LINGUISTIC FEATURE SPREAD IN ONLINE SOCIAL NETWORKS

by

HAYDEN YGARTUA

A THESIS

Presented to the Department of Linguistics
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Arts

June 2018

An Abstract of the Thesis of

Hayden Ygartua for the degree of Bachelor of Arts
in the Department of Linguistics to be taken May 2018

Title: Linguistic Feature Spread in Online Social Networks

Approved: _______________________________________

Spike Gildea

Sociolinguists studying computer-mediated communication often study the effects of categorical variables on online language use, but studies of the effects of community membership have lagged behind. Such studies of traditional, in-person communities have found that individuals tend to speak more like those in their immediate community, and less like those with whom they are distant, but comparable studies have not been carried out for online communities. To explore whether this trend may hold true online, I have conducted a study of the use of conversational, text-based communication on the social network site Twitter. I find that a variety of linguistic features characteristic of computer-mediated communication correlate with community membership in their usage, indicating that Twitter users do tend to "speak" more similarly to those in their immediate social circles.

Acknowledgements

I would like to thank my Thesis Committee Professors Spike Gildea, Charlotte Vaughn, and Casey Shoop for helping me solidify my approach, explore ideas from different perspectives, and always explain thoroughly what I am doing. Thank you all for supporting me and answering my questions, even as I struggled to make things work.

I must also thank Professor Kendall, for teaching a class on corpus linguistics with R and lending me his copy of *Social Network Analysis: Methods and Applications*. Without his help, I would have had no idea how to begin this project.

I would also like to thank the CHC's Miriam Rigby & Miriam Jordan for providing rich thesis resources. The formatting guidelines were a real lifesaver, and the theses available through Scholars' Bank were a huge help.

Finally, I would like to give special thanks to my mother and my friends, for helping me unwind, listening to my worries, and providing me with hot meals.

Table of Contents

List of Figures

List of Tables

1. Introduction

Language as a communicative system is inherently social; communication, by its nature, requires at least two parties. Language is thus influenced by a variety of social factors, such as class, race, and gender, as well as the positions of the speaker and listener relative to social groups and hierarchies.

Historically, the study of the effect of social factors on language, or sociolinguistics, has focused on the spoken word; speech is the primary form of communication, predating the written word by millennia. Written word has generally been more prestigious and less spontaneous than speech, and hasn't always been accessible across class lines. However, with the advent of the internet, the role of written language is beginning to change. Colloquial language on the internet shares many properties with the spoken word: it is largely unedited, sometimes synchronous, and rich with sociolinguistic variation.

Writing is the primary form of communication on the internet, requiring significantly less bandwidth than audio content. Many relationships and communities have been formed solely on the basis of written communication. A study of why people use the social network site Twitter found that one eighth of posts on the site were conversational messages between individuals, indicating that communication and interpersonal connections are a key function of Twitter messages (Tamburrini et al., 2015).

However, the limitations of text-based communication have forced internet users to develop a form of conversational writing that can convey prosody[1], gesture, pronunciation, and other hallmarks of spoken word with vital social functions. The internet "has given language new stylistic varieties, in particular increasing a language's expressive range at the informal end of the spectrum" for this exact purpose (Crystal, 2005, p. 2). Many of these features are orthographic in nature, or even what one might even consider paralinguistic, such as the innovative use of punctuation. These features are rife with communicative purpose, whether to specifically emulate speech, to increase efficiency, or to convey particular meanings through signals arbitrarily decided upon by the online community at large. For instance, it is often difficult to convey sarcasm through text, but at least some pockets of the internet have agreed to signal sarcasm by surrounding the phrase in question with tildes and asterisks. This choice is ultimately arbitrary, as are most relationships between form and meaning; what is important is that the members of a given community agree on that form-meaning relationship.

The kinds of developments in language use currently occurring online are perfectly understandable; anyone with a passing knowledge of sociolinguistics recognizes that language change and variation are both natural and inevitable. However, to the casual observer or pundit, language change may seem like an aberration: "One major narrative thread in public discourse about emerging technologies involves concerns about the way language is affected… any perceived threats to conventional or standard language practices are invariably met with the same anxiety people have about

[1] Patterns of stress and intonation.

all language change" (Thurlow, 2006, p. 668) "Netspeak" has been accused of "dumbing down the English language," "signal[ing] the slow death of language," and heralding the collapse of society (Thurlow, 2006, p. 677). These overblown reactions ignore the fact that language use on the internet is functionally identical to any other kind of language use; the aim of all linguistic systems is the conveyance of meaning.

One point of comparison between oral and text-based speech[2] that has not yet been fully explored, however, is the relationship between community membership and language usage. With regard to oral speech, it is generally agreed that "People in regular contact with one another tend to share more linguistic features, and tend to borrow more features of each others' language varieties, even in situations where those varieties are different languages. Likewise, people who have less contact with one another tend to share fewer linguistic features with one another" (Paolillo, 1999). My aim in the present study is to investigate whether this claim holds true in the case of computer-mediated, text-based communication, specifically with regards to linguistic features representative of the conversational writing used online. In particular, I aim to explore whether Twitter users tend to conform to linguistic norms of the immediate social circles with regard to their use of "netspeak."

1.1. Background

1.1.1. Sociolinguistics

Sociolinguistics is the descriptive study of language in relation to social factors. Sociolinguistic theory "is mainly concerned with integrated models to account for the

[2] From here on, I refer to both types of speech as "speech," and specify the variety as needed.

links between linguistic variation, linguistic change and social structures" (Marshall, 2004, p. 15). Traditionally, sociolinguistic research focuses on the qualities of spoken language, such as phonology and syntax, or how words are said and arranged respectively. Some oft-discussed variables include the Northern Cities Vowel Shift, copula deletion, was/were leveling, use of hedge words, and so on. Phonological change, in particular, "is of primary importance to variationist sociolinguists," especially since phonological features are frequent enough that "the researcher can be confident of obtaining a sufficient number of instances from all speakers to be able to conduct quantitative analysis" (Mallinson et al., 2013, p. 11). In studies of spoken word, frequency is key, as there is usually a limited amount of data, and combing through audio for relevant features is costly and time-consuming.

