# What to do about bad language on the internet

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

## Abstract

The rise of social media has brought computational linguistics in ever-closer contact with *bad language*: text that defies our expectations about vocabulary, spelling, and syntax. This paper surveys the landscape of bad language, and offers a critical review of the NLP community's response, which has largely followed two paths: normalization and domain adaptation. Each approach is evaluated in the context of theoretical and empirical work on computer-mediated communication. In addition, the paper presents a quantitative analysis of the lexical diversity of social media text, and its relationship to other corpora.

## 1 Introduction

As social media becomes an increasingly important application domain for natural language processing, we encounter language that is substantially different from many benchmark corpora. The following examples are all from the social media service Twitter:

- *Work on farm Fri. Burning piles of brush WindyFire got out of control. Thank God for good naber He help get undr control Pants-BurnLegWound.* (Senator Charles Grassley)
- *Boom! Ya ur website suxx bro*
(Sarah Silverman)
- *...dats why pluto is pluto it can neva b a star*
(Shaquille O'Neil)
- *michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.*
(Ozzie Guillen)

These examples are selected from celebrities (for privacy reasons), but they contain linguistic challenges that are endemic to the medium, including non-standard punctuation, capitalization, spelling, vocabulary, and syntax. The consequences for language technology are dire: a series of papers has detailed how state-of-the-art natural language processing (NLP) systems perform significantly worse on social media text. In part-of-speech tagging, the accuracy of the Stanford tagger (Toutanova et al., 2003) falls from 97% on Wall Street Journal text to 85% accuracy on Twitter (Gimpel et al., 2011). In named entity recognition, the CoNLL-trained Stanford recognizer achieves 44% F-measure (Ritter et al., 2011), down from 86% on the CoNLL test set (Finkel et al., 2005). In parsing, Foster et al. (2011) report double-digit decreases in accuracy for four different state-of-the-art parsers when applied to social media text.

The application of language technology to social media is potentially transformative, leveraging the knowledge and perspectives of millions of people. But to deliver on this potential, the problems at the core of the NLP pipeline must be addressed. A growing thread of research takes up this challenge, including a shared task and workshop on "parsing the web," with new corpora which appear to sit somewhere between the Wall Street Journal and Twitter on the spectrum of bad language (Petrov and McDonald, 2012). But perhaps surprisingly, very little of this research has considered *why* social media language is so different. This review paper attempts to shed some light on this question, surveying a strong tradition of empirical and theoreti-



*Proceedings of NAACL-HLT 2013*, pages 359–369,
Atlanta, Georgia, 9–14 June 2013. 

cal research on computer-mediated communication (CMC). I argue that the two main computational approaches to dealing with bad language — normalization and domain adaptation — are based on theories of social media language that are not descriptively accurate. I have worked and continue to work in both of these areas, so I make this argument not as a criticism of others, but in a spirit of self-reflection. It is hoped that a greater engagement with sociolinguistic and CMC research will lead to new, nuanced approaches to the challenge of bad language.

**Why so much Twitter?** Most of the examples in this paper will focus on Twitter, a microblogging service. Munro and Manning (2012) argue that Twitter has unfairly dominated recent research, at the expense of email and SMS text messages, which they found to be both linguistically distinct from Twitter and significantly more prevalent (in 2010). This matches earlier research arguing that email contained relatively little "neography," compared with text messages and chat (Anis, 2007).

A crucial advantage for Twitter is that it is public by default, while SMS and email are private. This makes Twitter data less problematic from a privacy standpoint,[1] far easier to obtain, and more amenable to target applications such as large-scale mining of events (Sakaki et al., 2010; Benson et al., 2011) and opinions (Sauper et al., 2011). Similar argument could be made on behalf of other public social media, such as blog comments (Ali-Hasan and Adamic, 2007), forums, and chatrooms (Paolillo, 2001). The main advantage of Twitter over these media is convenience in gathering large datasets through a single streaming interface. More comparative evaluation is needed to determine linguistic similarities and differences between Twitter and these other media; Section 4 presents an evaluation of the lexical similarity between Twitter and political blogs.

## 2 A tour of bad language

While many NLP researchers and engineers have wrestled with the difficulties imposed by bad language, there has been relatively little consideration of *why* language in social media is so different from our other corpora. A survey of laypeople found that more than half of the respondents agreed with the following partial explanations for non-standard spelling on the internet: "people are unsure of the correct spellings," "it's faster," "it's become the norm," and "people want to represent their own dialects and/or accents" (Jones, 2010). Let us now consider the evidence for these and other potential explanations.

