Systematic patterning in phonologically-motivated orthographic variation[*]

Jacob Eisenstein
School of Interactive Computing
Georgia Institute of Technology
85 5th St NE, Atlanta GA 30312, USA
+1 617 913 2859
jacobe@gatech.edu

April 13, 2015

Abstract Social media features a wide range of nonstandard spellings, many of which appear inspired by phonological variation. However, the nature of the connection between variation across the spoken and written modalities remains poorly understood. Are phonological variables transferred to writing on the level of graphemes, or is the larger system of contextual patterning also transferred? This paper considers orthographic coda deletions corresponding to the phonological variables of (ing) and (t,d). In both cases, orthography mirrors speech: reduction of the *-ing* suffix depends on the word's syntactic category, and reduction of the *-t,-d* suffix depends on the succeeding phonological context. These spellings are more frequently used in informal conversational contexts, and in areas with high proportions of African Americans, again mirroring the patterning of the associated phonological variables. This suggests a deep connection between variation in the two modalities, necessitating a new account of the production of cross-modal variation. (150 words)

Keywords variation; orthography; social media; computer-mediated communication

Word count 7687

[*] This is a preprint of an article to appear in the **Journal of Sociolinguistics**. The research reported in this article was supported by National Science Foundation award IIS-1111142. Thanks to Brendan O'Connor for collaborating on obtaining the Twitter dataset, and to Lauren Hall-Lew, Shawn Ling Ramirez, Tyler Schnoebelen, and Sali A. Tagliamonte for helpful discussions of this research. Thanks also to the editors and anonymous reviewers for their thoughtful feedback, which has greatly improved the paper.

1 Introduction

The widespread adoption of online social media for informal communication provides new contexts for the use of written language (Crystal, 2006; Androutsopoulos, 2011), and new opportunities for the study of language variation. While much has been written in the popular and academic press about the abbreviations and emoticons that characterize online writing (Werry, 1996; Tagliamonte and Denis, 2008; Dresner and Herring, 2010; Varnhagen, McFall, Pugh, Routledge, Sumida-MacDonald and Kwong, 2010), this paper focuses on unconventional spellings of traditional words (Androutsopoulos, 2000), a phenomenon that Anis (2007) calls **neography**. Many such spellings can be considered "phonetic", in that they cohere, in one way or another, with plausible pronunciations. For example, the spelling *wut* would seem to indicate an identical pronunciation as *what*, for speakers who merge /hw/ and /w/. But the logic that produces such simplified spellings can also be employed to emphasize the variability of pronunciation. Of particular interest from a sociolinguistic perspective are variable spellings that correspond to social or dialectal variation in speech, such as the "dropped g" that connotes informality (Fischer, 1958), or the *t,d*-deletion that characterizes African-American English (Green, 2002) and a number of other dialects (Guy, 1991; Tagliamonte and Temple, 2005).

The capacity of such spellings to carry sociocultural meanings is well-documented in prior literature, which emphasizes how the use of orthography to represent dialectal or colloquial speech can index minority and subcultural identities (Androutsopoulos, 2000; Sebba, 2007), and can convey resistance to standard language ideologies (Jaffe, 2000). But while variation can serve a broad range of social purposes, it is embedded in a linguistic system that constrains the "palette" of orthographic choices available to the writer (Shortis, 2007). This is particularly true for phonologically-motivated respellings, as they may be subject to forces exerted by both the "host" modality of writing, as well as the original modality of speech. Understanding the possibilities for phonologically-motivated orthographic variation therefore requires clarifying the nature of the relationship between non-standard speech and spelling.

Just what does it mean to say that a respelling is "phonetic"? A straightforward explanation would be that phonemes are cognitively aligned to graphemes or grapheme sequences, so that patterns of substitution or deletion on the phonetic side give rise to analogous patterns in orthography. For example, *g*-deletion corresponds to an alternation between the coronal and velar nasal; this phonetic substitution would then license an analogous respelling in which the *-ing* grapheme is substituted by *-in*. Note however that the phonological variable (ing)[1] does not simply permit free variation between the associated phonemes, but is subject to a system of constraints and preferences: for example, it cannot apply to monomorphemic words like *thing*, and is more frequent in verbs than in nouns and adjectives (Houston, 1985). A model in which phonetic respellings are executed purely on the graphemic level would not

[1] Parentheticals, such as (ing), are conventionally used as shorthands for phonological variables. So as not to presuppose an identity between these phonological variables and the spelling variation studied here, the use of this notation is reserved for spoken-language phenomena in this paper.

account for these factors, and would predict an identical frequency of phonetic respelling across all classes of words that end with *-ing*.

A more nuanced model would be that phonetic respellings apply on the word level, so that writers are unlikely to produce respellings that would correspond to infelicitous pronunciations of words. Thus, *goin* would be an acceptable respelling for *going*, but *rin* would not be an acceptable respelling for *ring*. This word-level model would also predict a greater frequency of coda *g*-deletion in verbs rather than in adjectives or nouns, and it is capable of expressing within-word phonological conditioning, such as the increased frequency of *t,d*-deletion in coronal stops that follow sibilants (Guy, 1991). In exchange, this model imposes more cognitive work on the writer, who must assess (consciously or otherwise) the word-level phonetic plausibility of each respelling — at least until such time as the respelling is sufficiently conventionalized to be directly accessible.

The system of phonological variation also includes patterning across multiple words, and the word-level model would predict that this patterning would **not** be reflected in phonologically-motivated respellings. For example, in *t,d*-deletion in the coda (e.g., *jus*, *ol*), it is well known that the succeeding phonological segment exerts a strong effect, with vowel-initial segments inhibiting deletion (Guy, 1991). Because the succeeding phonological segment is in the next word, accounting for this systematic patterning would require the writer to assess (again, consciously or otherwise) the phonological plausibility of the **entire utterance** while producing the alternative spelling. Conventionalization is an unlikely explanation in this case, since it would imply encoding alternative spellings for millions of word pairs.

This paper offers an empirical evaluation of phonologically-motivated orthographic variation, in the context of the spoken language variables (ing) and (t,d), and their associated written forms, *g*-deletion and *t,d*-deletion. In the case of *g*-deletion, the goal is to determine whether the respelling is sensitive to the grammatical function of the word in which it is employed, thus testing whether phonologically-motivated respellings could be explained by graphemic substitution. In the case of *t,d*-deletion, the goal is to assess sensitivity to the phonological context implied by the succeeding word, thus testing whether these respellings could be explained at the word level. In addition, by examining the aggregate demographic statistics of the geographical regions in which these spellings are used, it is possible to obtain an approximate estimate of the demographic environments in which the users of these respellings live. This provides a new, large-scale quantitative perspective on the demographics of neography, complementing prior work on spelling variation in other contexts (Jaffe, Androutsopoulos, Sebba and Johnson, 2012).

The remainder of paper is structured as follows. I first survey the literature on the relationship between phonological and orthographic variation, touching also on previous studies of language variation in social media. Next, I describe the Twitter-based social media corpus used in this research, while briefly discussing the challenges and opportunities offered by social media data for variationist sociolinguistic investigation. The heart of the paper analyzes *g*-deletion and *t,d*-deletion, focusing on how phonological and grammatical contexts modulate the use of these spellings. These quantitative analyses include both frequency comparisons and mixed effects logistic regression, using the `rbrul` software package (Johnson, 2009). In both cases, the central question is the extent to which the system of variation from spoken language transfers to the written medium: when a final character is deleted, is this simply a graphemic substitution, or do the resulting spellings display the same systematic patterning as the phonetic variables that they imply? I conclude with a discussion of the implications of these results for a more unified understanding of phonological and orthographic variation, and for the use of social media corpora in sociolinguistic research.