Variables are generally studied with regard to categorical social variables, "usually categories with only two values (e.g., male/female, Catholic/Protestant, native/immigrant, local/cosmopolitan)" (Murray, 1993, p. 162). However, these categories alone do not paint a full picture of social effects on language use. Despite William Labov's (1972) assertion that "the primary influence and major control on linguistic behavior is exercised by hang-out groups," sociolinguistic research that focuses on interactional communities remains rarer than that which focuses on categorical social variables (p. 276). Nevertheless, there exists a strain of research built on the principles of social network analysis, "first introduced by Radcliffe-Brown in 1940, and elaborated by Barnes in 1954" (Marshall, 2004, p. 18).

Social network analysis is the study of "the linkages among social entities and the implications of these linkages" (Wasserman, 1994, p. 17). In other words, it is the study of individuals or groups' interactions with each other, and how those interactions shape behavior. The smallest unit in a social network is the "actor," who is linked by relational ties to other actors. Ties may be unidirectional, such as one actor "following" another on social media, or reciprocal, such as a relationship between friends. Ties may also be "weak" or "strong," based on the frequency, duration, and multiplicity[3] of their connection. Based on the ties between actors, each may be placed in a subgroup or cluster comprising a set of actors with associated ties. A social network is then "a finite set or sets of actors and the relation or relations defined on them" (Wasserman, 1994, p. 20). Networks may be classified by their interactional and structural properties, such as size (the number of actors) and density (the number of actual links between actors divided by the number of possible links).

Social network analysis, as applied to linguistics, has historically been concerned with small communities based on geographical proximity, or communities of practice. The data for these studies is usually collected by interviewing or observing a relatively small (n<100) selection of actors in the community, and relations are gleaned by asking each actor who they interact with and how often. An early example of network-based sociolinguistic research is Labov's 1961 study of inhabitants of Martha's Vineyard, in which he proposed "that members [of the 'in-group'] use linguistic variability to indicate their affiliation with the group" (Marshall, 2004, p. 23). The Milroys (1985), who have carried out multiple network studies, conclude that "the closer the individual's ties to a local community network, the more likely he is to

[3] The number of distinct ties between two individuals. For instance, a pair may consist of a store-owner and their customer, but beyond the owner-customer relation, they may also communicate in other contexts, such as peers in a class or members of a book club.

approximate to vernacular norm," and that "closeknit network[s can] maintain linguistic norms of a non-standard kind" (p. 359). In other words, an individual who is deeply entrenched in a community, made up of family, friends, coworkers, and so forth, will adopt the speech norms of that community. It follows that "People in regular contact with one another tend to share more linguistic features, and tend to borrow more features of each others' language varieties, even in situations where those varieties are different languages. Likewise, people who have less contact with one another tend to share fewer linguistic features with one another."

That said, "Neither stratificational analysis nor network analysis alone is capable of answering all questions; they must be considered as two approaches of quantifying certain aspects of a complex picture" (Marshall, 2004, p. 28). Both categorization based on gender, race, class, etc. and social networks have an impact on language variation and change. Furthermore, "in different speech communities social and linguistic factors are linked not only in different ways, but to different degrees, so that the imbrication of social and linguistic structure in a given speech community is a matter for investigation and cannot be taken as given" (Romaine, 192, p. 13). As such, the effects of stratification and network properties are not fixed, and sociolinguists must continue to study their effects, individually and together, to gain a comprehensive understanding of the social factors involved in language use.

1.1.2. Internet Linguistics

The field of internet linguistics, sometimes called "CMC (computer-mediated communication)" or "netspeak," began in earnest in the 1990s; since then, it has evolved alongside the internet itself. An early proponent for its study, David Crystal

(2005) writes “The Internet has given language new stylistic varieties, in particular increasing a language’s expressive range at the informal end of the spectrum… Rather than condemning it, therefore, we should be exulting in the fact that the Internet is allowing us to once more explore the power of the written language in a creative way” (p. 2). Given the vastness and diversity of the internet, it naturally encompasses a great deal of disparate research. The field “is divided into sub-varieties that are related to different communication modes,” such as email, instant messaging, and forums (Androutsopoulos, 2006, p. 419). Though these different modes may share some linguistic features, each have their own communicative norms, and so research tends to focus on only one mode at a time. There has also been a great deal of CMC scholarship outside the field of linguistics that focuses on “the dynamics of interpersonal and group communication rather than the specifics of linguistic practice” (Thurlow, 2006, p. 669).

Like traditional sociolinguistic studies, CMC studies often focus on categorical variables such as gender, age, and race. The precise extent to which social categories influence language use online is yet unclear, as is whether the effect is the same as that on other forms of speech. Of gender, Kapidzic & Herring (2011) write that “little evidence has been found of gender differences on the grammatical or word level in CMC,” compared to other kinds of speech or text (p. 41). However, they also write that “research has repeatedly found evidence of gender differences in CMC at the discourse and stylistic levels,” such as tone and use of emoji (p. 42).

Compared to the study of oral speech, linguistic variables in CMC tend to differ. Since the medium is largely text-based, phonology is generally irrelevant. Frequently cited variables include “emoticons, acronyms (IMHO = in my humble opinion, AFK =

away from keyboard), speedwriting and abbreviations (4 = for, g = grin) and conventions for simulating prosodic features (e.g. upper case = loud voice)" (Beißwenger, 2008, p. 2).

1.1.3. Twitter

A popular venue for research is social media, or social network sites (SNSs), "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (boyd & Ellison, 2007, p. 211). All of these features make social media a prime source for data collection, and especially for social network analysis. Since SNSs "enable users to articulate and make visible their social networks," they allow researchers to quantify relationships between actors in ways that traditional network studies do not allow (boyd & Ellison, 2007, p. 211). Furthermore, whereas "Early public online communities such as Usenet and public discussion forums were structured by topics or according to topical hierarchies… social network sites are structured as personal (or 'egocentric') networks, with the individual at the center of their own community," more accurately mirroring "unmediated social structures, where 'the world is composed of networks, not groups'" (boyd & Ellison, 2007, p. 219).

