Phonological Factors in Social Media Writing

Jacob Eisenstein
jacobe@gatech.edu
School of Interactive Computing
Georgia Institute of Technology

Abstract

Does phonological variation get transcribed into social media text? This paper investigates examples of the phonological variable of consonant cluster reduction in Twitter. Not only does this variable appear frequently, but it displays the same sensitivity to linguistic context as in spoken language. This suggests that when social media writing transcribes phonological properties of speech, it is not merely a case of inventing orthographic transcriptions. Rather, social media displays influence from structural properties of the phonological system.

1 Introduction

The differences between social media text and other forms of written language are a subject of increasing interest for both language technology (Gimpel et al., 2011; Ritter et al., 2011; Foster et al., 2011) and linguistics (Androutsopoulos, 2011; Dresner and Herring, 2010; Paolillo, 1996). Many words that are endogenous to social media have been linked with specific geographical regions (Eisenstein et al., 2010; Wing and Baldridge, 2011) and demographic groups (Argamon et al., 2007; Rao et al., 2010; Eisenstein et al., 2011), raising the question of whether this variation is related to spoken language dialects. Dialect variation encompasses differences at multiple linguistic levels, including the lexicon, morphology, syntax, and phonology. While previous work on group differences in social media language has generally focused on lexical differences, this paper considers the most purely "spoken" aspect of dialect: phonology.

Specifically, this paper presents evidence against the following two null hypotheses:

- H0: Phonological variation does not impact social media text.
- H1: Phonological variation may introduce new lexical items into social media text, but not the underlying structural rules.

These hypotheses are examined in the context of the phonological variable of *consonant cluster reduction* (also known as consonant cluster simplification, or more specifically, -/t,d/ deletion). When a word ends in cluster of consonant sounds — for example, *mist* or *missed* — the cluster may be simplified, for example, to *miss*. This well-studied variable has been demonstrated in a number of different English dialects, including African American English (Labov et al., 1968; Green, 2002), Tejano and Chicano English (Bayley, 1994; Santa Ana, 1991), and British English (Tagliamonte and Temple, 2005); it has also been identified in other languages, such as Quebecois French (Côté, 2004). While some previous work has cast doubt on the influence of spoken dialects on written language (Whiteman, 1982; Thompson et al., 2004), this paper presents large-scale evidence for consonant cluster reduction in social media text from Twitter — in contradiction of the null hypothesis H0.

But even if social media authors introduce new orthographic *transcriptions* to capture the sound of language in the dialect that they speak, such innovations may be purely lexical. Phonological variation is governed by a network of interacting preferences that include the surrounding linguistic context. Do

Proceedings of the Workshop on Language in Social Media (LASM 2013), pages 11–19,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

these structural aspects of phonological variation also appear in written social media?

Consonant cluster reduction is a classic example of the complex workings of phonological variation: its frequency depends on the morphology of the word in which it appears, as well as the phonology of the preceding and subsequent segments. The variable is therefore a standard test case for models of the interaction between phonological preferences (Guy, 1991). For our purposes, the key point is that consonant cluster reduction is strongly inhibited when the subsequent phonological segment begins with a vowel. The final *t* in *left* is more likely to be deleted in *I left **the** house* than in *I left **a** tip*. Guy (1991) writes, "prior studies are unanimous that a following consonant promotes deletion more readily than a following vowel," and more recent work continues to uphold this finding (Tagliamonte and Temple, 2005).

Consonant cluster reduction thus provides an opportunity to test the null hypothesis H1. If the introduction of phonological variation into social media writing occurs only on the level of new lexical items, that would predict that reduced consonant clusters would be followed by consonant-initial and vowel-initial segments at roughly equal rates. But if consonant cluster reduction is inhibited by adjacent vowel-initial segments in social media text, that would argue against H1. The experiments in this paper provide evidence of such context-sensitivity, suggesting that the influence of phonological variation on social media text must be deeper than the transcription of invidual lexical items.