2 Background

Written realizations of spoken dialects can be considered from at least two perspectives. First, there are transcriptions of speech events produced by third-party observers, such as novelists or researchers. This sort of non-standard orthography is termed "imitative" by Miethaner (2000), and the status of such transcriptions as representations of the original speech has been a topic of concern for a number of scholars (Preston, 1985; Jaffe and Walton, 2000; Bucholtz, 2007). This paper focuses instead on orthographic variation as a primary literacy practice (Sebba, 2007). Early studies of the relationship between dialect and writing found relatively little evidence of phonologically-motivated variation in written language. Whiteman (1982) gathered a multimodal dataset of interview transcripts and classroom compositions, finding many examples of final consonant deletion in the written part of the corpus: 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 seen as an omission of the inflectional suffix, rather than as a written analogue of phonological variation, which Whiteman finds to be very rare in cases where morphosyntactic factors are not in play. She writes (page 164), "nonstandard phonological features rarely occur in writing, even when these features are extremely frequent in the oral dialect of the writer." Similar conclusions are drawn by Thompson, Craig and Washington (2004), who compare the spoken and written language of 50 third-grade students 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, 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. (page 280; emphasis in the original)

The lack of print examples of AAE does not imply that speakers of this dialect do not incorporate it in their writing: both Whiteman and Thompson *et al.* observe morphosyntactic variables from spoken AAE in student writing, and Ball (1992) has shown how discourse-level features of African American English lead to distinctive regularities in expository writing. But these prior works do **not** find evidence of phonological patterning making its way into writing. In the intervening years since this research was published, social media has come to provide a new public forum for written AAE. Unlike classroom settings, social media is informal and outside the scope of school control and other "language police" (Tagliamonte and Denis, 2008, page 27). This medium therefore provides new opportunities for the expression of phonological dialect features, and for convergence of these forms of expression into regular norms.

In other contexts, researchers have noticed the presence of alternative spellings that seem unambiguously connected to phonological variation (e.g. Androutsopoulos, 2000). Of particular relevance for the present study are attempts to quantify the relationship of such spellings to their phonological and grammatical context. Hinrichs and White-Sustaíta (2011) investigate four non-standard spellings (*yu, mi, dem, neva*) from Jamaican Creole, and perform a variable rules analysis to identify the role of social variables such as gender and author residence, as well as grammatical factors relating to Creole syntax. Of particular interest in their results are interactions between social variables and contextual factors, with Jamaicans living abroad tending to reserve nonstandard orthography for lexical items that have Creole grammatical functions — such as using the spelling *mi* for the first person singular subject pronoun, and the spelling *me* for the first person singular object pronoun, as in standard English. This suggests that the orthography reflects differing conceptions of the relationship between Jamaican Creole and standard English. Similarly, Honeybone and Watson (2013) propose that alternative spellings can reveal which features of phonological variation are perceptually salient, investigating the rate of transcription for various features in a corpus of "humorous localized dialect literature" from Liverpool (CHLDL).

More broadly, the present paper represents an attempt to add a new, structural perspective to our understanding of the relationship between computer-mediated communication (CMC) and other linguistic modalities. The position of CMC with respect to speech and writing has been a topic of active interest for at least two decades; for example, Yates (1996) compared the quantitative properties of instant messages (IM) with speech and formal writing. Tagliamonte and Denis (2008) compare IM text with speech, using a dataset of transcripts from teenagers in both modes. They find that while IM does reflect some ongoing changes from spoken English (such as the intensifier *so*), it is more conservative than speech in other ways (such as with respect to ongoing changes in the quotative system), and is therefore best seen as a novel, "hybrid form." Baron (2004) used an IM corpus to compute the frequency of stereotypical "netspeak" features (Crystal, 2006), such as abbreviations and emoticons, finding them to be relatively rare. Other researchers have focused on the pragmatic function of features such as emoticons (Walther and D'Addario, 2001; Dresner and Herring, 2010; Schnoebelen, 2012), and Herring (2012) offers a survey of structural properties of online language. Recent work on email touches on the connection between writing and speech, showing that expressive letter repetitions such as *soooo* are typically "articulable", in the sense that the repeated character is less likely to refer to a plosive consonant (Kalman and Gergle, 2014).

Social media would seem to offer a wealth of new data for variationist sociolinguistics, yet the use of this resource has been only lightly explored. Paolillo (2001) describes how the use of code-switching, vernacular variants, and netspeak features relate to social network positions in an Internet Relay Chat channel. A number of studies address the relationship between language and gender, typically with the goal of identifying linguistic features that distinguish the writing of women and men (Herring and Paolillo, 2006; Argamon, Koppel, Pennebaker and Schler, 2007; Bamman, Eisenstein and Schnoebelen, 2014). Other recent work has been concerned with disentangling the relationship between geographical variation in spoken language and in social media writing. Eisenstein, O'Connor, Smith and Xing (2010) describe a method for inducing dialect regions from language differences in online writing on the social media service Twitter, finding that these regions are characterized by a combination of entity names, spoken language dialect words, and locally-specific netspeak terms such as abbreviations, alternative spellings, and emoticons. Doyle (2014) also uses Twitter, showing that the geographical distributions of traditional dialect variables (such as the double modal *might could*) roughly matches the geographic distributions of the associated speech variables, through comparisons to the Atlas of North American English (Labov, Ash and Boberg, 2006). Most of these previous efforts are focused on variation that is lexical in nature, or can be captured through lexical-syntactic patterns. The present study targets the lower-level phenomenon of systematic spelling variation, emphasizing systematic conditioning on phonological and grammatical context.

3 Data

The use of social media data in sociolinguistic research brings both challenges and opportunities. It is easy to build very large corpora with billions of tokens of text from thousands or even millions of individuals. This makes it possible to study relatively rare phenomena, and offers robustness to individual outliers. Social media text is written outside an interview setting, and in many cases, the intended audience can be inferred from metadata, along with information such as the time and place of authorship. In this way, social media offers a unique perspective on meaningful communication in a natural setting. However, with these advantages come new challenges. Not all social media data is publicly available, and its use poses ethical questions, particularly when text can be tied to its author's real-life identity (boyd and Crawford, 2012). Conversely, when the text **cannot** be tied to any real-world person, the true demographic composition of the dataset is unknown. Of particular concern are automated accounts that post repetitive content for marketing or other purposes. Social media services such as Twitter provide several affordances for authors to quote each other's text, making the true count of observed features difficult to ascertain. Finally, social media has well-known demographic and geographic biases (Mislove, Lehmann, Ahn, Onnela and Rosenquist, 2011; Hecht and Stephens, 2014), making it difficult to draw robust conclusions about the linguistic profile of any particular population. In this study, a number of measures are taken to avoid or mitigate these concerns; however, the "purity" of social media data will probably never approach that of more traditional sociolinguistic study populations, and the implications of unanticipated biases must always be kept in mind. These biases are at least partially offset by the advantage of working with socially meaningful writing from a broad sample of individuals, outside the interview setting.

3.1 Twitter as a sociolinguistic corpus

The data in this study is drawn from Twitter, a social media platform in which users post short messages ("tweets"), limited to 140 characters. By default, a message in Twitter can be read by any other user of the service, but in practice, most messages will be read by only a small subset of "followers" who have created an explicit social tie to the author. Unlike Facebook, these social network connections are unidirectional, so that celebrities may have millions of followers (Kwak, Lee, Park and Moon, 2010). Twitter is often used as a broadcast medium, when messages are intended for anyone who happens to be following the author. However, Twitter also enables dialogues in which messages can be "addressed" to a specific individual by beginning the message with her username, e.g. *@barackobama hi.* By default, such messages are not shown to other followers of the author, unless they also follow the addressee. This functionality enables Twitter to host ongoing conversations (Huberman, Romero and Wu, 2008; Honeycutt and Herring, 2009), and more than 40% of the messages in the dataset used in this paper are addressed in this way. In the analysis that follows, the distinction between broadcast messages (not addressed to any specific reader) and conversational messages (addressed to another individual by beginning the message with a username) will be shown to correlate with the use of phonologically-motivated spelling variation. Twitter also allows private "direct" messages, but by their nature these messages are inaccessible to researchers.