### 2.1 Illiteracy

Some commentators have fixated on the proposal that the authors of non-standard language in social media are simply unaware or incapable of using more standard language (Thurlow, 2006). But empirical research suggests that many users of bad language are capable of using more traditional forms. Drouin and Davis (2009) find no significant differences in the literacy scores of individuals who do or do not use non-standard vocabulary in text messages. Tagliamonte and Denis (2008) review traces of instant messaging conversations among students, arguing that they "pick and choose ... from the entire stylistic repertoire of the language" in a way that would be impossible without skilled command of both formal and informal registers. While news text is usually more carefully composed and edited than much of the language in social media, there is little evidence that bad language results from an inability to speak anything else.

### 2.2 Length limits

In the case of Twitter, the limit of 140 characters for each message is frequently cited as an explanation for bad language (Finin et al., 2010). Does Twitter's character limit cause users to prefer shorter words, such as *u* instead of *you*? If so, one might expect shortening to be used most frequently in messages that are near the 140-character limit. Using a dataset of one million English-language tweets (Bamman et al., 2012), I have computed the average length of messages containing both standard words and their non-standard alternatives, focusing on the top five non-standard shortenings identified by the automatic method of Gouws et al. (2011a). The shortening *ur* can substitute for both *your* and *you're*. While *wit* and *bout* are also spellings for standard words, manual examination of one hundred randomly selected examples for each surface form revealed only one

[1] boyd and Crawford (2012) note that "public by default" data still raises important ethical considerations.

| standard | length | alternative | length |
|---|---|---|---|
| *your* | $85.1 \pm 0.4$ | *ur* | $81.9 \pm 0.6$ |
| *you're* | $90.0 \pm 0.1$ | | |
| *with* | $87.9 \pm 0.3$ | *wit* | $78.8 \pm 0.7$ |
| *going* | $82.7 \pm 0.5$ | *goin* | $72.2 \pm 1.0$ |
| *know* | $86.1 \pm 0.4$ | *kno* | $78.4 \pm 1.0$ |
| *about* | $88.9 \pm 0.4$ | *bout* | $74.5 \pm 0.7$ |

Table 1: Average length of messages containing standard forms and their shortenings

case in which the standard meaning was intended for *wit*, and none for *bout*.

The average message lengths are shown in Table 1. In all five cases, the non-standard form tends to be used in shorter messages — not in long messages near the 140 character limit. Moreover, this difference is substantially greater than the saving of one or two characters offered by shortened form. This is not consistent with the explanation that Twitter's character limit is the primary factor driving the use of shortened forms. It is still possible that Twitter's length limitations might indirectly cause word shortenings: for example, by legitimizing shortened forms or causing authors to develop a habit of preferring them. But factors other than the length limit must be recruited to explain why such conventions or habits apply only to some messages and not others.

### 2.3 Text input affordances

Text input affordances — whether standard keyboards or predictive entry on mobile devices — play a role in computer-mediated communication that is perhaps under-appreciated. Gouws et al. (2011b) investigate orthographic variation on Twitter, and find differences across devices: for example, that messages from iPhones include more contractions than messages from Blackberries, and that tweets sent from the web browser are more likely to drop vowels. While each affordance facilitates some writing styles and inhibits others, the affordances themselves are unevenly distributed across users. For example, older people may prefer standard keyboards, and wealthier people may be more likely to own iPhones. Affordances are a moving target: new devices and software are constantly becoming available, the software itself may adapt to the user's input, and the user may adapt to the software and device.

### 2.4 Pragmatics

**Emoticons** are frequently thought of as introducing an expressive, non-verbal component into written language, mirroring the role played by facial expressions in speech (Walther and D'Addario, 2001), but they can also be seen as playing a pragmatic function: marking an utterance as facetious, or demonstrating a non-confrontational, less invested stance (Dresner and Herring, 2010). In many cases, **phrasal abbreviations** like *lol* (*laugh out loud*), *lmao* (*laughing my ass off*), *smh* (*shake my head*), and *ikr* (*i know, right?*) play a similar role: *yea she dnt like me lol*; *lmao I'm playin son*. A key difference from emoticons is that abbreviations can act as constituents, as in *smh at your ignorance*. Another form of non-standard language is **expressive lengthening** (e.g., *coooollllll*), found by Brody and Diakopoulos (2011) to indicate subjectivity and sentiment. In running dialogues — such as in online multiplayer games — the symbols * and ^ can play an explicit pragmatic function (Collister, 2011; Collister, 2012).

