and by the distribution

[2] https://pypi.python.org/pypi/TwitterSearch/

[3] https://pypi.python.org/pypi/langid

Feature	word pairs	male	black	15-24	citysize	highschool	income	poverty	unemployment
/r/ → /Ø/ or /ə/	brotha/brother	***	**	–	–	**	–	–	–
	foreva/forever	**	***	–	–	–	–	**	–
	hea/here	–	***	**	***	***	***	***	*
	lova/lover	–	–	–	–	***	*	**	–
	motha/mother	–	**	–	*	–	**	–	–
	ova/over	***	***	–	–	–	***	***	–
	sista/sister	*	***	–	–	**	–	–	–
	wateva/whatever	***	***	–	–	–	***	***	–
	wea/where	**	***	***	***	***	***	***	*
	total	***	***	***	***	***	***	***	–
/str/ → /skr/	skreet/street	–	–	–	**	*	**	*	**
	skrong/strong	**	***	–	*	**	**	**	*
	skrip/strip	*	–	*	***	***	–	***	***
	total	***	***	–	***	***	***	–	-
/ð/ → /d/ or /v/	brova/brother	***	***	***	***	***	–	***	***
	dat/that	–	***	–	–	–	**	**	–
	deez/these	–	–	–	**	***	–	**	***
	dem/them	*	***	**	**	***	–	–	–
	dey/they	***	***	**	*	**	**	***	–
	dis/this	–	***	**	–	–	–	*	*
	mova/mother	***	***	***	–	***	***	–	***
	total	***	***	***	–	***	–	***	***
/θ/ → /t/ or /f/	mouf/mouth	**	–	–	–	–	–	–	–
	nuffin/nothing	***	***	***	***	***	***	–	***
	souf/south	***	–	**	–	**	–	***	***
	teef/teeth	–	–	–	–	**	–	–	–
	trough/through	–	–	–	–	**	–	*	*
	trow/throw	*	–	–	***	**	*	**	**
	total	***	***	***	***	–	**	–	*

– = p $\geq$ 0.05, * = 0.05 $>$ p $\geq$ 0.01, ** = p $\leq$ 0.01, , *** = p $\leq$ 0.0005

Shading corresponds to negative correlations

Table 3: Demographic correlations

of African Americans (with explained variances in the range of 0.03-0.05).

Our data suggests that **H2**, namely that AAVE is more prevalent in the Gulf states, is probably true. Hypothesis **H1**, that AAVE is an urban ethnolect, lends some support in our data, but the correlation with urbanicity is weaker (and negatively correlated or non-significant in half of the cases).

Our data only lends limited support to the first half of hypothesis **H3**. While derhotacization and /str/ correlate (negatively) significantly with income levels, we see no significant correlations within /ð/ and a positive correlation within /θ/. However, our data does not suggest that **H3** is false, either. Our data does lend support to the more specific hypothesis **H5**, namely that derhoticization is sensitive to income level, while the strong correlation with the distribution of African Americans lends support to **H4**.

More interestingly, our data suggests that *women use AAVE features more often than men*, i.e., there is a negative correlation between male gender and AAVE features, contrary to the second half of **H3**, namely that AAVE is more frequently appropriated by men. Note, however, that our gender ratios are aggregated for city areas, and with the demographic bias of Twitter, these correlations should be taken with a grain of salt. Considering the small gender ratio differences, we also compute correlations between our linguistic features and gender using the Rovereto Twitter N-gram Corpus (RTC) (Herdagdelen and Baroni, 2011).[4] The RTC corpus contains information about the gender of the tweeter associated with n-grams. While there is too little data in the corpus to correlate gender and backing in /str/, derhotacization and both interdental fricative mutations (/ð/ → /d/ or /v/ and /θ/ → /t/ or /f/) correlate significantly with women. Out of our words, 10 correlate sig-

[4] `http://clic.cimec.unitn.it/amac/twitter_ngram/`

Feature	word pairs	latitude	longitude	urban	Gulf
/r/	brotha/brother	***	***	***	***
	foreva/forever	***	**	–	***
	hea/here	***	***	*	***
	lova/lover	***	***	**	***
	motha/mother	–	–	***	–
	ova/over	***	–	–	***
	sista/sister	–	***	**	***
	wateva/whatever	***	***	**	***
	wea/where	***	***	–	***
	total	***	***	***	***
/str/	skreet/street	***	–	***	***
	skrong/strong	***	*	***	***
	skrip/strip	***	–	***	***
	total	***	**	–	***
/ð/	brova/brother	***	***	***	***
	dat/that	***	*	–	***
	deez/these	*	***	–	–
	dem/them	***	***	–	***
	dey/they	***	***	–	***
	dis/this	***	–	–	***
	mova/mother	*	***	***	***
	total	***	***	***	***
/θ/	mouf/mouth	***	–	–	***
	nuffin/nothing	***	***	***	***
	souf/south	***	***	***	***
	teef/teeth	**	–	**	***
	trough/through	–	***	–	–
	trow/throw	***	**	–	***
	total	*	***	***	***