The dataset used in this study was gathered by Eisenstein, O'Connor, Smith and Xing (2014) from the Decahose (née Gardenhose) version of Twitter's streaming API (Application Programming Interface), over the time period from June 2009 to May 2012. Only messages containing GPS metadata are retained, so that analysis can be restricted to messages that are geolocated within the United States. The streaming API ostensibly offers a 10% sample of public posts, although Morstatter, Pfeffer, Liu and Carley (2013) show that messages containing GPS metadata are sampled at a much higher rate, roughly 90%.

A number of filters are employed so as to focus on original content. **Retweets** — repetitions of previously-posted messages — are eliminated by filtering on Twitter metadata as well as string matching on the "RT" token, which is a common practice among Twitter users to indicate a retweet (Kooti, Yang, Cha, Gummadi and Mason, 2012). To remove messages from automated marketing-oriented accounts, all tweets containing URLs were eliminated. This heuristic has the effect of removing the overwhelming majority of "spam" and other automated messages, which are typically intended to draw the reader to another website; of course, many non-automated messages are also lost. Accounts with more than 1000 followers or followees were removed for similar reasons. The application rate of each of these filters is shown in Table 1. The filtered dataset contains a total of 114 million geotagged messages from 2.77 million different user accounts.

Text contains RT or MT (retweet)	0.88%
Text contains URL hyperlink	14.19%
Author has more than 1000 followers	13.60%
Author has more than 1000 followees	10.80%
Remaining messages	71.30%

Table 1: Percentage of geotagged messages in the United States that match each filter. A message may match multiple filters, so the percentages do not sum to 100.

Text was tokenized using the publicly-available `Twokenize` program, and part-of-speech tags were obtained from the CMU Twitter Part-of-Speech Tagger; both systems are described by Owoputi, O'Connor, Dyer, Gimpel, Schneider and Smith (2013). All text was downcased, but no other textual preprocessing was performed. The corpus contains a substantial amount of non-English text, but the analyses that follow are restricted to subsets of messages that include English-language words.

This paper focuses on the phenomena of *g*-deletion and *t,d*-deletion, which can often be identified using relatively simple string-matching techniques. But as we will see, even for these variables there is ambiguity: several words, such as *wan* ('want'), had to be excluded. Other phonologically-motivated spellings might be harder to distinguish without recourse either to manual annotation or to text mining algorithms that are significantly more complex than those employed here. The analysis of *g*-deletion relies on an automatic part-of-speech tagger, but as described in Section 4, manual examination of specific cases revealed that the distribution of automatically-identified part-of-speech tags was markedly different than the distribution given by a manual annotation. Social media is known to be particularly challenging for automatic natural language processing, so such corpora require special attention (Foster, Cetinoglu, Wagner, Le Roux, Nivre, Hogan and van Genabith, 2011; Eisenstein, 2013). These observations are a reminder that while language technology can offer a valuable lever to increase the scale of sociolinguistic analysis, it is still essential to validate the main claims on manually annotated subsets of the target data.

3.2 Demographics

The demographics of Twitter have been surveyed repeatedly by the Pew Internet Research center (Duggan and Smith, 2013). In 2013, 18% of Internet users in the United States reported visiting Twitter, with similar usage rates across men and women. Blacks were significantly more likely to use Twitter, at a rate of 29%, versus 16% for Whites and Hispanics. Young people also used Twitter at a much higher rate (31% for ages 18-29, 19% for ages 30-49, and below 10% for older respondents). Differences across education level and income were not significant, but Twitter is used significantly less often in rural areas. An important caveat is that these are per-user statistics; the **per-message** demographics may be substantially different, if the usage rate also varies with demographics. More recent surveys show that these biases are becoming less pronounced (Duggan, Ellison, Lampe, Lenhart and Madden, 2015), but the 2013 demographics are most relevant to the dataset considered in this paper.

Demographic statistics can be obtained for each message, based on the United States Census statistics of the county in which the message was geolocated. In the 2010 census, there were 3,144 counties or county equivalents (in Louisiana and Alaska, the terms "parish" and "borough" are used instead), and county-level demographics display considerable variation: for example, while 78.1% of Americans identify as White, the county-level percentages range from a minimum of 3.9% to a maximum of 99.7%. The use of aggregate census demographics in place of individual records is typical in public health (e.g., Krieger, 1992), and provides a methodology for making approximate demographic inferences from sociolinguistic datasets in which speaker-level information is unavailable. An important caveat is that the distribution of authors is unlikely to be a balanced sample within each county. Therefore, these aggregated statistics can only portray the author's demographic **environment**, and do not provide an unbiased estimate of the demographics of the authors themselves.

There are several possible approaches that might come closer to capturing the exact demographics of the author set, although none is capable of completely solving the problem. Names are often informative of race and gender, and prior work has employed census statistics to compute probabilities over these attributes for individual user accounts (Chang, Rosenn, Backstrom and Marlow, 2010; Mislove et al., 2011; Bamman et al., 2014). This approach relies on the name being accurately stated; it may also unfairly downweight individuals whose names are ambiguous, and seems to offer less help for measuring socioeconomic status. An alternative and complementary source of information is the social network structure, since online social networks are often assortative on demographic dimensions (Al Zamal, Liu and Ruths, 2012). A combination of these two signals might yield increasingly accurate estimates for the race and gender of individual user accounts, but quantification of this accuracy remains a challenge: while some computational researchers have attempted to build expert-annotated datasets of the race and gender of social media users (e.g. Culotta, Ravi and Cutler, 2015), it seems a better approach to rely on self-identifications from the users themselves. As Bucholtz and Hall (2004, page 371) note, "it is crucial to attend closely to speakers' own understandings of their identities", particularly in the study of groups of which the researchers are not themselves members. While census data is also necessarily an aggregation of institutionally-defined categories, it has at least the advantage of being based on self-reports.

In this paper, the following county-level demographic statistics are considered: median income; population density, computed as the number of persons per square mile; and the proportions of individuals who identify their race or ethnicity as Asian, Black, White, and Hispanic. (The United States Census groups demonyms into lists of alternatives; this paper uses the first element of each list, thus "Black" rather than "African American", and "Hispanic" rather than "Latino".) Each demographic statistic is converted in a categorical variable by partitioning the set of instances (tweets) into quartiles. For median income and population density, the factor groups include three levels: "high", indicating the top quartile of tweets; "low", indicating the bottom quartile; and "medium", indicating the middle two quartiles. For race and ethnicity, the factor groups include two levels: "high", indicating the top quartile, and "not high", indicating the remaining three quartiles. Because the race and ethnicity factor groups are mutually competing — e.g., counties with high proportions of Blacks tend to have low proportions of Whites — the bottom quartile can be omitted as a level, simplifying interpretation of the results.

3.3 Phonological variables in Twitter

This bulk of this paper focuses on *g*-deletion and *t,d*-deletion, assessing whether the systematic patterning of their spoken equivalents is reproduced in writing. Before presenting this analysis, it is worth briefly summarizing the many other ways in which phonetic variation is reflected in Twitter writing, which demonstrates the diversity of the medium and may suggest useful directions for future research.

***th*-stopping** is transcribed in the replacement of word-initial *th* with *d* and occasionally *dh*, as in *da/dha* ('the'), *dat/dhat* ('that'), *dis/dhis* ('this'), and *doe* ('though'). In spoken language, this variable is associated with both African American English (Green, 2002) as well as a number of geographical dialects, such as New York City (Labov, 2006) and Cajun Louisiana (Dubois and Horvath, 1998). In Twitter (in the United States), this substitution is observed almost exclusively in words in which *th* represents the voiced fricative (e.g., *this*, *the*), and not the unvoiced fricative (e.g., *thank*, *thing*). Dialect respellings such as *dis* and *dat* are well attested in historical literature, and their presence on Twitter is likely influenced by existing conventions.