### 2.5 Social variables

A series of papers has documented the interactions between social media text and social variables such as age (Burger and Henderson, 2006; Argamon et al., 2007; Rosenthal and McKeown, 2011), gender (Burger et al., 2011; Rao et al., 2010), race (Eisenstein et al., 2011), and location (Eisenstein et al., 2010; Wing and Baldridge, 2011). From this literature, it is clear that many of the features that characterize bad language have strong associations with specific social variables. In some cases, these associations mirror linguistic variables known from speech — such as geographically-associated lexical items like *hella*, or transcriptions of phonological variables like "g-dropping" (Eisenstein et al., 2010). But in other cases, apparently new lexical items, such as the abbreviations *ctfu*, *lls*, and *af*, acquire surprisingly strong associations with geographical areas and demographic groups (Eisenstein et al., 2011).

A robust finding from the sociolinguistics literature is that non-standard forms that mark social vari-

ables, such as regional dialects, are often inhibited in formal registers (Labov, 1972). For example, while the Pittsburgh spoken dialect sometimes features the address term *yinz* (Johnstone et al., 2006), one would not expect to find many examples in financial reports. Other investigators have found that much of the content in Twitter concerns social events and self presentation (Ramage et al., 2010), which may encourage the use of less formal registers in which socially-marketed language is uninhibited.

The use of non-standard language is often seen as a form of *identity work*, signaling authenticity, solidarity, or resistance to norms imposed from above (Bucholtz and Hall, 2005). In spoken language, many of the linguistic variables that perform identity work are phonological — for example, Eckert (2000) showed how the northern cities vowel shift was used by a subset of suburban teenagers to index affiliation with Detroit. The emergence of new linguistic variables in social media suggests that this identity work is as necessary in social media as it is in spoken language. Some of these new variables are transcriptions of existing spoken language variables: like *finna*, which transcribes *fixing to*. Others — abbreviations like *ctfu* and emoticons — seem to be linguistic inventions created to meet the needs of social communication in a new medium. In an early study of variation in social media, Paolillo (1999) notes that code-switching between English and Hindi also performs this type of identity work.

Finally, it is an uncomfortable fact that the text in many of our most frequently-used corpora was written and edited predominantly by working-age white men. The Penn Treebank is composed of professionally-written news text from 1989, when minorities comprised 7.5% of the print journalism workforce; the proportion of women in the journalism workforce was first recorded in 1999, when it was 37% (American Society of Newspaper Editors, 1999). In contrast, Twitter users in the USA contain an equal proportion of men and women, and a higher proportion of young adults and minorities than in the population as a whole (Smith and Brewer, 2012). Such demographic differences are very likely to lead to differences in language (Green, 2002; Labov, 2001; Eckert and McConnell-Ginet, 2003).

Overall, the reasons for language diversity in social media are manifold, though some of the most frequently cited explanations (illiteracy and length restrictions) do not hold up to scrutiny. The increasing prevalence of emoticons, phrasal abbreviations (*lol*, *ctfu*), and expressive lengthening may reflect the increasing use of written language for ephemeral social interaction, with the concomitant need for multiple channels through which to express multiple types of meaning. The fact many such neologisms are closely circumscribed in geography and demographics may reflect diffusion through social networks that are assortative on exactly these dimensions (Backstrom et al., 2010; Thelwall, 2009). But an additional consideration is that non-standard language is deliberately deployed in the performance of identity work and stancetaking. This seems a particularly salient explanation for the use of lexical variables that originate in spoken language (*jawn*, *hella*), and for the orthographic transcription of phonological variation (Eisenstein, 2013). Determining the role and relative importance of social network diffusion and identity work as factors in the diversification of social media language is an exciting direction for future research.

## 3 What can we do about it?

Having surveyed the landscape of bad language and its possible causes, let us now turn to the responses offered by the language technology research community.

### 3.1 Normalization

One approach to dealing with bad language is to turn it good: "normalizing" social media or SMS messages to better conform to the sort of language that our technology expects. Approaches to normalization include the noisy-channel model (Cook and Stevenson, 2009), string and distributional similarity (Han and Baldwin, 2011; Han et al., 2012), sequence labeling (Choudhury et al., 2007; Liu et al., 2011a), and machine translation (Aw et al., 2006). As this task has been the focus of substantial attention in recent years, labeled datasets have become available and accuracies have climbed.