In particular, the SNS Twitter has become a popular site for research. Twitter is "a microblogging platform launched in 2006 that allows users to publish 'tweets,' text messages of 140 characters or less," although this limit expanded to 280 characters in November 2017 (Jones, 2015, p. 403). One reason for the prevalence of research on Twitter is its large user base: "According to the Pew Research Center, as of 2013 just under 20% of the online adult population of the United States uses Twitter," and that number is even higher for those under 30 (Jones, 2015, p. 406). It also has "nearly identical usage among women (15% of female internet users are on Twitter) and men (14%)" and "an even distribution across income and education levels" (Bamman, 2014, p. 139). Therefore, it is a useful tool for investigating a large swath of the population. Further, the Twitter API[4] makes data collection simple and accessible.

Even more relevant for my purposes is its communicative function: "about one eighth of posts [are] conversational messages rendering Twitter as a prime resource for public access to naturally occurring communication" (Tamburrini, 2015, p. 184). As such, Twitter is an ideal environment to study the effects of social interaction and community on language.

Another benefit to Twitter (and other online communities) is that data can be collected for many individuals with relative ease. The sample sizes for traditional network studies tend to be fairly low: Labov's study of Martha's Vineyard had 70 participants; Milroy's study of Belfast had 46 participants; and Cheshire's study of Reading had 25 participants. In contrast, studies of Twitter users may survey hundreds or even thousands of individuals.

However, certain characteristics of Twitter, such as anonymity, "[raise] problems for traditional variationist methods which assume that reliable information about participant gender, age, social class, race, geographical location, etc., is available

[4] An "API," or "application programming interface," is a set of tools or methods for building software or applications, often in relation to an existing website or database. The Twitter API allows programmers to interface directly with content published on Twitter using Python, R, or other languages, without using generic web scraping tools, or contacting the Twitter to ask for permissions. However, rate limits apply, and a maximum of 3,200 tweets may be collected from each Twitter user's timeline. For details, see https://developer.twitter.com/en/docs/basics/things-every-developer-should-know.

to the researcher" (Androutsopoulos, 2006, p. 424). On the ability to conceal geographical location, Jones (2015) writes:

> of the tweets sampled, between 2.5% and 7% had location services enabled, which left between 150 and 800 mappable tweets for each word or phrase in the United States… Moreover, while it is possible for users to specify their location in their profiles, and many do, Twitter is very careful to emphasize that there is no way of knowing if such data are accurate (and in many cases, they are obviously not; "tha hood," "sesame street," "your mom's house," all show up, as well as places that no longer exist, like the notorious projects "Cabrini Green," and the obviously fake 0.0 by 0.0, which is in the ocean off the coast of Morocco) (p. 407).

As such, most studies of Twitter users limit themselves to the extremely small percentage of users for whom personal data is available. This approach could be problematic, as only including users who have their personal information accessible may bias the data in yet unforeseen ways; this remains an area in which more research must be done. Furthermore, the tweets of users whose accounts are protected or "locked" are not publicly accessible, also limiting options for researchers.

1.2. Literature Review

In the following section, I aim to discuss the strengths and failings of three studies of online social networks, and discuss in greater depth the types of linguistic features that have been discussed in the context of CMC.

1.2.1. Bamman et al. (2014)

Using a corpus of 14,000 Twitter users, Bamman et al. grouped users into clusters based on style and topical interests, and analyzed these clusters with regard to gender. Gender of users was determined using historical census information to assign genders to users' names, only selecting users whose names occurred over 1,000 times in the census data.[5] These clusters tended to be mostly male or female, and so the results were used to make generalizations about male and female language use.

Bamman et al.'s results show certain correlations between gender and feature usage. For instance, some features associated with female speakers include: alternative pronoun spellings, like "u, ur, yr," friendship terms like "bestie, bff, and bffs," abbreviations like "lol and omg," "ellipses, expressive lengthening… exclamation marks, question marks, and backchannel sounds like ah, hmmm, ugh, and grr," and hesitation words like "um and umm," "assent terms okay, yes, yess, yesss, yessss," the "abbreviated form of with… w/a, w/the w/my," and "non-standard spelling," while male-associated features include "yessir," "nah, nobody, and ain't," "swears and taboo words," and "2" for "to" (p. 143). Emoticons were also found to be associated with women, although the general consensus is mixed, as, among teenagers, Huffaker & Calvert found them to be used more by boys than girls. The tendency for many frequently noted features of CMC to be associated with women's speech may be indicative of the fact that women's speech is often stigmatized and thus more noticed, and that women are often at the forefront of linguistic change (see William Labov's 1990 "The intersection of sex and social class in the course of linguistic change").

However, Bamman et al. write:

> While most of the clusters are strongly gendered, none are 100 percent male or female. What can we say about the 1,242 men who are part of female- dominated clusters and the 1,052 women who are part of male-dominated clusters? These individuals could be dismissed as outliers or statistical noise. Because their language aligns more closely with the other gender, they are particularly challenging cases for machine

[5] This seems like a limited approach to selecting Twitter users as it restricts the possible pool of users to those who have their display name set to their actual name, only includes those with common names, and assumes that all users are male or female.

> learning. But rather than ask how we can improve our algorithms to divine the 'true' gender of these so-called outliers, we might step back and ask how their linguistic choices participate in the construction of gendered identities (p. 148).

Gender is a complex social phenomenon, so the lack of uniformly gendered clusters is unsurprising. Determining all the factors that account for why some individuals do or do not conform to gendered linguistic conventions would require a much more nuanced analysis than what may be provided in such a large study, with a relatively crude approach to evaluating author gender in the first place.

Furthermore, although this study involves clusters of users, these clusters are based on style and content, not on users' relations to each other. The individuals' actual links, such as whether they follow or communicate with each other, are unclear. Rather than generating clusters based on language use and then investigating the traits of the individuals in these clusters, my aim is to generate clusters based on social relations, and see if these clusters have a meaningful effect on language use. Bamman et al.'s study may show that social factors can be gleaned from language use, but it then seems necessary to explore the converse: whether language use can be gleaned from network factors.