2 Word pairs

The following word pairs were considered:

- *left / lef*
- *just / jus*
- *with / wit*
- *going / goin*
- *doing / doin*
- *know / kno*

The first two pairs display consonant cluster reduction, specifically t-deletion. As mentioned above, consonant cluster reduction is a property of African American English (AAE) and several other English dialects. The pair *with/wit* represents a stopping of the interdental fricative, a characteristic of New York English (Gordon, 2004), rural Southern English (Thomas, 2004), as well as AAE (Green, 2002). The next two pairs represent "g-dropping", the replacement of the velar nasal with the coronal nasal, which has been associated with informal speech in many parts of the English-speaking world.[1] The final word pair *know/kno* does not differ in pronunciation, and is included as a control.

These pairs were selected because they are all frequently-used words, and because they cover a range of typical "shortenings" in social media and other computer mediated communication (Gouws et al., 2011). Another criterion is that each shortened form can be recognized relatively unambiguously. Although *wit* and *wan* are standard English words, close examination of the data did not reveal any examples in which the surface forms could be construed to indicate these words. Other words were rejected for this reason: for example, *best* may be reduced to *bes*, but this surface form is frequently used as an acronym for *Blackberry Enterprise Server*.

Consonant cluster reduction may be combined with morphosyntactic variation, particularly in African American English. Thompson et al. (2004) describe several such cases: zero past tense (*mother kiss(ed) them all goodbye*), zero plural (*the children made their bed(s)*), and subject-verb agreement (*then she jump(s) on the roof*). In each of these cases, it is unclear whether it is the morphosyntactic or phonological process that is responsible for the absence of the final consonant. Words that feature such ambiguity, such as *past*, were avoided.

Table 1 shows five randomly sampled examples of each shortened form. Only the relevant portion of each message is shown. From consideration of many examples such as these, it is clear that the shortened forms *lef, jus, wit, goin, doin, kno* refer to the standard forms *left, just, with, going, doing, know* in the overwhelming majority of cases.

[1] Language Log offers an engaging discussion of the linguistic and cultural history of "g-dropping." http://itre.cis.upenn.edu/~myl/languagelog/archives/000878.html

1. *ok **lef** the y had a good workout*
   (ok, left the YMCA, had a good workout)
2. *@USER **lef** the house*
3. *eat off d wol a d rice and **lef** d meat*
   (... left the meat)
4. *she nah **lef** me*
   (she has not left me)
5. *i **lef** my changer*

6. ***jus** livin this thing called life*
7. *all the money he **jus** took out the bank*
8. *boutta **jus** strt tweatin random shxt*
   (about to just start tweeting ...)
9. *i **jus** look at shit way different*
10. *u **jus** fuckn lamee*

11. *fall in love **wit** her*
12. *i mess **wit** pockets*
13. *da hell **wit** u*
    (the hell with you)
14. *drinks **wit** my bro*
15. *don't fuck **wit** him*

16. *a team that's **goin** to continue*
17. *what's **goin** on tonight*
18. *is reign stil **goin** down*
19. *when is she **goin** bck 2 work?*
20. *ur not **goin** now where*
    (you're not going nowhere)

21. *u were **doin** the same thing*
22. *he **doin** big things*
23. *i'm not **doin** shit this weekend*
24. *oh u **doin** it for haiti huh*
25. *i damn sure aint **doin** it in the am*

26. *u **kno** u gotta put up pics*
27. *i **kno** some people bout to be sick*
28. *u already **kno***
29. *you **kno** im not ugly pendeja*
30. *now i **kno** why i'm always on netflix*

Table 1: examples of each shortened form

3 Data

Our research is supported by a dataset of microblog posts from the social media service Twitter. This service allows its users to post 140-character messages. Each author's messages appear in the newsfeeds of individuals who have chosen to "follow" the author, though by default the messages are publicly available to anyone on the Internet. Twitter has relatively broad penetration across different ethnicities, genders, and income levels. The Pew Research Center has repeatedly polled the demographics of Twitter (Smith and Brewer, 2012), finding: nearly identical usage among women (15% of female internet users are on Twitter) and men (14%); high usage among non-Hispanic Blacks (28%); an even distribution across income and education levels; higher usage among young adults (26% for ages 18-29, 4% for ages 65+).