That said, it is surprisingly difficult to find a precise definition of the normalization task. Writing before social media was a significant focus for NLP, Sproat et al. (2001) proposed to replace non-

standard words with "the contextually appropriate word or sequence of words." In some cases, this seems clear enough: we can rewrite *dats why pluto is pluto* with *that's why...* But it is not difficult to find cases that are less clear, putting would-be normalizers in a difficult position. The labeled dataset of Han and Baldwin (2011) addresses a more tractable subset of the normalization problem, annotating only token-to-token normalizations. Thus, *imma* — a transcription of *I'm gonna*, which in turn transcribes *I'm going to* — is not normalized in this dataset. Abbreviations like *LOL* and *WTF* are also not normalized, even when they are used to abbreviate syntactic constituents, as in *wtf is the matter with you?* Nor are words like *hella* and *jawn* normalized, since they have no obvious one-word transcription in standard English. These decisions no doubt help to solidify the reliability of the annotations, but they provide an overly optimistic impression of the ability of string edit distance and related similarity-based techniques to normalize bad language. The resulting gold standard annotations seem little more amenable to automated parsing and information extraction than the original text.

But if we critique normalization for not going far enough, we must also ask whether it goes too far. The logic of normalization presupposes that the "norm" can be identified unambiguously, and that there is a direct mapping from non-standard words to the elements in this normal set. On closer examination, the norm reveals itself to be slippery. Whose norm are we targeting? Should we normalize *flvr* to *flavor* or *flavour*? Where does the normal end and the abnormal begin? For example, Han and Baldwin normalize *ain* to *ain't*, but not all the way to *isn't*. While *ain't* is certainly well-known to speakers of Standard American English, it does not appear in the Penn Treebank and probably could not be used in the Wall Street Journal, except in quotation.

Normalization is often impossible without changing the meaning of the text. Should we normalize the final word of *ya ur website suxx bro* to *brother*? At the very least, this adds semantic ambiguity where there was none before (is she talking to her biological brother? or possibly to a monk?). Language variation does not arise from passing standard text through a noisy channel; it often serves a pragmatic and/or stancetaking (Du Bois, 2007) function. Eliminating variation would strip those additional layers of meaning from whatever propositional content might survive the normalization process. Sarah Silverman's *ya ur website suxx bro* can only be understood as a critique from a caricatured persona — the type of person who ends sentences with *bro*. Similarly, we can assume that Shaquille O'Neil is capable of writing *that's why Pluto is Pluto*, but that to do so would convey an undesirably didactic and authoritative stance towards the audience and topic.

This is not to deny that there is great potential value in research aimed at understanding orthographic variation through a combination of local context, string similarity, and related finite-state machinery. Given the productivity of orthographic substitutions in social media text, it is clear that language technology must be made more robust. Normalization may point the way towards such robustness, even if we do not build an explicit normalization component directly into the language processing pipeline. Another potential benefit of this research is to better understand the underlying orthographic processes that lead to the diversity of language in social media, how these processes diffuse over social networks, and how they impact comprehensibility for both the target and non-target audiences.

### 3.2 Domain adaptation

Rather than adapting text to fit our tools, we may instead adapt our tools to fit the text. A series of papers has followed the mold of "NLP for Twitter," including part-of-speech tagging (Gimpel et al., 2011; Owoputi et al., 2013), named entity recognition (Finin et al., 2010; Ritter et al., 2011; Liu et al., 2011b), parsing (Foster et al., 2011), dialogue modeling (Ritter et al., 2010) and summarization (Sharifi et al., 2010). These papers adapt various parts of the natural language processing pipeline for social media text, and make use of a range of techniques:

- **preprocessing** to normalize expressive lengthening, and eliminate or group all hashtags, usernames, and URLs (Gimpel et al., 2011; Foster et al., 2011)
- **new labeled data**, enabling the application of semi-supervised learning (Finin et al., 2010; Gimpel et al., 2011; Ritter et al., 2011)