1.2.2. Tamburrini et al. (2015)

A study that does investigate the linguistic role of social relations on Twitter is Tamburrini et al.'s 2015 study of word usage in replies.[6] The study investigates a network of 189,000 Twitter users partitioned into 424 communities using a modularity

[6] Replies are tweets directed at a particular Twitter user, possibly containing "@" followed by their username. Users may reply to replies, continuing the conversation, and more than two users may join in a single conversation.

maximization algorithm, which determines which set of groupings have the greatest possible density. Of these, Tamburrini et al. then studied 24 groups consisting of at least 250 English-speaking users. They found a significant difference in the use of words and apostrophes in the messages between members of the same group vs. the messages between members of different groups; they also investigated word-endings, but did not find significant results. Their study found that senders of replies tended to conform to the language use of the recipient. These are promising results, and suggest that community membership on Twitter does play a role in language use.

However, I would like to investigate users' speech overall, rather than only their speech directed at specific recipients, and take a more granular approach with regard to linguistic variables. In looking at such a large number of users, Tamburrini et al. understandably keep their variables relatively simple, focusing only on word usage, apostrophes, and word-endings. However, it strikes me that word frequencies are not the best measure to determine linguistic similarity, as they may reflect the contents of users' conversations rather than their linguistic styles.

Further, their discussion of their variables is quite vague. It is unclear what word-endings they investigated, and their approach to investigating apostrophe usage is overly broad. They write, "apostrophe frequencies were calculated" using "the frequency of apostrophes per word used," indicating they looked only at the raw number of apostrophes used in comparison to total word count (p. 85). However, the quantity of apostrophes used is not enough to adequately measure a user's style, as a low count may indicate either the lack of contractions, or the use of contractions without apostrophes. These possibilities are diametrically opposed in their formality, so it does not seem appropriate to place them in the same category. As such, I would like to build upon this research by taking a more detailed look at the variables in question.

1.2.3. Testimonies from CMC users

Outside of academia, many users of Twitter and other forms of CMC have informally discussed their use of language. I would be remiss in my study of CMC to not consider the observations made by these individuals, so I am including here a few of these observations regarding what kinds of linguistic features are characteristic of current text-based communication.

A tweet by □the wiggler□ with over 7,000 retweets and 14,000 likes posits “aaaaa & AAAAAAAAAAAA,” “any variation of SFDSGJFFJDGDGB,” and “Talking Like This for No Reason” as examples of common features of the speech of gay Twitter users, though it remains to be seen how these features are influenced by sexual orientation, or whether they apply to a broader contingent of internet users. In either case, the tweet implies that lengthening, keysmashes, and nonstandard capitalization are all salient features.

Other tweets, such as those by bec 198 and sleepy bitch nisha, frame linguistic features as qualities of the author’s personal speech, such as “random capitals for Emphasis,” “double,, comma,” “unnecessary… elipses[sic],” “ing but spelled ign,” “switching to all caps in the middle OF A WORD,” “!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!,” “lol,” and “‘you’ and ‘u’ in the same sentence,” indicating punctuation, capitalization, acronyms or abbreviations, and intentional misspellings are salient features. Although framed as personal quirks, the large number of likes and retweets indicate that these features are not unique to them.

Users of the SNS Tumblr, of which there is much overlap with Twitter's userbase, have also written about their own use of language. A post originally by user steveogers mentions “typing in a cresCENDO TO EXPRESS EXCITEMENT” and “…………..unnecessarily……. long……….. ellipsis.” Other users add more features to the post, such as “unnecessary!!!! punctuation marks???????,” “unneeded™ trademark symbols™,” “using commas,,,,,, for ellipsis,” “B I G S P A C E S F O R E M P H A S I S,” and “wHaTeVvEr ThiS isS.” Another post by crtter, tertiusiii, and anexperimentallife mentions “[r]eplacing ‘ck’ with ‘cc,’” “[s]witching the ‘n’ and ‘g’ in a word ending with ‘ing,’” and “[g]oing from lower case to caps in the middle of a word,” among others.

Tumblr users have also made posts discussing the communicative functions and structures of keysmashes and punctuation.[7] In reply to user a6’s query, “u kno when u keysmash but the jumble of letters dont convery the right Feeling so u gotta backspace and re-keysmash to turn ur HKELSXPXA to a JKFSDKAS,” stanzicapparatireplayers writes that “a deliberate keysmash and an accidental one need to be distinguishable,” and “a deliberate keysmash will nearly always use keys only in the home row, and usually in a particular order that isn’t likely to have happened purely accidentally.” In light of the lack of current scholarship on keysmashes, these assertions have not yet been investigated in an academic context, but they are compelling pieces of anecdotal evidence that features like keysmashes display formal constraints or tendencies.

[7] For a good source of similar analyses of features of internet speech, visit https://allthingslinguistic.com/tagged/language+on+the+interwebz/.

User averagefairy describes their confusion in interpreting the tone of texts from "old people," who do not know the norms regarding punctuation; "Yay for you…." Comes across as "passive aggressive." In response, user feynites details her bewilderment at her mother's punctuation use:

> So whereas I might sent a response that looked something like:
> "Yay! That sounds great - where are we meeting?"
> My mother, whilst meaning the exact same thing, would go:
> 'Yay. That sounds great… where are we meeting?"
> And when I look at both of those texts, mine reads like 'happy/approval' to my eye, whereas my mother's looks *flat.* Positive phrasing delivered in a completely flat tone of voice is almost always sarcastic when spoken aloud, so written down, it looks sarcastic or passive-aggressive.

In other words, she and other digital natives understand they must work within the formal limitations of text-based communication to adequately convey their intonation and emotions. The textual communication gap between feynites and her mother exemplifies the matter in which the communicative functions of orthographic features, such as punctuation, are socially determined. Those outside the community that uses these features in such a fashion do not necessarily understand the norms of their use. It stands to reason that community membership factors into the use of orthographic features intended to represent or suggest features of oral speech.

1.2.4. Linguistic features

What follows is a list of types of orthographic features that have been mentioned or studied in a collection of research papers. These are not the only features mentioned in these papers, but they are the ones I would consider unique to or associated with CMC, as opposed to obscenity, code-switching, and dialectal features, which are also frequently studied. I compare these features with those mentioned in the testimonials of CMC users for reference.