R-lessness is transcribed mainly in the coda, yielding frequent examples such as *holla*, *togetha*, *brotha*, *betta*, *neva*, *whateva*, and *otha*. Medial r-lessness is transcribed in *lawd* (*lord*), *yaself* (*yourself*), and *shawty* (*shorty*).

Vowel substitutions include *tha* ('the'), *ta* ('to'), *ya* and *yuh* ('you', 'your'), *naw* and *na* ('no'), *thang* ('thing'), *dranks* ('drinks'), *bruh* ('bro'), and *mayne* ('man'), as in

(1) @NAME sorry mayne, had ta bounce.

Relaxed pronunciations of individual words include *prolly* ('probably'), *aight* ('alright'), and *lil* ('little').

Multiword phenomena include *gonna*, *gunna*, *imma*, *finna*, and *fitna*. Such examples, which Preston (1985) calls "allegro spellings", may be best viewed as derived from the system of tense and aspect in African American English (Green, 2002). Other examples from outside the tense and aspect system include *wassup* ('what's up'), *whatcha* ('what are you'), *wanna* ('want to'), *bouta*, *tryna* ('trying to'), *ion* ('i don't'), and *iono* ('i don't know').

alternative spelling	**rate**	**gloss**	**relative freq.**
wanna	1063	'want to'	1.81
ya	1157	'you'	0.0937
da	1507	'the'	0.0399
dat	2652	'that'	0.0570
tryna	4037	'trying to'	0.783
dis	4542	'this'	0.0460
tha	5459	'the'	0.0110
nah	6247	'no'	0.0621
naw	7904	'no'	0.0491
wassup	9286	'what's up'	0.903
bruh	11076	'bro'	0.260
doe	12462	'though'	0.191
dats	12594	'that's'	0.0609
na	13690	'no'	0.0283
prolly	13817	'probably'	0.350
betta	13817	'better'	0.0846
neva	15167	'never'	0.0693
aight	17098	'alright'	0.526
ta	17918	'to'	0.00352
holla	18415	'holler'	11.4
bouta	20931	'about to'	0.140
shawty	21266	'shorty'	1.68
yuh	24441	'you'	0.00444
thang	25429	'thing'	0.0682
ion	27101	'i don't'	0.0374
dha	47427	'the'	0.00127
lawd	47616	'lord'	0.177
brotha	49799	'brother'	0.130
otha	59461	'other'	0.0379
dhat	92649	'that'	0.00163
mayne	101286	'man'	0.00877
whatcha	107674	'what are you'	0.194
yaself	119518	'yourself'	0.0663
whateva	127147	'whatever'	0.0672
iono	161511	'i don't know'	0.0496
watcha	202573	'what are you'	0.103
dhis	298795	'this'	0.000699
togetha	412131	'together'	0.0131
dranks	796788	'drinks'	0.0196

Table 2: Transcripts of some of the more frequent phonetic variables not covered by the two main cases in this paper, *g*-deletion and *t,d*-deletion. "Rate" indicates the number of tokens per incidence of each word. "Relative freq." indicates the ratio of the counts for the alternative form divided by the sum of the counts for the gloss. Some cases, such as *doe* and *ion*, have other senses, so this ratio must be interpreted with caution.

The frequencies for some of the more common phonologically-inspired orthographic variables are shown in Table 2. This list was selected by identifying words in the Twitter corpus which were not present in the `Aspell` Unix spelling dictionary, and manually searching in order of frequency. In the table, "rate" refers to the number of tokens per instance of the word, so *wanna* appears roughly once per 1078 tokens; more often, in fact, than the bigram *want to*. Left to future work is the question of whether these variables also display systematic patterning that matches their spoken forms.

4 *g*-DELETION

The (ing) variable corresponds to alternation between the coronal and velar nasal, and is commonly referred to as "g-dropping." It is used in spoken language throughout the English-speaking world, and is often associated with informal or relaxed speech (Trudgill, 1974). The variable observes a grammatical constraint: the *-in* form is most frequent in verbs (especially the progressive), and less frequent in nouns and adjectives (Fischer, 1958). In addition, the *-in* form is only possible for unstressed syllables, and therefore cannot apply to monosyllabic words such as *ring* and *thing* (Houston, 1985).

Deletion of the final *g* in *-ing* ending words — henceforth, "*g*-deletion" — is widespread in the Twitter corpus: for example, the word *goin* is the 292nd most frequent overall (just behind *try* and *coming*, and ahead of *live* and *might*), appearing at a rate of roughly once per 220 messages. Do the phonological and syntactic constraints from spoken language apply to the (ing) variable in Twitter writing? This question was posed using a subsample of the corpus described above, consisting of one hundred thousand (10^5) tweets. Each tweet contains at least one of the 200 most frequent *-ing* ending words, in either *-ing* or *-in* form. Three words were excluded because the shortened form was also a commonly-used word: *sing*, *thing*, and *king*. The CMU Twitter Part-of-Speech Tagger (Owoputi et al., 2013) was used to identify the coarse-grained grammatical category for each token. This software automatically selects from a set of 25 possible tags for each word, but only three tags are of interest here: N (common noun), V (verb), and A (adjective).[2]

4.1 Frequency analysis

Figure 1 shows the probability of *g*-deletion across all 197 words, with color and shape determined by the most common grammatical category for each word type. Verbs generally show a higher deletion rate than nouns or adjectives, with the notable exceptions of *fucking* and its "anti-swear" companion, *freaking*. While these words are mostly commonly used as adjectives, their verb senses are also frequent: according to the automatic part-of-speech tagger, 47% of the usages of *fuckin* and 44% of the usages of *freakin* are tagged as verbs. In 200 randomly-selected examples of the word *fuckin*, 38 were manually annotated as verbs; for 200 randomly-selected examples of the word *fucking*, only seventeen were annotated as

[2] Unlike many other tagsets, such as the one used in Penn Treebank (Marcus, Marcinkiewicz and Santorini, 1993), verb inflections and noun number are not differentiated by the CMU Twitter Part-of-Speech tagger.

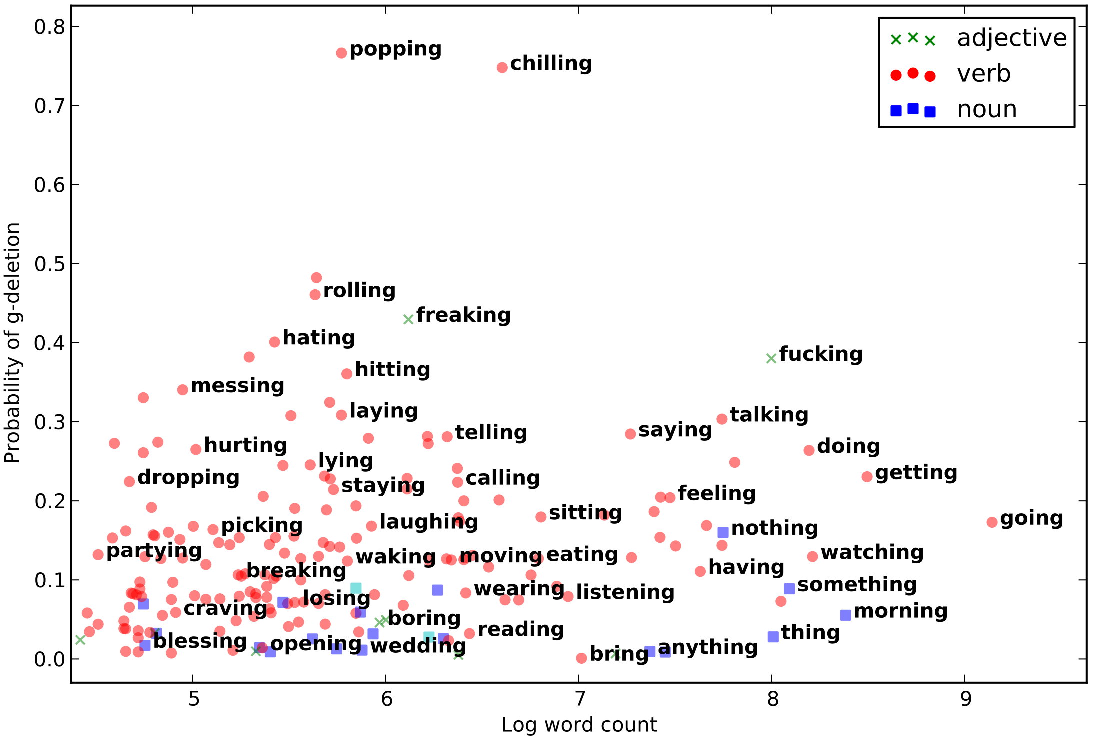

Figure 1: Probability of *g*-deletion across words in the dataset. Each mark corresponds to a word, with the shape indicating the word's most common grammatical category. The x-axis corresponds to the logarithm of the total count for each word. To maintain legibility, only a subset of words are written out.