- **new annotation schemes** specifically customized for social media text (Gimpel et al., 2011)
- **self-training** on unlabeled social media text (Foster et al., 2011)
- **distributional features** to address the sparsity of bag-of-words features (Gimpel et al., 2011; Owoputi et al., 2013; Ritter et al., 2011)
- **joint normalization**, incorporated directly into downstream application (Liu et al., 2012)
- **distant supervision**, using named entity ontologies and topic models (Ritter et al., 2011)

Only a few of these techniques (normalization and new annotation systems) are specific to social media; the rest can found in other domain adaptation settings. Is domain adaptation appropriate for social media? Darling et al. (2012) argue that social media is not a coherent domain at all, and that a POS tagger for Twitter will not necessarily generalize to other social media. One can go further: Twitter itself is not a unified genre, it is composed of many different styles and registers, with widely varying expectations for the degree of standardness and dimensions of variation (Androutsopoulos, 2011). I am the co-author on a paper entitled "Part-of-speech tagging for Twitter," but if we take this title literally, it is impossible on a trivial level: Twitter contains text in dozens or hundreds of languages, including many for which no POS tagger exists. Even within a single language — setting aside issues of code-switching (Paolillo, 1996) — Twitter and other social media can contain registers ranging from hashtag wordplay (Naaman et al., 2011) to the official pronouncements of the British Monarchy. And even if all good language is alike, bad language can be bad in many different ways — as Androutsopoulos (2011) notes when contrasting the types of variation encountered when "visiting a gamer forum" versus "joining the Twitter profile of a rap star."

## 4 The lexical coherence of social media

The internal coherence of social media — and its relationship to other types of text — can be quantified in terms of the similarity of distributions over bigrams. While there are many techniques for comparing word distributions, I apply the relatively simple method of counting out-of-vocabulary (OOV) bigrams. The relationship between OOV rate and domain adaptation has been explored by McClosky et al. (2010), who use it as a feature to predict how well a parser will perform when applied across domains.[2]

Specifically, the datasets $A$ and $B$ are compared by counting the number of bigram tokens in $A$ that are unseen in $B$. The following corpora are compared:

- **Twitter-month**: randomly selected tweets from each month between January 2010 to October 2012 (Eisenstein et al., 2012).
- **Twitter-hour**: randomly selected tweets from each hour of the day, randomly sampled during the period from January 2010 to October 2012.
- **Twitter-#**: tweets in which the first token is a hashtag. The hashtag itself is not included in the bigram counts; see below for more details on which bigrams are included.
- **Twitter-@**: tweets in which the first token is a username. The username itself is not included in the bigram counts.
- **Penn Treebank**: sections 2-21
- **Infinite Jest**: the text of the 1996 novel by David Foster Wallace (Wallace, 2012). Consists of only 482,558 tokens.
- **Blog articles**: A randomly-sampled subset of the American political blog posts gathered by Yano et al. (2009).
- **Blog comments**: A randomly-selected subset of comments associated with the blog posts described above.

In all corpora, only fully alphabetic tokens are counted; thus, all hashtags and usernames are discarded. The Twitter text is tokenized using Tweet-

[2] A very recent study compares Twitter with other corpora, using a number of alternative metrics, such as the use of high and low frequency words, pronouns, and intensifiers (Hu et al., 2013). This is complementary to the present study, which focuses on the degree of difference in the lexical distributions of corpora gathered from various media.

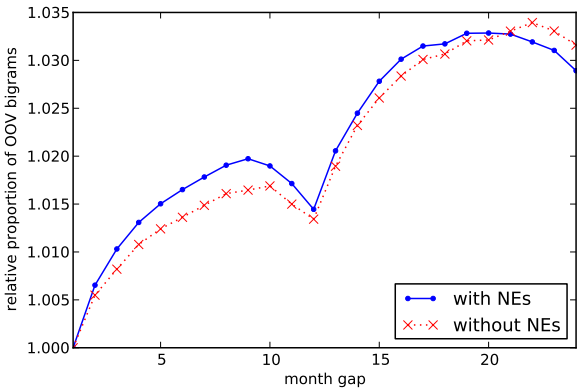


Figure 1: Lexical mismatch increases over time, as social media language evolves.