Twitter's streaming API delivers an ongoing random sample of messages from the complete set of public messages on the service. The data in this study was gathered from the public "Gardenhose" feed, which is claimed to be approximately 10% of all public posts; however, recent research suggests that the sampling rate for geolocated posts is much higher (Morstatter et al., 2013). This data was gathered over a period from August 2009 through the end of September 2012, resulting in a total of 114 million messages from 2.77 million different user accounts (Eisenstein et al., 2012).

Several filters were applied to ensure that the dataset is appropriate for the research goals of this paper. The dataset includes only messages that contain geolocation metadata, which is optionally provided by smartphone clients. Each message must have a latitude and longitude within a United States census block, which enables the demographic analysis in Section 6. Retweets are excluded (both as identified in the official Twitter API, as well as messages whose text includes the token "RT"), as are messages that contain a URL. Grouping tweets by author, we retain only authors who have fewer than 1000 "followers" (people who have chosen to view the author's messages in their newsfeed) and who follow fewer than 1000 individuals.

Specific instances of the word pairs are acquired by using `grep` to identify messages in which the shortened form is followed by another sequence of purely alphabetic characters. Reservoir sampling (Vitter, 1985) was used to obtain a randomized set of at most 10,000 messages for each word. There were only 753 examples of the shortening *lef*; for all other words we obtain the full 10,000 messages. For each shortened word, an equal number of samples for the standard form were obtained through the same method: `grep` piped through a reservoir sampler. Each instance of the standard form must also be followed by a purely alphabetic string. Note that the total number of instances is slightly higher than the number of messages, because a word may appear multiple times within the same message. The counts are shown in Table 2.

4 Analysis 1: Frequency of vowels after word shortening

The first experiment tests the hypothesis that consonant clusters are less likely to be reduced when followed by a word that begins with a vowel letter. Table 2 presents the counts for each term, along with the probability that the next segment begins with the vowel. The probabilities are accompanied by 95% confidence intervals, which are computed from the standard deviation of the binomial distribution.All differences are statistically significant at $p < .05$.

The simplified form *lef* is followed by a vowel only 19% of the time, while the complete form *left* is followed by a vowel 35% of the time. The absolute difference for *jus* and *just* is much smaller, but with such large counts, even this 2% absolute difference is unlikely to be a chance fluctuation.

The remaining results are more mixed. The shortened form *wit* is significantly *more* likely to be followed by a vowel than its standard form *with*. The two "g dropping" examples are inconsistent, and troublingly, there is a significant effect in the control condition. For these reasons, a more fine-grained analysis is pursued in the next section.

A potential complication to these results is that cluster reduction may be especially likely in specific phrases. For example, *most* can be reduced to *mos*, but in a sample of 1000 instances of this reduction, 72% occurred within a single expression: *mos def*. This phrase can be either an expression of certainty (*most definitely*), or a reference to the performing artist of the same name. If *mos* were observed to be more likely to be followed by a consonant-initial word than *most*, this might be attributable to this one expression.

word	N	N(vowel)	P(vowel)
lef	753	145	0.193 ± 0.028
left	757	265	0.350 ± 0.034
jus	10336	939	0.091 ± 0.006
just	10411	1158	0.111 ± 0.006
wit	10405	2513	0.242 ± 0.008
with	10510	2328	0.222 ± 0.008
doin	10203	2594	0.254 ± 0.008
doing	10198	2793	0.274 ± 0.009
goin	10197	3194	0.313 ± 0.009
going	10275	1821	0.177 ± 0.007
kno	10387	3542	0.341 ± 0.009
know	10402	3070	0.295 ± 0.009

Table 2: Term counts and probability with which the following segment begins with a vowel. All differences are significant at $p < .05$.