verbs. The low number of verbs in the manual annotation suggests that the tagger may misclassify these words at a higher rate than the overall 10% error rate reported by Owoputi et al. (2013). But even with these manual part-of-speech annotations, the verb form is significantly more likely to be reduced ($p < .01$ by Fisher's exact test), demonstrating that the grammatical constraint operates between multiple grammatical categories for a single written word.

The verbs with the highest rate of *g*-deletion are *popping* and *chilling*, and not the progressive *going*. In an informal register, the very common bigram *going to* can be further reduced from *goin to* to *gonna*, which suppresses the counts for the shortened form *goin*. This issue will arise again in the discussion of phonologically-conditioned variation in the next section; for now, it helps to explain the relatively low frequency of *g*-deletion for this word. To understand the high frequency of deletion for the words *popping* and *chilling*, consider some examples of these words in use:

(2) @NAME what's poppin w u .. ('what's happening with you')

(3) @NAME man dat shit was poppin last night

(4) @NAME chillin .. at my 3rd cookout #fat

(5) gwenn stefannii is chillin with us

These terms are frequently used to describe informal social situations; the bigram *what's poppin* accounts for roughly 20% of the occurrences of *poppin*. The popularity of *g*-deletion for these words suggests that the written form of this variable maintains the same social function as its spoken form, connoting informal speech.

4.2 Multivariate analysis

A mixed effects logistic regression offers a more precise characterization of how multiple linguistic and social factors modulate the frequency of *g*-deletion in social media writing. The following factor groups were considered:

Grammatical category For each instance, the **token-level** part-of-speech label is used as a predictor. Note that this differs from Figure 1, which categorizes words by the most frequently-tagged category for the word **type**.

Number of syllables The number of syllables for each word type was identified by consulting the CMU Pronouncing Dictionary (version 0.7a), as included in the NLTK Natural Language Toolkit (Bird, Klein and Loper, 2009). In cases with multiple pronunciations, the one with the fewest syllables was selected.

Message type A binary predictor indicates whether the message is conversational or broadcast (see Section 3.1). Since conversational messages begin with *@username*, the notation "@-message" is used to refer to such messages throughout the remainder of the paper.

Demographics County-level statistics for race, ethnicity, income, and population density are all included as categorical variables; see Section 3.2 for details.

Random effects Recent methodological work has demonstrated the utility of **mixed effects models**, which can correct for idiosyncratic lexical variation and account for variability across speakers or writers (Gorman and Johnson, 2013). Random intercepts were included for each of the 197 words, for the four years in which the data was collected (2009-2012), and for the 538 authors with at least twenty instances in the dataset. The purpose of author random effects is to account for the disproportionate impact of a few high-volume writers. It is not necessary to include random effects for every author, since low-volume authors will exert a negligible impact on the estimated coefficients for the fixed effects. This can be seen as a form of variable selection, which

	Weight	Log odds	%	N
Linguistic features				
Verb	.556	.227	.200	89,173
Noun	.497	-.013	.083	18,756
Adjective	.447	-.213	.149	4,964
Monosyllable	.071	-2.57	.001	108,804
Message type				
@-message	.534	.134	.205	36,974
County demographics				
High % White	.452	-.194	.117	28,017
High % Black	.536	.145	.241	27,022
High % Asian	.491	-.034	.190	27,444
High % Hispanic	.493	-.029	.174	26,438
High pop density	.514	.055	.228	27,773
Low pop density	.496	-.017	.144	28,228
Total			.178	112,893

Table 3: mixed effects logistic regression of factors for *g*-deletion. Median county income was not selected as statistically significant in either forward selection or backward elimination. As is traditional in variable rules analysis, the first column ("weight") corresponds to the logistic transformation of the estimated log odds.

is frequently employed in high-dimensional settings (Hastie, Tibshirani and Friedman, 2009).

All factor groups shown in Table 3 were selected as statistically significant at $p < .01$; the county's median income was not selected as significant in either forward selection or backward elimination. The likelihood of *g*-deletion in written language is jointly affected by syntax (part-of-speech; $p \ll .001$), phonetics (number of syllables; $p \ll .001$), pragmatics (message type; $p \ll .001$), and demographics ($p < .01$ for % Hispanic; stronger p-values were obtained for all other demographic variables shown in Table 3). The role of pragmatics — favoring conversational messages over broadcasts — coheres with the high frequency of *g*-deletion in words that suggest informal contexts, as in Examples (2-5). Demographically, the strongest effect is racial: *g*-deletion spelling is inhibited in counties with large proportions of Whites, and occurs more frequently in counties with large numbers of Blacks. Similar demographic patterning is reported for the (ing) variable in spoken language (Shuy, Wolfram and Riley, 1967; Green, 2002).[3]

[3] The county demographics are not independent of message types: counties with high population density and with high proportions of minorities, particularly Hispanics, tended to use more conversational messages.

To compare coefficients for levels within a factor group, parametric bootstrap confidence intervals were computed using the `lme4` package (Bates, Mächler, Bolker and Walker, 2014). All pairwise differences between grammatical categories are statistically significant, with the least certainty in the comparison between nouns and verbs ($p < .01, z \approx 2.5$, one-tailed). If we discard the random intercepts for words, then the relationship between nouns and adjectives is reversed, with nouns appearing to inhibit *g*-deletion more than adjectives. However, this effect is largely due to a single word — *fucking* — which shows an unusually high frequency of *g*-deletion, as discussed above. The robustness to such lexical outliers is a key advantage of mixed effects models (Gorman and Johnson, 2013).

5 *t,d*-DELETION

The (t,d) variable corresponds to the deletion of the final stop in words such as *just*, *don't*, and *passed*. This variable has been observed in a number of English dialects, including African American English (Labov, Cohen, Robins and Lewis, 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). The (t,d) variable is a classic example of patterned 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). I will focus on just one contextual factor, the inhibition of (t,d) when the following phonological context is a vowel: for example, the final *t* in *left* is more likely to be deleted in *I left* ***the*** *house* than in *I left* ***a*** *tip*. Guy (1991, page 233) 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). If dialect respellings in social media are purely lexical or graphemic, then we would predict that written *t,d*-deletion would occur at roughly equal rates, regardless of the following phonological context.

5.1 Frequency analysis

To isolate the potential impact of phonology, the quantitative analysis focuses on words that are unambiguous under *t,d*-deletion, and for which the deletion is clearly phonological – excluding words such as *missed*, because the reduction to *miss* could be caused by omission of the past tense rather than deletion of the final phoneme. The following five deletions were selected by examining non-dictionary words in order of frequency: *jus(t)*, *ol(d)*, *ain(t)*, *nex(t)*, and *tol(d)*. Excluded words include: *an(d)*, because *an* is a frequently-used standard word; *bes(t)*, because *bes* is often used to reference *Blackberry Enterprise Server* in this corpus;

However, even after controlling for message type, the differences in the rate of *g*-deletion across county-level demographics were still significant (by Fisher's exact test); conversely, the differences by message type were still significant even after controlling for demographics.

and *wan(t)*, because *wan* is mainly used in constructions such as *wan go*, and because it is occasionally used in both the standard sense and as an acronym for *wide area network*.