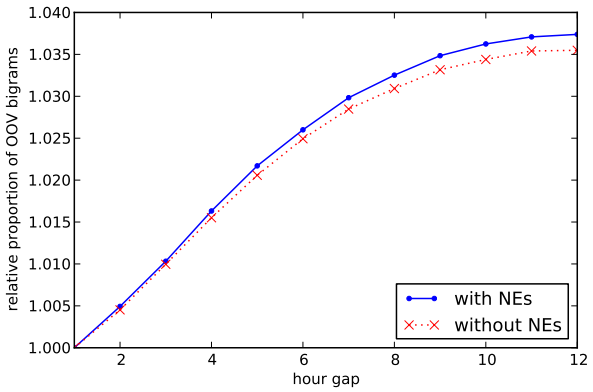


Figure 2: Different times of day have unique lexical signatures, reflecting differing topics and authors.

motif;[3] the Penn Treebank data uses the gold standard tokenization; Infinite Jest and the blog data are tokenized using NLTK (Bird et al., 2009). All tokens are downcased, and sequences of three or more consecutive identical characters are reduced to three characters (e.g., *coooool* → *coool*). All Twitter corpora are subject to the following filters: messages must be from the United States and should be written in English,[4] they may not include hyperlinks (eliminating most marketing messages), they may not be retweets, and the author must not have more than 1,000 followers or follow more than 1,000 people. These criteria serve to eliminate text from celebrities, businesses, or automated bots.

**Twitter over time** Figure 1 shows how the proportion of out-of-vocabulary bigrams increases over time. It is possible that the core features of language are constant but the set of named entities that are mentioned changes over time. To control for this, the CMU Twitter Part-of-Speech tagger (Owoputi et al., 2013) was used to identify named entity mentions, and they were replaced with a special token. The OOV rate is standardized with respect to a one-month time gap, where it is 24.4% when named entities are included, and 21.3% when they are not. These rates reach maxima at 25.2% and 22.0% respectively, with dips at 12 and 24 months indicating cyclic yearly effects. While the proportion of OOV tokens is smaller when named entities are not included, the rate of growth is similar in each case. The steadily increasingly rate of OOV bigrams suggests that we cannot annotate our way out of the bad language problem. An NLP system trained from data gathered in January 2010 will be increasingly outdated as time passes and social media language continues to evolve.

One need not wait months to see language change on Twitter: marked changes can be observed over the course of a single day (Golder and Macy, 2011). A quantitative comparison is shown in Figure 2. Here the OOV rate is standardized with respect to a one-hour gap, where it is 24.2% when named entities are included, and 21.1% when they are not. These rates rise monotonically as the time gap increases, peaking at 25.1% and 21.9% respectively. Such diurnal changes may reflect the diverse language of the different types of authors who post throughout the day.

**Types of usage** The **Twitter-#** and **Twitter-@** corpora are designed to capture the diversity of ways in which social media is used to communicate. **Twitter-#** contains tweets that begin with hashtags, and are thus more likely to be part of running jokes

[3] https://github.com/brendano/tweetmotif

[4] Approximate language detection was performed as follows. We first identify the 1000 most common words, then sort all *authors* by the proportion of these types that they used, and eliminate the bottom 10%. This filtering mechanism eliminates individuals who never write in English, but a small amount of foreign language still enters the dataset via code-switching authors. The effect of more advanced language detection methods (Bergsma et al., 2012) on these results may be considered in future work.

or trending topics (Naaman et al., 2011). **Twitter-@** contains tweets that begin with usernames — an addressing mechanism that is used to maintain dialogue threads on the site. These datasets are compared with a set of randomly selected tweets from June 2011, and with several other corpora: Penn Treebank, the novel Infinite Jest, and text and comments from political blogs. There was no attempt to remove named entities from any of these corpora, as such a comparison would merely reflect the different accuracy levels of NER in each corpus.

The results are shown in Table 2. A few observations stand out. First, the Penn Treebank is the clear outlier: a PTB dictionary has by far the most OOV tokens for all three Twitter domains and Infinite Jest, although it is a better match for the blog corpora than Infinite Jest is. Second, the social media are fairly internally coherent: the Twitter datasets better match each other than any other corpus, with a maximum OOV rate of 33.4 for **Twitter-#** against **Twitter-@**, though this is significantly higher than the OOV rate of 27.8 between two separate generic Twitter samples drawn from the same month. Finally, the OOV rate increase between Twitter and blogs — also social media — is substantial. Contrary to expectations, the **Blog-body** corpus was no closer to the **PTB** standard than **Blog-comment**.