	Researchers						CMC users						
	Androtsopolous 2006	Bamman et al. 2014	Beißwenger et al. 2008	Lyddy et al. 2014	Paolillo 1999	Tamburrini et al. 2015	a6 2017	anexperimentallife 2017	averagefairy 2017	bec 2016	cappuccinohowell 2018	sleepy bitch nisha 2017	□the wiggler□ 2018
Acronyms			✓	✓									
Abbreviations		✓	✓	✓	✓								
Capitalization				✓				✓		✓	✓	✓	✓
Emoticons	✓		✓										
Keysmashes							✓		✓				✓
Nonstandard spelling	✓	✓		✓									
Punctuation		✓		✓		✓	✓		✓	✓	✓	✓	
Simulating speech[8]	✓	✓	✓	✓									✓
Suffixes[9]				✓	✓	✓				✓	✓	✓	

Table 1. Frequently-discussed features

As is evident above, the features singled out by CMC users are not always featured prominently in sociolinguistic literature. In particular, keysmashes have received little to no scholarly attention, and capitalization also remains understudied. Admittedly, these discrepancies are in part due to the relative recency of the social

[8] This category contains expressive lengthening, hesitation, and backchanneling though some of these papers only refer to it in the broadest of terms.

[9] This refers to a subcategory of nonstandard spellings of suffixes, such as “-z” for the plural suffix, or “-in” for “-ing.”

media posts cited compared to the scholarly articles, which are all from 2015 or earlier. Like the internet itself, the linguistic norms of CMC have evolved quickly, so it could well be that some of these features have only appeared in the past two years. However, I did make an attempt to find newer literature, and did not see significantly different results. I was unable to find a single paper on keysmashes.

1.3. Hypothesis

I hypothesize that CMC speech, like oral speech, should evolve to express the kinds of “paralinguistic” information that can be conveyed with gesture or intonation. Further, as these new codes develop, they will spread initially in close networks of communicators, and as such will reflect, or even create, the norms of these networks. In the context of the present study, I theorize that Twitter users adopt similar linguistic features as their peers, and that social network analysis may shed light on linguistic variability on Twitter in ways that other sociolinguistic methods cannot. I also hypothesize that densely linked communities of Twitter users will display less variance in feature-usage than more loosely-linked, or unrelated groups of individuals, since it potentially follows that a greater degree of conformity leads to a lower degree of variation. In other words, I aim to test the propositions that “People in regular contact with one another tend to share more linguistic features,” and “people who have less contact with one another tend to share fewer linguistic features with one another” with regards to language on Twitter.

I theorize that Twitter users are more likely to produce tweets with a higher proportion of nonstandard stylistic features if they belong to a community where others also make use of nonstandard features. “Nonstandard stylistic features” encompasses an intentionally broad variety of conventions characteristic of “netspeak,” such as acronyms, abbreviations, and nonstandard spelling. This approach may then shed light on how certain linguistic features come to be adopted by the larger online community. I accomplish this using the framework of social network analysis, and the computational tools of text mining.

Twitter is a particularly fascinating vehicle for social network analysis, given its high degree of direct communication between users and wide userbase. Furthermore, communities on Twitter are not necessarily bounded by the constraints of location, class, etc., and relations between Twitter users are chosen, not based on proximity or necessity. Moreover, the process of measuring or quantifying community membership on Twitter is simpler than in geographical communities, given the accessibility of data concerning user relations. Thus, Twitter offers a comparatively pure environment for social network analysis.

2. Methodology

In this section, I aim to delineate the steps I have taken to collect and analyze my data. This includes the initial collection of user information, the partitioning of these users into groups, and the collection, processing, and analysis of the text of the tweets themselves. I outline both the theoretical reasons for my approach, as well as the specific computational processes used.

2.1. Data

My data consists of the most recent original tweets as of the time of their collection from a set of Twitter users. The following sections detail how I selected Twitter users, and how I then collected their tweets.

2.1.1. Data collection

Data was collected using the programming language R's twitteR package[10], as well as Python's Tweepy package.[11]

Twitter users were collected was done using twitteR's getFriends[12] and getFollowers methods. To amass a list of individuals, I used a snowball-sample where, for each user sampled, beginning with a single user, all of their Friends and followers

[10] R is an open-source programming language designed for statistical computing, which also contains the necessary tools for text-mining, such as the ability to count instances of specific words, phrases, or patterns. The twitteR package allows the user to interface with the Twitter API, and collect tweets and user information.

[11] Python is a general-purpose programming language. The Tweepy package is comparable in function to the twittR package.

[12] twitteR's "getFriends" method obtains a list of all users a user is following. The name is misleading, as they may not necessarily be "friends." Taking a page from boyd and Ellison (2007), I will capitalize this usage of "Friends" to distinguish it "from the colloquial term 'friends'" (p. 225).

are added to a list of users. Then, I obtain Friends and followers for all members of that list, and so on until I had gathered 92,075 users.

To obtain a diverse sample, I opted to carry out multiple stages of snowball-sampling so as to ensure that users in the final sample would be from sufficiently distant communities, so that there would be no overlap between subgroups. For the second stage, I took three random samples of size 1000 from this list of 92,075, and gathered more detailed user information for these 3000 users. I then gathered a list of their mutuals,[13] capping my list of users at 10,000. For these 10,000 users, I collected detailed information, and filtered out accounts that did not meet my specifications. Their accounts must:

1. Be unprotected, meaning that they can be accessed by anyone
2. Be unverified, indicating that they are not a public figure, so as to avoid accounts that are moderated by a PR team or are otherwise compromised with regards to naturalness
3. Have more than 1000 statuses, to ensure they use Twitter regularly
4. Have at least one Friend and at least one follower, to decrease the chance of collecting bot accounts
5. Have under 1000 Friends and under 1000 followers, to ensure that no one user has too much influence
6. Have their locale set as English[14]
7. Not have the default Twitter icon, again to ensure regular Twitter usage and lessen the prevalence of bots

[13] "Mutual followers," or "mutuals," are users who follow each other. By this, I mean that for each of the 3000 users, I gathered the intersection of their Friends and followers.

[14] Twitter provides an option to set the language of one's Twitter interface; however, not all languages are available, and users do not always have their locale set to the language they speak most often.

This filter resulted in a list of 2703 users. For these users, I generated a table with 2703 rows (representing followers) and 2703 columns (representing Friends) to visualize the relations between users. I removed empty rows and columns, such that only users with at least one connection in the set remained.