An inverse effect could explain the high likelihood that *goin* is followed by a vowel. Given that the author has chosen an informal register, the phrase *goin to* is likely to be replaced by *gonna*. One might hypothesize the following decision tree:

- If formal register, use *going*
- If informal register,
  - If next word is *to*, use *gonna*
  - else, use *goin*

Counts for each possibility are shown in Table 3; these counts are drawn from a subset of the 100,000 messages and thus cannot be compared directly with Table 2. Nonetheless, since *to* is by far the most frequent successor to *going*, a great deal of *going*'s preference for consonant successors can be explained by the word *to*.

5 Analysis 2: Logistic regression to control for lexical confounds

While it is tempting to simply remove *going to* and *goin to* from the dataset, this would put us on a slippery slope: where do we draw the line between lexical confounds and phonological effects? Rather than

	total	... *to*	percentage
going	1471	784	53.3%
goin	470	107	22.8%
gonna	1046	n/a	n/a

Table 3: Counts for *going to* and related phrases in the first 100,000 messages in the dataset. The shortened form *goin* is far less likely to be followed by *to*, possibly because of the frequently-chosen *gonna* alternative.

word	μ_β	σ_β	z	p
lef/left	-0.45	0.10	-4.47	3.9×10^{-6}
jus/just	-0.43	0.11	-3.98	3.4×10^{-5}
wit/with	-0.16	0.03	-4.96	3.6×10^{-7}
doin/doing	0.08	0.04	2.29	0.011
goin/going	-0.07	0.05	-1.62	0.053
kno/know	-0.07	0.05	-1.23	0.11

Table 4: Logistic regression coefficients for the VOWEL feature, predicting the choice of the shortened form. Negative values indicate that the shortened form is less likely if followed by a vowel, when controlling for lexical features.

excluding such examples from the dataset, it would be preferable to apply analytic techniques capable of sorting out lexical and systematic effects. One such technique is logistic regression, which forces lexical and phonological factors to **compete** for the right to explain the observed orthographic variations.[2]

The dependent variable indicates whether the word-final consonant cluster was reduced. The independent variables include a single feature indicating whether the successor word begins with a vowel, and additional lexical features for all possible successor words. If the orthographic variation is best explained by a small number of successor words, the phonological VOWEL feature will not acquire significant weight.

Table 4 presents the mean and standard deviation of the logistic regression coefficient for the VOWEL feature, computed over 1000 bootstrapping iterations (Wasserman, 2005).[3] The coefficient has the largest magnitude in cases of consonant cluster reduction, and the associated p-values indicate strong statistical significance. The VOWEL coefficient is also strongly significant for *wit/with*. It reaches the $p < .05$ threshold for *doin/doing*, although in this case, the presence of a vowel indicates a preference for the shortened form *doin* — contra the raw frequencies in Table 2. The coefficient for the VOWEL feature is not significantly different from zero for *goin/going* and for the control *kno/know*. Note that since we had no prior expectation of the coefficient sign in these cases, a two-tailed test would be most appropriate, with critical value $\alpha = 0.025$ to establish 95% confidence.

6 Analysis 3: Social variables

The final analysis concerns the relationship between phonological variation and social variables. In spoken language, the word pairs chosen in this study have connections with both ethnic and regional dialects: consonant cluster reduction is a feature of African-American English (Green, 2002) and Tejano and Chicano English (Bayley, 1994; Santa Ana, 1991); th-stopping (as in *wit/with*) is a feature of African-American English (Green, 2002) as well as several regional dialects (Gordon, 2004; Thomas, 2004); the velar nasal in *doin* and *goin* is a property of informal speech. The control pair *kno/know* does not correspond to any sound difference, and thus there is no prior evidence about its relationship to social variables.

The dataset includes the average latitude and longitude for each user account in the corpus. It is possible to identify the county associated with the latitude and longitude, and then to obtain county-level demographic statistics from the United States census. An **approximate** average demographic profile for each word in the study can be constructed by aggregating the demographic statistics for the counties of residence of each author who has used the word. Twitter users do not comprise an unbiased sample from each county, so this profile can only describe the demographic environment of the authors, and not the demographic properties of the authors themselves.