These results suggest that the Penn Treebank corpus is so distant from social media that there are indeed substantial gains to be reaped by adapting from news text towards generic Twitter or Blog target domains. The internal differences within these social media — at least as measured by the distinctions drawn in Table 2 — are much smaller than the differences between these corpora and the PTB standard. However, in the long run, the effectiveness of this approach will be limited, as it is clear from Figure 1 that social media is a moving target. Any static system that we build today, whether by manual annotation or automated adaptation, will see its performance decay over time.

## 5 What to do next

Language is shaped by a constant negotiation between processes that encourage change and linguistic diversity, and countervailing processes that enforce existing norms. The decision of the NLP community to focus so much effort on news text is eminently justified on practical grounds, but has unintended consequences not just for technology but for language itself. By developing software that works best for standard linguistic forms, we throw the weight of language technology behind those forms, and against variants that are preferred by disempowered groups. By adopting a model of "normalization," we declare one version of language to be the norm, and all others to be outside that norm. By adopting a model of "domain adaptation," we confuse a medium with a coherent domain. Adapting language technology towards the median Tweet can improve accuracy on average, but it is certain to leave many forms of language out.

Much of the current research on the relationship between social media language and metadata has the goal of using language to predict the metadata — *revealing* who is a woman or a man, who is from Oklahoma or New Jersey, and so on. This perspective on social variables and personal identity ignores the *local categories* that are often more linguistically salient (Eckert, 2008); worse, it strips individuals of any agency in using language as a resource to create and shape their identity (Coupland, 2007), and conceals the role that language plays in creating and perpetuating categories like gender (Bucholtz and Hall, 2005). An alternative possibility is to reverse the relationship between language and metadata, using metadata to achieve a more flexible and heterogeneous domain adaptation that is sensitive to the social factors that shape variation. Such a reversal would help language technology to move beyond false dichotomies between normal and abnormal text, source and target domains, and good and bad language.

## Acknowledgments

This paper benefitted from discussions with David Bamman, Natalia Cecire, Micha Elsner, Sharon Goldwater, Scott Kiesling, Brendan O'Connor, Tyler Schnoebelen, and Yi Yang. Many thanks to Brendan O'Connor and David Bamman for providing Twitter datasets, Tae Yano for the blog comment dataset, and Byron Wallace for the *Infinite Jest* dataset. Thanks also to the anonymous reviewers for their helpful feedback.

| | Tw-June | Tw-@ | Tw-# | Blog-body | Blog-comment | Infinite-Jest | PTB |
|---|---|---|---|---|---|---|---|
| Tw-June | | 28.7 | 29.3 | 47.1 | 48.6 | 54.0 | 63.9 |
| Tw-@ | 25.9 | | 29.7 | 47.8 | 49.9 | 56.3 | 66.4 |
| Tw-# | 29.8 | 33.4 | | 49.6 | 51.0 | 54.7 | 66.2 |
| Blog-body | 41.9 | 44.1 | 43.8 | | 27.2 | 49.1 | 48.0 |
| Blog-comment | 47.4 | 49.6 | 49.2 | 30.2 | | 53.0 | 48.4 |
| Infinite-Jest | 49.4 | 51.1 | 49.9 | 48.3 | 47.4 | | 55.5 |
| PTB | 72.2 | 73.1 | 72.7 | 64.5 | 61.9 | 71.9 | |

Table 2: Percent OOV bigram tokens across corpora. Rows are the dataset providing the tokens, columns are the dataset providing the dictionary.

## References


Noor Ali-Hasan and Lada Adamic. 2007. Expressing social relationships on the blog through links and comments. In *Proceedings of ICWSM*.

American Society of Newspaper Editors. 1999. *1999 Newsroom Census: Minority Employment Inches up in Daily Newspapers*. American Society of Newspaper Editors, Reston, VA.

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.

Jacques Anis. 2007. Neography: Unconventional spelling in French SMS text messages. In Brenda Danet and Susan C. Herring, editors, *The multilingual internet: Language, culture, and communication online*, pages 87 – 115. Oxford University Press.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday*, 12(9).

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of ACL*, pages 33–40.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW*, pages 61–70.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2012. Gender in twitter: Styles, stances, and social networks. Technical Report 1210.4567, arXiv, October.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of ACL*.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pages 65–74, June.

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

danah boyd and Kate Crawford. 2012. Critical questions for big data. *Information, Communication & Society*, 15(5):662–679, May.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of EMNLP*.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

John D. Burger and John C. Henderson. 2006. An exploration of observable features related to blogger age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

John D. Burger, John C. Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10(3):157–174.