For each of these 1844 users, I attempted to gather 1000 original tweets[15] for initial testing. Of these, I successfully collected tweets from 1740 users. The users for whom I could not collect data may have deleted or locked their accounts, or else deleted all but under 1000 of their tweets. Based on the aforementioned requirements, all remaining users should have had 1000 or more tweets, which is why I tried to gather 1000 for each user. However, the "number of tweets" field always includes the number of tweets a user has retweeted from other users, meaning that not all users with over 1000 tweets have over 1000 original tweets. I excluded users for whom I obtained no original tweets, but even then, only 467 users had yielded more than 100 tweets. Furthermore, twitteR was only able to collect the first 140 characters of 280-character tweets, as the package has not been updated since the character limit expanded. For this reason, I used this dataset to generate user relations and filter out accounts that do not primarily tweet in English, but not for extracting data from the tweets themselves.

To filter out non-English accounts, I removed all accounts whose most frequent words did not include at least one of the following: "is," "and," "the," "you," "be," "are," "who," "by," "him," "this," "my," and "from." Many of the remaining non-

[15] These are tweets written by the user posting them, as opposed to retweets, which originate with another user. I am also exclusively including tweets, not direct messages, which are not publicly available.

English tweets were in Indonesian[16], so I removed all that included at least one of the following among their most frequent words: “aku,” “gua,” “di,” “dia,” “yg,” “tu,” “tak,” “nak,” “yang,” and “aja.”

2.1.2. Network generation

At this point, I restricted the table I had made previously to include only the remaining 908 users and converted that table to one in which each row represents a possible connection between two users. I then used functions from the network analysis package igraph to generate communities from that data frame, using one of two methods: walktrap and edge betweenness. These methods can be conceptualized by visualizing the network as a collection of nodes (representing actors) connected by lines, representing connections. “Edge betweenness” is the number of shortest paths between nodes that run alone a given line. Lines with a high degree of edge betweenness generally connect separate communities to each other, so the edge betweenness method eliminates these edges until only self-contained clusters remain. In contrast, the walktrap method does the opposite, detecting the communities themselves rather than the edges that delineate them. The algorithm can be visualized as randomly “walking” along paths, and getting “trapped” in highly interconnected groups of nodes, which then form clusters.

[16] This prevalence is likely for a few reasons: Indonesia is the fourth most populous country in the world; many Indonesians are bilingual, and a significant portion of their tweets are in English; and, despite the prevalence of the language, Twitter does not have Indonesian as an option for its interface, meaning many Indonesians use the site in English. Furthermore, while there were a couple of users I did not exclude whose Tweets were a mix of English and Chinese or Japanese, I opted to exclude Indonesian because it uses the Roman alphabet, and as such is more likely to interfere with variable counts.

For each method, I generated models with 40, 50, 60, 65, and 99 communities. All of the methods ended up with relatively similar results, so I opted for the approach with the greatest number of groups. Of the 99 communities, only the first 31 had more than 3 members, so I restricted further analyses to only these 31 communities. I also checked the number of replies between users in groups, and the in-group vs. out-group followers of each user. Each group was fairly self-contained in these regards, though four of the larger groups were a little looser.

Having sufficiently narrowed down my pool of users, I used Python's Tweepy package to gather tweets from the 815 members of groups 1 through 31. This was done in Python to circumvent the aforementioned limitations of twitteR. Tweets were collected between February 10th and February 14th, 2018.

Once tweets were recovered for each user, I converted the data to a more machine-readable format by fixing the text-formatting and file encoding, and by replacing line break characters with "[line break]." At this point, I found that 11 of the groups still contained a significant number of Indonesian tweets, so these groups were removed from the dataset. Two groups also contained some Chinese and Japanese words, but I did not remove these groups because, as these languages do not use the Roman alphabet, it is easy to isolate only the English text for analysis. I also removed one user who tweeted regularly in French.

This set (Set 1) contains 19 groups of users, with the smallest group consisting of 4 users and the largest of 129. The average density (actual follow connections over total possible follow connections) for each group is also quite low, with the density for some of the larger groups being less than 0.01, indicating that fewer than 1% of the possible connections between users are actually present. To better control for these effects of group size and density, I manually generated a second set of groups (Set 2), consisting of the smaller (n<20) groups from Set 1, as well as smaller, denser subgroups from the larger groups. The resulting set consists of 26 groups with between 3 and 17 users, with a much greater average density.

Set	Users	Average users per group	Tweets	Average tweets per group	Words	Average words per group	Average Density
1	686	36.105 (SD 33.396)	1367240	71,960 (SD 63,120.352)	13,970,410	735,284.737 (SD 611704.485)	0.285 (SD 0.207)
2	208	8 (SD 3.868)	429169	16,506.5 (SD 8,978.610)	4,282,542	164,713.154 (SD 100642.606)	0.601 (SD 0.228)

Table 2. Network information

Although Set 2 is much smaller, I consider it a far superior dataset for analyzing the effects of community. Since there is much less variance in terms of group size and density, I can better compare the groups to each other, without worrying about the possible effects of size and density. Further, I have greater confidence that the groups in Set 2 do, in fact, represent coherent communities of users where many of the members follow each other, and which are relatively isolated from the other groups. As such, I used Set 2 for data collection and analysis.

2.2. Linguistic variables

Having generated a network of users and collected their tweets, I then investigated the language used in the text of the tweets themselves. The following sections describe the linguistic features that I looked for, and the methods by which I searched for them in the text.

2.2.1. Identifying variables

I mined the final set of tweets for a diverse selection of variables. My variable list is hardly comprehensive, but rather comprises a selection of features drawn from previous research, testimony from Twitter users, and personal experience as a member of the speech community, as well as careful investigation of the data itself.

I am restricting my variables to those that occur within the text of the tweets themselves, and as such I will not be investigating the use of images, whether animated gifs, reaction images, memes, etc. I also do not include syntactic variation, since my data will not be parsed for parts of speech. To even parse informal CMC data to begin with would be problematic, as “Tools developed for the automatic annotation of linguistic data (sentencizers, [part of speech] taggers, lemmatizers, chunk parsers) cannot be used for processing CMC data without being adapted” to accommodate “netspeak” (Beißwenger, 2008, p. 12).