Instances are selected from messages that contain either the standard or shortened forms, excluding cases in which the word appears as the final token (pause contexts are not considered in this study). The phonological context is determined by searching for the following token in the CMU Pronouncing Dictionary (version 0.7a), as included in the NLTK Natural Language Toolkit (Bird, Klein and Loper, 2009). If the following token does not appear in the dictionary, the entire message is excluded. The frequency and count for the five selected *t,d*-deletion cases are shown in Table 4. A Fisher's exact test shows that the deletion rate differences are statistically significant at $p < .01$ for *just*, *old*, and *told*, but not for *aint* and *next*. While the difference in the combined counts is statistically significant at $p \ll .001$, this is unsatisfying because it is dominated by the most frequent word, *jus(t)*.

As a point of comparison, two alternative groups of coda character deletions are considered. From the *-ing* ending words, I again select the five most frequent unambiguous cases: *goin(g)*, *gettin(g)*, *fuckin(g)*, *talkin(g)*, and *doin(g)*. Also considered are words whose coda character deletion does not appear to reflect any phonetic difference: *kno(w)*, *love/luv*, *hav(e)*, *tru(e)*, and *mayb(e)*. For these spellings, which are a form of **eye dialect** (Preston, 1985), we expect no sensitivity to the phonological context. Nonetheless, the differences in deletion rate, while proportionally smaller than the strongest cases of *t,d*-deletion, are all statistically significant, due to the high counts.

More confusingly, the patterns of deletion within the *-ing* group seem to point in opposite directions: for example, *goin* appears much more often in vowel contexts, while *talkin* appears significantly less often. On reflection, it should be clear that such counts and frequencies can easily be confounded by lexical effects. For example, *going* is most frequently followed by the word *to*, and then by *on* and *in*. As noted above, the bigram *goin to* can be further reduced to *gonna*, which suppresses the counts for *goin* followed by a consonant. Because there is no equivalent shortened form for *goin on* or *goin in*, the deletion rate appears to be much higher in vowel contexts. A related phenomenon helps to explain the frequency of deletion for *true*, where the deleted form is 2.5 times more frequent before the first-person pronoun *I*, and 5.6 times more frequent before the shortened form of the second-person pronoun, *u*: e.g., *@NAME tru u did tweet me this morn.* By adding bigrams as random effects, it is possible to isolate the systematic phonological patterns that would otherwise be obscured by these confounds.

5.2 Multivariate analysis

For mixed effects logistic regression analysis, the dataset was subsampled to obtain at most 10,000 messages for each spelling. As there were only a few thousand examples of *tol* and *nex* in the original dataset, no subsampling was performed for these words. The subsampling procedure selects entire messages, so for some words, more than 10,000 examples are obtained, because some messages contain more than one instance of a shortened form. Three

	Before consonants		**Before vowels**	
	Del. rate	**count**	**Del. rate**	**count**
(t,d)				
just	.0853	4,989,358	.0735	552,349
next	.00499	634,397	.00474	29,343
aint	.0189	477,190	.0195	78,641
old	.0912	393,554	.0368	82,664
told	.0101	353,037	.00825	26,300
(ng)				
going	.153	1,640,855	.285	419,337
getting	.238	775,190	.255	312,230
doing	.258	513,059	.248	185,161
fucking	.379	552,574	.363	88,108
talking	.429	311,293	.193	230,923
non-phonetic				
have	.00758	2,930,902	.00890	1,306,755
know	.139	1,825,722	.143	727,724
love	.0376	1,985,194	.0397	421,392
maybe	.0288	217,273	.0323	166,259
true	.0845	153,850	.112	26,281

Table 4: Frequency and count of final character deletions across three ending groups, depending on subsequent phonological context.

(t,d)	**Weight**	**Log odds**	**%**	**N**
vowel succeeding context	.483	-.066	.385	9,004
@-message	.519	.075	.436	35,240
County demographics				
High % White	.422	-.313	.311	19,992
High % Black	.516	.065	.508	19,854
High % Hispanic	.480	-.080	.388	19,090
High median income	.473	-.107	.388	20,653
Low median income	.532	.127	.482	25,386
High pop density	.505	.021	.456	21,295
Low pop density	.520	.078	.416	22,870
total			**.423**	**89,174**

Table 5: Mixed effects logistic regression on coda character deletion for the *t,d*-deletion words *just, next, aint, told,* and *old*. The percentage of Asian Americans in the county was not selected as statistically significant in either forward selection or backward elimination. Vowel succeeding context was selected as statistically significant at $p < .01$.

separate analyses were run using `Rbrul`, one for each word group. In each analysis, the following factor groups were considered:

Phonological context Vowel and consonant contexts are distinguished using a binary predictor.

Message type and demographics The same predictors are included here as in the analysis of *g*-deletion.

Random effects To address the concerns about lexical conditioning mentioned above, all 2665 word bigrams that appear at least ten times (across all ending groups) were included as random effects. Bigrams that are more rare will make little impact on the estimates for the fixed effects. As in the analysis of *g*-deletion, authors and years are included as random effects.

Results are shown in Tables 5, 6, and 7. Both *t,d*-deletion and *g*-deletion are sensitive to the phonological context ($p < .01$ in both cases). In contrast, "eye dialect" spellings such as *kno* and *luv* are not sensitive to this factor group ($p \approx .6$) — despite the fact that the differences in raw deletion rate were significant. The insensitivity of the control group to the phonological context is encouraging, suggesting that the statistical significance of phonological conditioning in the other word groups can indeed be attributed to the phonetic character of these variables, and not simply to the large sample size.

(ing)	**Weight**	**Log odds**	**%**	**N**
vowel succeeding context	.479	-.082	.477	27,005
@-message	.532	.082	.532	36,886
County demographics				
High % White	.454	-.184	.394	26,681
High % Black	.512	.046	.576	19,793
High % Asian	.492	-.032	.515	24,895
High % Hispanic	.478	-.086	.474	22,960
High median income	.480	-.079	.485	25,756
Low median income	.522	.089	.540	26,693
High pop density	.521	.083	.568	24,096
Low pop density	.481	-.075	.437	25,777
total			**.498**	102,262

Table 6: Mixed effects logistic regression on coda character deletion for the *g*-deletion words *going, getting, doing, fucking,* and *talking.* Vowel succeeding context was selected as statistically significant at $p < .01$.

The ranges of the weights for the phonological factor groups are considerably smaller than those reported in prior work. Tagliamonte and Temple (2005) obtain a range of .60 by distinguishing obstruents, glides, nasals, and liquids, rather than grouping them as consonants, as done here; however, Guy (1991) groups all consonants and still obtains a range of .42. Another possible reason for the smaller range is that mixed effects models are more "conservative," because they allocate weight both to fixed effects as well the lexical random effects (Johnson, 2009). But given the relatively small differences in raw frequencies (Table 4), it seems indisputable that the impact of phonology on social media spellings, while statistically significant, is weaker than in speech. Jaffe (2000, page 506) argues that "few texts that make use of non-standard orthographic forms apply these forms consistently or rigorously," and with less consistency, the effects of phonologically-motivated contextual patterning will necessarily be less evident.