Results are shown in Figure 1. The confidence intervals reflect the Bonferroni correction for multiple comparison, setting $\alpha = 0.05/48$. The con-

[2] (Stepwise) logistic regression has a long history in variationist sociolinguistics, particularly through the ubiquitous VARBRUL software (Tagliamonte, 2006).

[3] An L2 regularization parameter was selected by randomly sampling 50 training/test splits. Average accuracy was between 58% and 66% on the development data, for the optimal regularization coefficient.

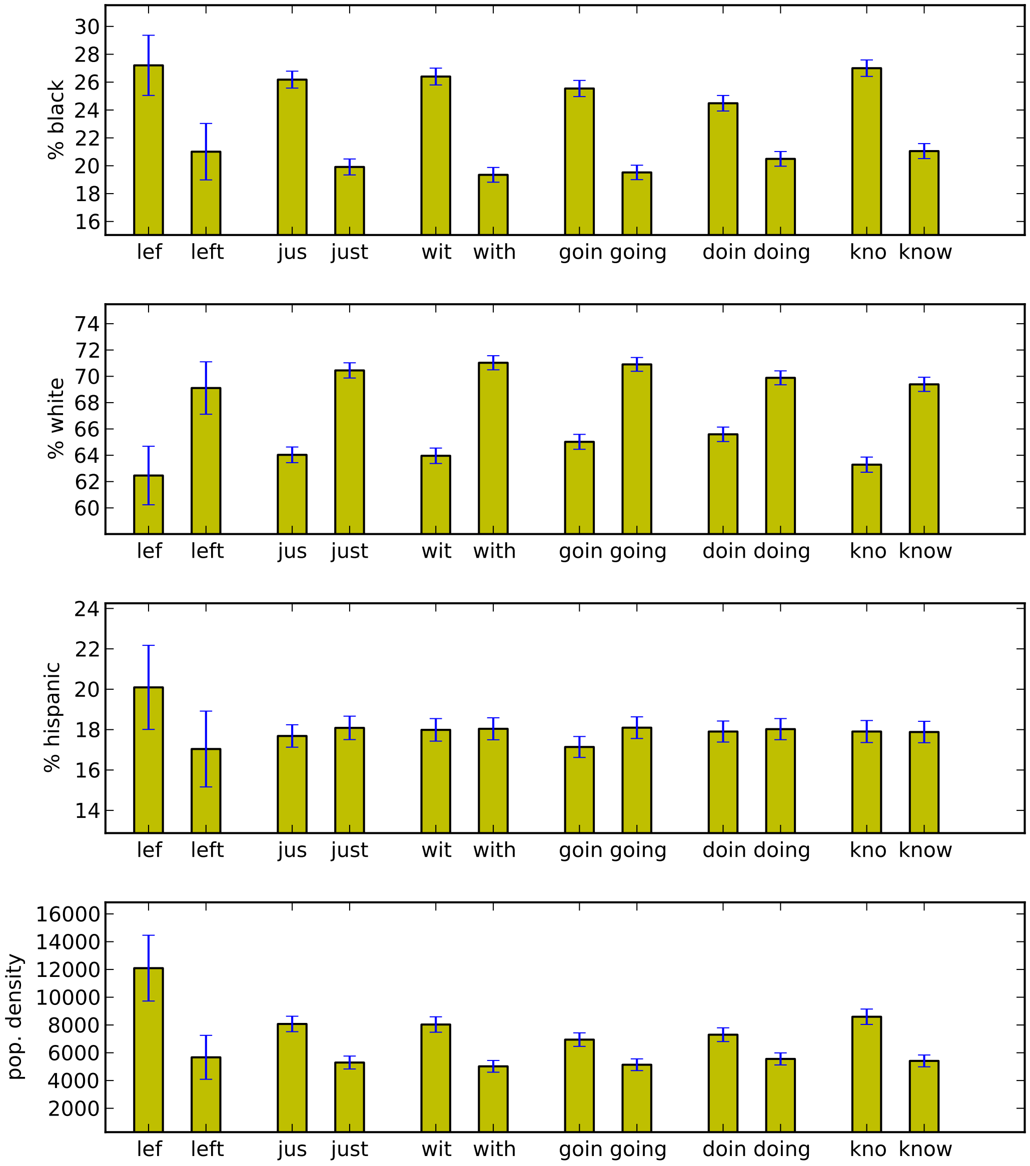

Figure 1: Average demographics of the counties in which users of each term live, with 95% confidence intervals