– = $p \geq 0.05$, * = $0.05 > p \geq 0.01$, ** = $p \leq 0.01$, *** = $p \leq 0.0001$

Shading corresponds to negative correlations

Table 2: Geographic correlations

nificantly with female speakers; seven with male. The correlations are found in Table 4. For each feature, certain words correlate significantly with female speakers, while others correlate significantly with male speakers. Consequently, neither our Twitter data not the Twitter data in the RTC suggest that AAVE is more often appropriated by men. We discuss whether our data provides a basis for falsifying the second half of **H3** in §3.1.

The high correlation between mutations of /ð/ and longitude supports the presence of these mutations of /ð/ in non-standard northern varieties (Rickford, 1999). The mutation of /θ/ is also correlated with longitude, and with latitude, suggesting an Eastern American feature rather than a distinct Southern feature (Rickford, 1999). The variation in mutations could possibly be explained by both geography as well as the distribution og African Americans.

There is evidence in our data that backing in /str/ (to /skr/) is appropriated more often by AAVE speakers than by speakers of other dialects (**H8**). There is also a negative correlation between latitude and backing in /str/ as well as a strong positive correlation with the Gulf states, suggesting that backing in /str/ is a feature primarily seen in this region. The data thereby suggests that the feature is appropriated significantly more by African Americans than by speakers of the Southern dialect.

In sum, while our data lends support to several of the common hypotheses from the sociolinguistics literature, we found one unexpected tendency, going against the second half of **H3**, namely that AAVE features were found more often with females. We now discuss this finding in light of the methodological problems discussed in §1.2.

Feature	word pairs	male
/r/ ⟶ /Ø/ or /ə/	brotha-brother	**
	foreva-forever	**
	hea-here	*
	lova-lover	–
	motha-mother	**
	ova-over	**
	sista-sister	–
	wateva-whatever	–
	wea-where	**
ð ⟶ /d/ or /v/	brova-brother	*
	dat-that	**
	deez-these	**
	dem-them	**
	dey-they	**
	dis-this	**
	mova-mother	–
θ ⟶ /f/ or /t/	mouf-mouth	**
	nuffin-nothing	**
	souf-south	**
	teef-teeth	–
	trough-through	**
	trow-throw	**

– = $p \geq 0.05$, * = $0.05 > p \geq 0.01$, ** = $p \leq 0.01$

Shading corresponds to negative correlations

Table 4: Gender correlations in RTC

3.1 Is AAVE *not* male-dominated?

We now discuss whether our data falsifies the second half of **H3**, one methodological problem at a time (see §1.3). If WRITTEN BIAS were to bias our conclusions, one gender should be more likely to exhibit more phonologically motivated spelling variation. This may actually be true, since it is well-established that women tend to be more linguistically creative and have larger vocabularies (Labov, 1990; Brizendine, 2006). Whether women are also more meta-linguistic (META-USE BIAS), has to the best of our knowledge not been studied. Since genders are almost equally geographically distributed, and since Twitter is generally considered gender-balanced, neither USER POPULATION BIAS nor GALTON'S PROBLEM is likely to bias our conclusions. TOPIC BIAS, on the other hand, may. While our semantically equivalent pairs control for topic, the pragmatics sometimes differ. Just like code-switching is a strategy for bilinguals, using the spelling *motha* instead of *mother* could mean something, say irony, which one gender is more prone for. In sum, while we do believe that our data should lead sociolinguists to question whether AAVE is male-dominated, our findings may be biased by WRITTEN BIAS.

4 POS tagging

We need automated syntactic analysis to study morpho-syntactic dialectal variation. We ran a state-of-the-art POS tagger trained on newswire[5] (STANFORD), as well as two state-of-the-art POS taggers adapted to Twitter, namely GATE[6] and ARK[7], on our data. We had one professional annotator manually annotate 100 positive (AAVE) and 100 negative (non-AAVE) sentences using the coarse-grained tags proposed by Petrov et al. (2011). We map the tagger outputs to those tags and report tagging accuracies. See Table 5 for results, with $\Delta(+,-)$ being the absolute difference in performance from non-AAVE to AAVE.