Demographically, all three forms of coda deletion are patterned similarly, with lower frequency in counties that have high proportions of Whites and a high medium income, and higher frequency in counties that have high proportions of Blacks and a low medium income. This coheres with previous studies on (t,d) in spoken language (Labov et al., 1968; Green, 2002). Counties with high proportions of Hispanics exert an inhibitory effect on coda deletion in all three cases; the effect for Asian Americans is generally weak, and falls below the level of statistical significance for the *t,d*-deletion words. The role of population density is more mixed: the non-phonetic deletions are most strongly favored in high-density

non-phonetic	**Weight**	**Log odds**	**%**	**N**
vowel succeeding context	[.503]	[.011]	.518	28,864
@-message	.516	.065	.526	50,707
County demographics				
High % White	.439	-.245	.354	25,974
High % Black	.529	.118	.601	21,304
High % Asian	.491	-.036	.533	26,493
High % Hispanic	.490	-.038	.487	23,775
High median income	.496	-.017	.497	26,918
Low median income	.517	.066	.541	27,106
High pop density	.551	.204	.606	26,672
Low pop density	.401	-.143	.464	25,134
total			**.495**	103,575

Table 7: mixed effects logistic regression on non-phonetic coda character deletion for the words *have, know, love, maybe,* and *true.* Vowel succeeding context is not selected as statistically significant ($p \approx .6$), but is shown here for comparison with the other two results tables. The weakest selected predictor is the percentage of Hispanics in the county, which is $p < .001$.

counties and disfavored in low-density counties; the effect is similar but weaker for the *g*-deletion words, and is inverted for the *t,d*-deletion words, where medium-density counties exert the strongest inhibitory force (the coefficient is -.097). In all cases, coda deletion is significantly more frequent in conversationally addressed "@-messages"; just as we saw earlier in the analysis of *g*-deletion, the conversational nature of these messages seems to encourage more informal spellings.

6 Conclusion

The quantitative analyses in this paper paint a consistent picture of the nature of phonologically-motivated respellings in online social media: when alternative spelling is linked to phonetic variation, it acquires at least the residue of the systems of phonological, grammatical, and social patterning present in speech. The analysis of *g*-deletion shows that phonologically-motivated spelling variation is not simply a matter of graphemic substitution: nouns and verbs show deletion at significantly different rates, even for different senses of the same word. The analysis of *t,d*-deletion indicates that phonological context plays a role in spelling variation, operating in the same way as in spoken language — although the force of phonological context is weaker in spelling variation than in speech.

In both *g*-deletion and *t,d*-deletion, we also see echoes of the system of socially-linked variation from spoken language, with higher frequencies of non-standard spellings in counties that include more individuals who identify as African Americans, and more standard spellings in counties that include more individuals who identify as White. This mirrors the pattern of social variation for the spoken-language variables of (ing) and (t,d), although it is important to observe that roughly the same demographic pattern holds for the "eye dialect" variables of *luv* and *kno* (Table 7). The writing style seen in these heavily African-American counties may therefore reflect multiple causal factors: both the desire to replicate phonological features of African American English, as well as a greater resistance to the pressures of language standards in writing. Other aspects of social patterning are less consistent across variables: the eye dialect variables are significantly more frequent in high-density (urban) counties, while the *t,d*-deletion variables are more frequent in low-density counties. Eye dialect spellings are not linked to spoken variation, and their widespread use may be a more recent innovation. If written variation follows the same pattern of urban innovation often identified in speech (Labov, 2001), this might explain the greater prevalence of these spellings in urban areas.

Finally, we observe a higher likelihood of non-standard spellings in messages that are addressed to individual readers rather than to the Twitter audience at large. This indicates that social media users customize their linguistic self-presentation depending on who they think the reader is likely to be, providing new evidence for audience design on a large scale (Bell, 1984), and linking to recent studies on code-switching between languages (Androutsopoulos, 2013; Johnson, 2013). Given the social meaning of (ing) and (t,d) in speech, this also suggests that social media users assign lower formality to these more closely-directed "conversational" messages, providing a new, quantitative perspective on previous qualitative CMC research about the adaptation of writing styles to the audience (e.g., Marwick and boyd, 2011). Forthcoming work by Pavalanathan and Eisenstein (2015) presents related evidence on the interaction between lexical variation and audience design.

With these findings in hand, we return to the question of what it means to describe a respelling as "phonetic". A simplistic model in which phonological variable rules are transferred to the associated graphemes fails to account for the syntactic patterning of *g*-deletion. Even a more complex model, in which phonetic variables apply at the level of spellings of individual words, fails to account for the phonological patterning of *t,d*-deletion, which is influenced by phonological factors beyond the word boundary. Conventionalization is also an unlikely explanation, as this would imply encoding permissible alternative spellings for thousands or millions of word pairs. One possible solution is that the writing process involves first conceptualizing the utterance in an "inner speech" (Vygotsky, 1987) that is itself shaped by the system of phonological and grammatical variation. This would require a remarkably strong hypothesis of cognitive primacy for speech, suggesting that the process of writing operates on a cognized utterance that already contains phonetic variation, and that **standard writing** — which even heavy users of non-standard orthography are capable of producing (Tagliamonte and Denis, 2008) — requires eliminating these variables. A more plausible explanation is that the production of phonologically-motivated non-standard spellings requires the writer to assess, whether consciously or not, the phonological coherence of the utterance as a whole. This assessment could take place in synchrony with production of writing, or in a *post hoc* editing process. Techniques such as keystroke logging (Leijten and Van Waes, 2013) might offer new insights on the processes underlying the production of phonologically-motivated orthographic variation.

In any case, what is clear is that the use of non-standard orthography as a form of social action does not license just **any** spelling: rather, the result must be comprehensible, both semantically and stylistically. When the non-standard form references a phonetic variable, it is not enough to respell the word so as to index that variable; the variable must be employed in a way that fits in with the system of phonological variation. This leads to the question of whether this systematic patterning impacts all authors to the same extent. For example, it is possible that phonologically-motivated respelling are employed by authors who do not produce the corresponding variable in speech, or for whom contact with this variable occurred as adults. An analysis of such individuals might reveal whether systematicity in written language depends on the writer's degree of familiarity with the associated spoken language variable. Quantitatively, a random slopes model could account for variation in the degree to which phonological and syntactic patterning is relevant for different authors (Gorman and Johnson, 2013), but this investigation must be left for future work.

Finally, with the aforementioned methodological limitations in mind, this study demonstrates the potential of large social media corpora for the analysis of language variation. The size of these corpora enables the study of rare lexical phenomena, and the identification of relatively subtle effects. Moreover, the **nature** of this data offers a unique perspective on writing, capturing an ongoing cultural shift towards the use of literate forms for phatic communication. Inherent in this shift is a renegotiated relationship between writing and speech; by uncovering the terms of this negotiation, we can obtain new insights on the role of each modality in the production of language variation.

References

Al Zamal, Faiyaz, Wendy Liu and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 387–390. Menlo Park, California: AAAI Publications.

Androutsopoulos, Jannis. 2000. Non-standard spellings in media texts: The case of German fanzines. *Journal of Sociolinguistics* 4: 514–533.

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

Androutsopoulos, Jannis. 2013. Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism* .

Anis, Jacques. 2007. Neography: Unconventional spelling in French SMS text messages. In Brenda Danet and Susan C. Herring (eds.), *The Multilingual Internet: Language, Culture, and Communication Online*, 87–115. Oxford University Press.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday* 12.

Ball, Arnetha F. 1992. Cultural preference and the expository writing of African-American adolescents. *Written Communication* 9: 501 – 532.

Bamman, David, Jacob Eisenstein and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18: 135–160.

Baron, Naomi S. 2004. See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23: 397–423.

Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. Technical Report 1406.5823, ArXiv e-prints.

Bayley, Robert. 1994. Consonant cluster reduction in Tejano English. *Language Variation and Change* 6: 303–326.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13: 145–204.

Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural language processing with Python.* California: O'Reilly Media.

boyd, danah and Kate Crawford. 2012. Critical questions for big data. *Information, Communication & Society* 15: 662–679.

Bucholtz, Mary. 2007. Variation in transcription. *Discourse Studies* 9: 784–808.