Since I cannot manually parse 429,169 tweets, I have restricted my analysis to features I can easily parse using R’s text mining algorithms. The variables I have chosen may be sorted into several broad categories, each of which also contains subcategories. Categories include:

1. Stylistic
    a. Acronyms (such as “lol,” “tfw,” etc.)
    b. Abbreviations (“rly,” “ppl,” etc.)
    c. Apostrophe deletion (“dont” for “don’t,” etc.)

2. Emotive, not necessarily representational

   a. Nonstandard or excess use of punctuation marks (periods, commas, exclamation points, ellipses, question marks, asterisks, etc.)

   b. Use of upper and lowercase letters (the use of only capitals or lowercase, capitalization of seemingly random words, the change from lowercase to capitals mid-word, etc.)

   c. Keysmashes (sequences of random letters generated by pressing random keys, used to indicate a variety of emotions)

   d. Alternative suffixes (“ign” for “ing”, “cc” for “ck,” etc.)

   e. Alternative spellings (“cwying” for “crying,” etc.)

3. Emotive, representational

   a. Exclamations (“oh,” “yay,” etc.)

   b. Filler words (“umm,” “ermm,” etc.)

   c. Laughter (“haha,” “heehee,” etc.)

   d. Repeated letters (such as “ahhhh,” “oooo,” etc.)

These categories are necessarily vague and overlapping. Their chief purpose is to expository and organizational convenience; I do not intend for them to be definitive or comprehensive descriptive categories. My working definitions are that “stylistic” features primarily serve to save time and conform to space limitations; “emotive, non-representational” features express emotion or affect, but not in a way that is, or even can be, replicated in oral speech; and “emotive, representational” features specifically simulate the sounds of oral speech. Alternative spellings could be argued to be representational, as they often evoke a particular pronunciation, but I include them in the non-representational category as they often, in my experience, reflect pronunciations that are highly uncommon in oral speech. I treated words like “ppl” for “people” and "shoulda" for "should have" as abbreviations rather than alternate spellings because they a) contain significantly fewer letters than the unabbreviated form and, in the case of "shoulda" b) are a contraction of two words.

For abbreviations and apostrophe-deletion, I excluded ambiguous items. For instance, I did not compare instances of "he'll" with "hell," as the version without an apostrophe is a separate, relatively common word. I also decided not to include single-character abbreviations, such as "2" for "to/too," as I cannot automatically discern whether or not they are in fact abbreviations, or just representations of single letters or numbers.

A full list of variables used in this study can be found in Appendix B.

2.2.2. Variable collection

Using the stringr R package, I searched for variables using the str_count function, which counts the instances of a particular regular expression[17] in a body of text.

For the variables that occur in alternation with a more standard variant (abbreviations, apostrophe deletion, alternative spellings/suffixes), I collected counts for both the standard and nonstandard variants, and used the proportion of nonstandard variants over all (standard or nonstandard) tokens for later analysis. For the remaining variables, I collected raw counts and normalized them to number of instances per 1000

[17] Regular expressions are sequences of characters used by search algorithms to identify patterns of characters or character-types (such as punctuation, numerals, or uppercase characters). Detailed information can be found at https://regexr.com/, and relevant expressions for the present study are included in Appendix A.

words, or, in the case of punctuation and capitalization, which cannot be compared to single word tokens, per 1000 tweets.

For suffixes, to ensure that items were in fact in variation with a standard variant, I began by taking a list of all word types in the data with the standard suffix[18] ("ing," "y," "ck," or "cking"), and searching the data to see if a desired nonstandard variant appeared for each. I then took the proportion of nonstandard tokens over the total tokens, standard or nonstandard, but only counted standard word types that actually occurred in variation. For example, if "accounting" appeared but "accountign" did not, then "accounting" would be excluded from the total count when computing the proportion of "ign" suffixes.

The one variable that was not trivial to search for was keysmashes, which, by their very nature, are difficult to patternize. As such, I had to devise another method to identify keysmashes based on their common properties so that I could count their occurrence, or at least a method that would exclude actual words and misspellings. I began by generating a wordlist for the entire dataset. I removed all word types that included non-English characters, accent marks, numbers, or two of the same letter in a row, as these are not typical features of keysmashes in my experience. I then removed all items that appeared in the dictionary[19], as well as all items with 5 or fewer characters. I removed all items with common letter clusters (VCVC[20] or CVCV, "er" and "in"), and all items that contain a more frequent item of three or more characters. At

[18] "Suffix" is used loosely here to refer to any combination of letters that tends to occur at the end of a word.

[19] My dictionary was compiled from machine readable English dictionaries in .txt format, found at http://gwicks.net/dictionaries.htm and https://github.com/titoBouzout/Dictionaries

[20] "V" stands for vowel and "C" for consonant.

this point, I combed over the remaining items manually, and removed some items that were obvious misspellings and not keysmashes. My final list contains 3,877 keysmashes. Though I believe this approach to be reasonably accurate, it is not ideal. I erred on the side of caution with my algorithm, so it likely underestimates the true number and variety of keysmashes.

2.3. Data analysis

For each variable, I used the aov function from R’s built-in stats package to determine the likelihood that group membership predicts variable usage. This approach uses an “analysis of variance” or ANOVA approach to examine the dependency of a variable in response to other factors. In this case, the dependent variable is a numeric value representing the usage of a single feature by a single user, while the independent variable is categorical: the group the user is a member of. The approach then discerns the likelihood that the explanatory variable, group membership, influences the dependent variable, feature usage.

I will illustrate the process with a small sample of my data. Below are the per-1000 word frequencies of the acronym “af” (“as fuck”) for each user of groups A, M, Y, and Z. I have selected these groups such that the relation between group membership and frequency is immediately apparent. Note that members are unique to each group, and that the number of members in each group is not uniform.

	Group A	Group M	Group Y	Group Z
Frequen	0	0	0	0.112
	0.270	0.133	0	0.098

	0	0.181	0	0.122
	0	0.171	0	0
	0	--	--	0.205
	0.150	--	--	0.071
	0.159	--	--	0
	--	--	--	0.033

Table 3: Frequency of "af" per 1000 words for four groups

The analysis of variance function determines whether there are statistically significant differences between the average frequencies for each group. The result is conveyed with a P-value. A high P-value indicates that the null hypothesis (no significant difference between groups) is likely, while a low P-value indicates that the alternative hypothesis is likely. In accordance with standard practices, I took P-values under 0.05 as evidence of significance.