sonant cluster reduction examples are indeed preferred by authors from densely-populated (urban) counties with more African Americans, although these counties tend to prefer *all* of the non-standard variants, including the control pair *kno/know*. Conversely, the non-standard variants have aggregate demographic profiles that include fewer European Americans. None of the differences regarding the percentage of Hispanics/Latinos are statistically significant. Overall, these results show an association between non-standard orthography and densely-populated counties with high proportions of African Americans, but we find no special affinity for consonant cluster reduction.

7 Related work

Previous studies of the impact of dialect on writing have found relatively little evidence of purely phonological variation in written language. Whiteman (1982) gathered an oral/written dataset of interview transcripts and classroom compositions. In the written data, there are many examples of final consonant deletion: verbal *-s* (*he go- to the pool*), plural *-s* (*in their hand-*), possessive *-s* (*it is Sally- radio*), and past tense *-ed*. However, each of these deletions is morphosyntactic rather than purely phonological. They are seen by Whiteman as an omission of the inflectional suffix, rather than as a transcription of phonological variation, which she finds to be very rare in cases where morphosyntactic factors are not in play. She writes, "nonstandard phonological features rarely occur in writing, even when these features are extremely frequent in the oral dialect of the writer."

Similar evidence is presented by Thompson et al. (2004), who compare the spoken and written language of 50 third-grade students who were identified as speakers of African American English (AAE). While each of these students produced a substantial amount of AAE in spoken language, they produced only one third as many AAE features in the written sample. Thompson *et al.* find almost no instances of purely phonological features in writing, including consonant cluster reduction — except in combination with morphosyntactic features, such as zero past tense (e.g. *mother kiss(ed) them all goodbye*). They propose the following explanation:

> African American students have models for *spoken* AAE; however, children do not have models for written AAE... students likely have minimal opportunities to experience AAE in print (emphasis in the original).

This was written in 2004; in the intervening years, social media and text messages now provide many examples of written AAE. Unlike classroom settings, social media is informal and outside the scope of school control. Whether the increasing prevalence of written AAE will ultimately lead to widely-accepted writing systems for this and other dialects is an intriguing open question.

8 Conclusions and future work

The experiments in this paper demonstrate that phonology impacts social media orthography at the word level and beyond. I have discussed examples of three such phenomena: consonant cluster reduction, th-stopping, and the replacement of the velar nasal with the coronal ("g-dropping"). Both consonant cluster reduction and th-stopping are significantly influenced by the phonological context: their frequency depends on whether the subsequent segment begins with a vowel. This indicates that when social media authors transcribe spoken language variation, they are not simply replacing standard spellings of individual words. The more difficult question — *how* phonological context enters into writing — must be left for future work.

There are several other avenues along which to continue this research. The sociolinguistic literature describes a number of other systematic factors that impact consonant cluster reduction (Guy, 1991; Tagliamonte and Temple, 2005), and a complete model that included all such factors might shed additional light on this phenomenon. In such work it is typical to distinguish between different types of consonants; for example, Tagliamonte and Temple (2005) distinguish obstruents, glides, pauses, and the liquids /r/ and /l/. In addition, while this paper has equated consonant *letters* with consonant *sounds*, a more careful analysis might attempt to induce (or annotate) the pronunciation of the relevant words. The speech synthesis literature offers numerous such methods (Bisani and Ney, 2008), though social media text may pose new challenges, particularly for approaches that are based on generalizing from standard pronunciation dictionaries.