Bucholtz, Mary and Kira Hall. 2004. Language and identity. In Alessandro Duranti (ed.), *A Companion to linguistic anthropology*, 369–394. Blackwell.

Chang, Jonathan, Itamar Rosenn, Lars Backstrom and Cameron Marlow. 2010. ePluribus: Ethnicity on social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 18–25. Menlo Park, California: AAAI Publications.

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

Crystal, David. 2006. *Language and the Internet.* Second edition. Cambridge University Press.

Culotta, Aron, Nirmal Kumar Ravi and Jennifer Cutler. 2015. Predicting the demographics of Twitter users from website traffic data. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, in press. Menlo Park, California: AAAI Press.

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 98–106. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Dresner, Eli and Susan C Herring. 2010. Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory* 20: 249–268.

Dubois, Sylvie and Barbara M Horvath. 1998. Let's tink about dat: Interdental fricatives in Cajun English. *Language Variation and Change* 10: 245–261.

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart and Mary Madden. 2015. Social media update 2014. Technical report, Pew Research Center.

Duggan, Maeve and Aaron Smith. 2013. Social media update 2013. Technical report, Pew Research Center.

Eisenstein, Jacob. 2013. What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 359–369. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 1277–1287. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9.

Fischer, John L. 1958. Social influences on the choice of a linguistic variant. *Word* 14: 47–56.

Foster, Jennifer, 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 the International Joint Conference on Natural Language Processing (IJCNLP)*, 893–901. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Gorman, Kyle and Daniel Ezra Johnson. 2013. Quantitative analysis. In *The Oxford Handbook of Sociolinguistics*, 214–240. Oxford, U.K.: Oxford University Press.

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

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

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning*. Second edition. New York: Springer.

Hecht, Brent and Monica Stephens. 2014. A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 197–205. Menlo Park, California: AAAI Publications.

Herring, Susan C. 2012. Grammar and electronic communication. In Carol A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. Wiley.

Herring, Susan C. and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10: 439–459.

Hinrichs, Lars and Jessica White-Sustaíta. 2011. Global Englishes and the sociolinguistics of spelling: A study of Jamaican blog and email writing. *English World-Wide* 32: 46–73.

Honeybone, Patrick and Kevin Watson. 2013. Salience and the sociolinguistics of Scouse spelling: Exploring the phonology of the contemporary humorous localised dialect literature of Liverpool. *English World-Wide* 34.

Honeycutt, Courtenay and Susan C. Herring. 2009. Beyond microblogging: Conversation and collaboration via Twitter. In *Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS)*, 1–10. Los Alamitos, California: IEEE Computer Society.

Houston, Ann Celeste. 1985. *Continuity and change in English morphology: The variable (ING)*. Ph.D. thesis, University of Pennsylvania.

Huberman, Bernardo, Daniel M. Romero and Fang Wu. 2008. Social networks that matter: Twitter under the microscope. *First Monday* 14.

Jaffe, Alexandra. 2000. Introduction: Non-standard orthography and non-standard speech. *Journal of Sociolinguistics* 4: 497–513.

Jaffe, Alexandra, Jannis Androutsopoulos, Mark Sebba and Sally Johnson. 2012. *Orthography as Social Action: Scripts, Spelling, Identity and Power*. Berlin: Walter de Gruyter.

Jaffe, Alexandra and Shana Walton. 2000. The voices people read: Orthography and the representation of non-standard speech. *Journal of Sociolinguistics* 4: 561–587.

Johnson, Daniel Ezra. 2009. Getting off the goldvarb standard: Introducing rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–383.

Johnson, Ian. 2013. Audience design and communication accommodation theory: Use of Twitter by Welch-English biliterates. In Elin Haf Gruffydd Jones and Enrique Uribe-Jongbloed (eds.), *Social Media and Minority Languages: Convergence and the Creative Industries*, 99–118. Bristol, U.K.: Multilingual Matters.

Kalman, Yoram M. and Darren Gergle. 2014. Letter repetitions in computer-mediated communication: A unique link between spoken and online language. *Computers in Human Behavior* 34: 187–193.

Kooti, Farshad, Haeryun Yang, Meeyoung Cha, P. Krishna Gummadi and Winter A. Mason. 2012. The emergence of conventions in online social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 194–201. Menlo Park, California: AAAI Publications.

Krieger, Nancy. 1992. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *American Journal of Public Health* 82: 703–710.

Kwak, Haewoon, Changhyun Lee, Hosung Park and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the Conference on World-Wide Web (WWW)*, 591–600. New York: ACM.

Labov, William. 2001. *Principles of Linguistic Change*, volume 2: Social Factors. Wiley-Blackwell.

Labov, William. 2006. *The social stratification of English in New York City*. Cambridge, U.K.: Cambridge University Press.

Labov, William, Sharon Ash and Charles Boberg. 2006. *Atlas of North American English: Phonetics, Phonology, and Sound Change*. Berlin: Mouton de Gruyter.

Labov, William, 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.

Leijten, Mariëlle and Luuk Van Waes. 2013. Keystroke logging in writing research using inputlog to analyze and visualize writing processes. *Written Communication* 30: 358–392.

Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19: 313–330.

Marwick, Alice E. and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13: 114–133.

Miethaner, Ulrich. 2000. Orthographic transcriptions of non-standard varieties: The case of earlier african-american english. *Journal of Sociolinguistics* 4: 534–560.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 554–557. Menlo Park, California: AAAI Publications.

Morstatter, Fred, 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 the International Conference on Web and Social Media (ICWSM)*, 400–408. Menlo Park, California: AAAI Publications.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 380–390. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Paolillo, John C. 2001. Language variation on internet relay chat: A social network approach. *Journal of Sociolinguistics* 5: 180–213.

Pavalanathan, Umashanthi and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech* (in press).

Preston, Dennis R. 1985. The Li'l Abner syndrome: Written Representations of Speech. *American Speech* 60: 328–336.

Santa Ana, Otto. 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.

Schnoebelen, Tyler. 2012. Do you smile with your nose? Stylistic variation in Twitter emoticons. *University of Pennsylvania Working Papers in Linguistics* 18: 14.

Sebba, Mark. 2007. *Spelling and Society: The Culture and Politics of Orthography around the World*. Cambridge, U.K.: Cambridge University Press.

Shortis, Tim. 2007. Revoicing txt: Spelling, vernacular orthography and 'unregimented writing'. In *The texture of internet: Netlinguistics in progress*, 2–23. Newcastle: Cambridge Scholars Publishing.

Shuy, Roger W, Walt Wolfram and William K Riley. 1967. Linguistic correlates of social stratification in Detroit speech. Technical Report 6-1347, United States Office of Education.

Tagliamonte, Sali A. and Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83: 3–34.

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

Thompson, Connie A., 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: 269–282.

Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3: 215–246.

Varnhagen, Connie K., G. Peggy McFall, Nicole Pugh, Lisa Routledge, Heather Sumida-MacDonald and Trudy E. Kwong. 2010. Lol: New language and spelling in instant messaging. *Reading and Writing* 23: 719–733.

Vygotsky, Lev. 1987. *Thought and Language*. Cambridge, Massachusetts: MIT Press.

Walther, Joseph B. and Kyle P. D'Addario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review* 19: 324–347.

Werry, Christopher C. 1996. Linguistic and interactional features of internet relay chat. In Susan C. Herring (ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-cultural Perspectives*, 47–63. Amsterdam: John Benjamins.

Whiteman, Marcia Farr. 1982. Dialect influence in writing. In Marcia Farr Whiteman (ed.), *Writing: The Nature, Development, and Teaching of Written Communication*, volume 1: Variation in Writing. New York: Routledge.

Yates, Simeon J. 1996. Oral and written linguistic aspects of computer conferencing: A corpus based study. *Computer-Mediated Communication: Linguistic, social, and cross-cultural perspectives* 39: 29.