Lauren B. Collister. 2011. *-repair in online discourse. *Journal of Pragmatics*, 43(3):918–921, February.

Lauren B. Collister. 2012. The discourse deictics ^ and <-- in a world of warcraft community. *Discourse, Context & Media*, 1(1):9–19, March.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Linguistic Creativity*, pages 71–78.

Nikolas Coupland. 2007. *Style (Key Topics in Sociolinguistics)*. Cambridge University Press, July.

William M. Darling, Michael J. Paul, and Fei Song. 2012. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of EACL Workshop on Semantic Analysis in Social Media*.

Eli Dresner and Susan C. Herring. 2010. Functions of the non-verbal in cmc: Emoticons and illocutionary force. *Communication Theory*, 20(3):249–268.

Michelle Drouin and Claire Davis. 2009. R u txting? is the use of text speak hurting your literacy? *Journal of Literacy Research*, 41(1):46–67.

John W. Du Bois. 2007. The stance triangle. In Robert Engelbretson, editor, *Stancetaking in discourse*, pages 139–182. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Penelope Eckert and Sally McConnell-Ginet. 2003. *Language and Gender*. Cambridge University Press, New York.

Penelope Eckert. 2000. *Linguistic variation as social practice*. Blackwell.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

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. Technical Report 1210.5268, arXiv.

Jacob Eisenstein. 2013. Phonological factors in social media writing. In *Proceedings of the NAACL Workshop on Language Analysis in Social Media*.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Jenny R. Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370.

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*.

Scott A. Golder and Michael W. Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, September.

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

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011b. Contextual bearing on linguistic variation in social media. In *Proceedings of the ACL Workshop on Language in Social Media*.

Lisa Green. 2002. *African American English*. Cambridge University Press.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of ACL*, volume 1.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of EMNLP*.

Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. 2013. Dude, srsly?: The surprisingly formal nature of twitter's language. In *Proceedings of ICWSM*.

Barbara Johnstone, Jennifer Andrus, and Andrew E Danielson. 2006. Mobility, indexicality, and the enregisterment of pittsburghese. *Journal of English Linguistics*, 34(2):77–104.

Lucy Jones. 2010. The changing face of spelling on the internet. Technical report, The English Spelling Society.

William Labov. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

William Labov. 2001. *Principles of linguistic change. Vol.2 : Social factors.* Blackwell Publishers, Oxford.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011a. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of ACL*, pages 71–76.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011b. Recognizing named entities in tweets. In *Proceedings of ACL*.

Xiaohua Liu, Ming Zhou, Xiangyang Zhou, Zhongyang Fu, and Furu Wei. 2012. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of ACL*.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Proceedings of NAACL*, pages 28–36, June.

Robert Munro and Christopher D. Manning. 2012. Short message communications: users, topics, and in-language processing. In *Proceedings of the 2nd ACM Symposium on Computing for Development*.

Mor Naaman, Hila Becker, and Luis Gravano. 2011. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918.

Olutobi Owoputi, 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 NAACL*.

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

John C. Paolillo. 1999. The virtual speech community: Social network and language variation on irc. *Journal of Computer-Mediated Communication*, 4(4):0.

John C. Paolillo. 2001. Language variation on internet relay chat: A social network approach. *Journal of Sociolinguistics*, 5(2):180–213.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

Daniel Ramage, Sue Dumais, and D. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of ICWSM*.

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, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proceedings of NAACL*.

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

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and Post-Social media generations. In *Proceedings of ACL*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW*, pages 851–860.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2011. Content models with attitude. In *Proceedings of ACL*.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010. Summarizing microblogs automatically. In *Proceedings of NAACL*.

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

Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Sali A. Tagliamonte and Derek Denis. 2008. Linguistic ruin? lol! instant messaging and teen language. *American Speech*, 83(1):3–34, March.

Mike Thelwall. 2009. Homophily in MySpace. *J. Am. Soc. Inf. Sci.*, 60(2):219–231.

Crispin Thurlow. 2006. From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *J. Computer-Mediated Communication*, pages 667–701.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL*.

Byron Wallace. 2012. Multiple narrative disentanglement: Unraveling infinite jest. In *Proceedings of NAACL*.

Joseph B. Walther and Kyle P. D'Addario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19(3):324–347.

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

Tae Yano, William W. Cohen, and Noah A. Smith. 2009. Predicting response to political blog posts with topic models. In *Proceedings of NAACL*.