One might also ask whether the phonological system impacts all authors to the same extent. Labov (2007) distinguishes two forms of language change: *transmission*, where successive generations of children advance a sound change, and *diffusion*, where language contact leads adults to "borrow" aspects of other languages or dialects. Labov marshalls evidence from regional sound changes to show that transmission is generally more structural and regular, while diffusion is more superficial and irregular; this may be attributed to the ability of child language learners to recognize structural linguistic patterns. Does phonological context impact transcription equally among all authors in our data, or can we identify authors whose use of phonological transcription is particularly sensitive to context?

Acknowledgments

Thanks to Brendan O'Connor for building the Twitter dataset that made this research possible. Thanks to the reviewers for their helpful comments.

References

Jannis Androutsopoulos. 2011. Language change and digital media: a review of conceptions and evidence. In Nikolas Coupland and Tore Kristiansen, editors, *Standard Languages and Language Standards in a Changing Europe*. Novus, Oslo.

S. Argamon, M. Koppel, J. Pennebaker, and J. Schler. 2007. Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday*, 12(9).

Robert Bayley. 1994. Consonant cluster reduction in tejano english. *Language Variation and Change*, 6(03):303–326.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Commun.*, 50(5):434–451, May.

Marie-Hélène Côté. 2004. Consonant cluster simplification in Québec French. *Probus: International journal of Latin and Romance linguistics*, 16:151–201.

E. Dresner and S.C. Herring. 2010. Functions of the nonverbal in cmc: Emoticons and illocutionary force. *Communication Theory*, 20(3):249–268.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP*.

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

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2012. Mapping the geographical diffusion of new words, October.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proceedings of IJCNLP*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of ACL*.

Matthew J. Gordon, 2004. *A Handbook of Varieties of English*, chapter New York, Philadelphia, and other northern cities, pages 282–299. Volume 1 of Kortmann et al. (Kortmann et al., 2004).

Stephan Gouws, Dirk Hovy, and Donald Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90, Edinburgh, Scotland, July. Association for Computational Linguistics.

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press, September.

Gregory R. Guy. 1991. Contextual conditioning in variable lexical phonology. *Language Variation and Change*, 3:223–239, June.

Bernd Kortmann, Edgar W. Schneider, and Kate Burridge et al., editors. 2004. *A Handbook of Varieties of English*, volume 1. Mouton de Gruyter, Berlin, Boston.

William Labov, Paul Cohen, Clarence Robins, and John Lewis. 1968. A study of the Non-Standard english of negro and puerto rican speakers in new york city. Technical report, United States Office of Education, Washington, DC.

William Labov. 2007. Transmission and diffusion. *Language*, 83(2):344–387.

Fred Morstatter, Jurgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of ICWSM*.

John C. Paolillo. 1996. Language choice on soc.culture.punjab. *Electronic Journal of Communication/La Revue Electronique de Communication*, 6(3).

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of Workshop on Search and mining user-generated contents*.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of EMNLP*.

Otto Santa Ana. 1991. *Phonetic simplification processes in the English of the barrio: A cross-generational sociolinguistic study of the Chicanos of Los Angeles*. Ph.D. thesis, University of Pennsylvania.

Aaron Smith and Joanna Brewer. 2012. Twitter use 2012. Technical report, Pew Research Center, May.

Sali Tagliamonte and Rosalind Temple. 2005. New perspectives on an ol' variable: (t,d) in british english. *Language Variation and Change*, 17:281–302, September.

Sali A. Tagliamonte. 2006. *Analysing Sociolinguistic Variation*. Cambridge University Press.

Erik R Thomas, 2004. *A Handbook of Varieties of English*, chapter Rural Southern white accents, pages 87–114. Volume 1 of Kortmann et al. (Kortmann et al., 2004).

Connie A. Thompson, Holly K. Craig, and Julie A. Washington. 2004. Variable production of african american english across oracy and literacy contexts. *Language, speech, and hearing services in schools*, 35(3):269–282, July.

Jeffrey S. Vitter. 1985. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, March.

Larry Wasserman. 2005. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer, October.

Marcia F. Whiteman. 1982. Dialect influence in writing. In Marcia Farr Whiteman and Carl, editors, *Writing: The Nature, Development, and Teaching of Written Communication*, volume 1: Variation in writing. Routledge, October.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of ACL*.