	STANFORD	GATE	ARK
AAVE	61.4	**79.1**	77.5
non-AAVE	74.5	**83.3**	77.9
Δ(+,-)	13.1	4.2	0.4

Table 5: POS tagging accuracies (%)

While GATE is certainly better than STANFORD on our data, performance is generally poor and prohibitive of many downstream applications and variational studies. We also note that both the best and worst tagger perform significantly worse on AAVE tweets than on non-AAVE tweets. What are the sources of error in the AAVE data? One example is the word *brotha*, which is tagged as a both an adverb, a verb, and as X (foreign words, mark-up, etc.). Contractions like *finna* (”fixing to” meaning ”going to”) and *gimme* (”give me”) are often tagged as particles, but annotated as verbs or, as in the case of *witchu* (”with you”), as a preposition. Another interesting mistake is tagging adverbial *like* as a verb.

5 Conclusion

Large-scale variational studies of social media can be used to question received wisdom about dialects, lending support to some sociolinguistic research hypotheses and questioning others. However, we caution that our results were biased by several factors, including the representativity of the social media user bases. We also show how state-of-the-art POS taggers are more likely to fail on dialects in social media. The performance drops may be considered prohibitive of studying morph-syntactic patterns across dialects and as a challenge to us as a community.

References

Sharon Ash and John Myhill. 1986. Linguistic correlates of inter-ethnic contact. In David

[5] http://nlp.stanford.edu/software/tagger.shtml

[6] https://gate.ac.uk/wiki/twitter-postagger.html

[7] http://www.ark.cs.cmu.edu/TweetNLP/

Sankoff, editor, *Diversity and Diachronyc*, pages 33–44, Amsterdam and Philadelphia. John Benjamins Publishing Co.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18.

Louann Brizendine. 2006. *The Female Brain*. Morgan Road Books.

Phillip Carter. 2013. Shared spaces, shared structures: Latino social formation and African American English in the U.S. south. *Journal of Sociolinguistics*, 17:66–92.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *EACL*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.

Jacob Eisenstein, Noah A. Smith, and Eric Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *ACL*.

Jacob Eisenstein. 2013. Phonological factors in social media writing. In *NAACL Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta, Georgia. Association for Computational Linguistics.

Jacob Eisenstein. to appear. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*.

Mark Graham, Scott Hale, and Devin Gaffney. 2014. Where in the world are you? Geolocation and language identification on Twitter. *The Professional Geographer*, 66(4).

Amac Herdagdelen and Marco Baroni. 2011. Stereotypical gender actions can be extracted from web text. *Journal of the American Society for Information Science and Technology*, 62:1741–1749.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *ACL*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review-sites as a source for large-scale sociolinguistic studies. In *WWW*.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *CoNLL*.

William Labov, Sharon Ash, and Charles Boberg. 2005. *The Atlas of North American English Phonetics, Phonology and Sound Change*. Mouton de Gruyter, New York, NY.

William Labov. 1972a. *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press.

William Labov. 1972b. *Sociolingustic Patterns*. University of Pennsylvania Press, Philadelphia, PA.

William Labov. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2:205–254, 7.

William Labov. 2006. Unendangered dialects, endangered people. In Natalie Schilling-Estes, editor, *GURT'06*.

Miriam Meyerhof. 2006. *Introducing Sociolinguistics*. Routledge.

R Naroll. 1961. Two solutions to Galton's problem. *Philosophy of Science*, 28.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. CoRR abs/1104.2086.

K.E. Pollock, G. Bailey, M. Berni, D. Fletcher, L. Hinton, I. Johnson, J. Roberts, and R. Weaver. 1998. Phonological features of african american english. http://www.rehabmed.ualberta.ca/spa/phonology/features.htm.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, pages 37–44. ACM.

Sonya Rastogi, Tallese D. Johnson, Elizabeth M. Hoeffel, and Malcolm P. Drewery Jr. 2011. The black population: 2010. Technical report, US Census, September.

John Rickford. 1999. *African American Vernacular English: Features, Evolution, Educational Implications*. Blackwell, Malden, MA.

John Rickford. 2010. Geographical diversity, residential segregation, and the vitality of african american vernacular english and its speakers. *Transforming Anthropology*, 18(1):28–34.

Sean Roberts and James Winters. 2013. Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. *PLoS ONE*, 8(8).

Eric Thomas. 2007. Phonological and phonetic characteristics of african american vernacular english. *Language and Linguistic Compass*, 1(5):450–475.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *EMNLP*.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media (demo). In *AAAI*.

Walt Wolfram. 2004. The grammar of urban african american vernacular english. In Kormann B. and E. Schneider, editors, *Handbook of Varieties of English*, pages 111–132, Berlin. Mouton de Gruyter.