The approach is useful as it does not necessitate that each group have an equal number of data points, and it allows researchers to determine whether there is a relationship between a categorical variable and a numerical one, as opposed to simple correlation tests, which may only be used when both variables are numerical. Group membership cannot increase or decrease, so it cannot precisely be said to "correlate" with frequency, though it may explain or influence it.

Group density was calculated using the formula D=Na/N, "where D=density, Na=number of actual links, N=number of possible links," as described by Marshall (24). To see whether group factors like density affect the variability within groups, I computed p-values and correlation values for the relationship between the standard deviation of a feature within a group to a) the density of the group and b) the number of members in the group. Correlation values reflect the correlation of two continuous, numerical variables; unlike categorical variables, such as group membership, which associates each data point with a certain category, with no in-between values, the values here are numerical and may include fractions or decimals. The correlation value is the rate at which one variable changes relative to changes in the other variable, according to the general trend. For instance, a value of 1 means that each time the independent variable increases by one, the dependent variable does the same. A value of 0 indicates a lack of correlation.

3. Results and Discussion

3.1. Results

3.1.1. Variable-group correlations

P-values were quite low for many variables, indicating that group membership is a strong predictor for the usage of many features. However, no variable category encompassing multiple individual variables was uniform in significance. Rather, a portion of variables from each category showed significant results, while others did not.

The following table provides a breakdown of the 160 significant variables from each category. A "+" indicates that the preceding letter or syllable may repeat multiple times. For instance, "wahahaha+" includes "wahahaha," "wahahahaha," and so on, and "yaa+yyy+" includes "yaaayyyy," whereas "yayyy+" includes items with any number of y's, but only one a. Items that may be pluralized are appended with "(s)."

Category	Subcategory	Significant variables[21]
Stylistic	Acronyms	**af**, **bff(s)**, **btw**, **ftw**, **gtg**, **gdi**, **idc**, **idgaf**, **idk**, **idr**, ikr, imo, **jk**, **lmao**, **lmk**, **lol**, **omfg**, **omg**, omw, **rn**, **smh**, **stfu**, **tbh**, **tf**, **ty**, **wtf**, **wth**, **wyd**
	Abbreviations	**abt**, **bc**, **buncha**, **fav(s)**, **fave(s)**, **gonna**, **gotta**, **luv**, **msged**, **nbhd(s)**, **nvm**, **notif(s)**, **obv**, **ofc**, **ppl**, **pls**, prob, **rly**, shoulda, sth, **smth**, **thnks**, **tho**, **tmrw**, **w/e**, w/, **w/o**, **ur**, **urself**
	Apostrophe deletion	**im**, ive, **youre**, youll, **yall**, youd, **hes**, **shes**, **theyre**, **isnt**, dont, **wont**, **havent**, hasnt, **wouldnt**, shouldnt, **mustve**, **heres**, theres, **therell**, itll, **itd**, **thats**
Emotive, not representational	Punctuation	**additional periods, exclamation points, question marks, commas, semicolons, slashes, tildes, !?**

[21] Bolded items have a p-value under 0.01.

	Capitalization	**all uppercase**, **all lowercase**, **beginning with uppercase**, upper to lower, **lower to upper**, **mixed cases**[22]
	Keysmashes	**keysmashes**
	Alternative spellings	**blease**, pwease, **hewwo**, **cwying**, **fcked**, **gorl**, **smol**, **borger**, **cronch**
	Alternative suffixes	-in, -en, -ie, -cc, -kcing
Emotive, representational	Exclamations	**oh**, **ohh+**, **ugh**, **nooo+**, **nono+**, **woo+www+**, **ya**, **yaaa+**, yaya+, **yayaaa+**, **yayyy+**, **yaayyy+**, **ayyy+**, **yeaaa+**, **yeah**, **ack**, **pfff+t**, pfff+
	Filler	**uhh+**, umm+, **ah**, **ahh+**, ermm+, **huh**, **hm**
	Laughter	**haha**, **wahahaha+**, hehe, **teeheehee+**, **huhuhu+**, huehue+, **hohoho+**
	Repeated letters	**three or more repetitions (a, c, d, g, h, I, l, m, n, o, q, s, t, u, w, y)**

Table 4: Significant variables

That so many variables display a potential correlation with group membership supports my hypothesis that Twitter users' usage of innovative linguistic features is influenced by their immediate peer group.

3.1.2. Group size and density vs. in-group variance

No correlation was found between group size and standard deviation for variable use within groups, or between density and standard deviation. That said, I calculated correlations twice: once for groups 1-26, and once for groups 1-26 along with the standard deviation, density, and group size for the entire dataset. When the whole dataset was included, the p-values for correlation between standard deviation and group size or density were lower, indicating a higher chance of correlation. However, no p-values were less than 0.05, meaning they were not significant enough to draw conclusions.

[22] "Mixed cases" refers to various patterns of lower- and upper-case letters, such as "wHaTeVvEr ThiS isS," cited above.

	Groups 1-26		Groups 1-26 and whole set	
	Correlation	P-value	Correlation	P-value
Density	-0.015	0.179	-0.021	0.060
Group size	0.005	0.681	0.017	0.121

Table 5: Variance vs. group size and density

That said, the conclusions the data points to are that group density is a better predictor of variance than group size, and that density correlates negatively with variance, while size correlates positively. In other words, a large, low-density group is likely to have a greater variance in feature usage than a smaller, higher-density group. This trend leans toward but does not wholly support my initial hypothesis, and any correlations in the data are weak at best.

3.2. Discussion

3.2.1. P-value vs. frequency

To ensure that p-values were grounded in the data, and not the result of extremely low-frequency items (for instance, an item that appears to be highly related to group membership, but only appears once in a single group), I compared p-values with overall raw variable frequencies, or raw frequencies of nonstandard variants for those in alternation with a standard variant. The correlation value for group membership P-value and raw frequency was -0.0116, with a p-value of 0.062. The P-value, here indicating the likelihood of correlation, is not quite low enough to draw definitive conclusions, but still low enough to conservatively say that there may be a slight negative correlation. This is promising, as it implies that higher-frequency variables are generally influenced more by group membership, rather than less.

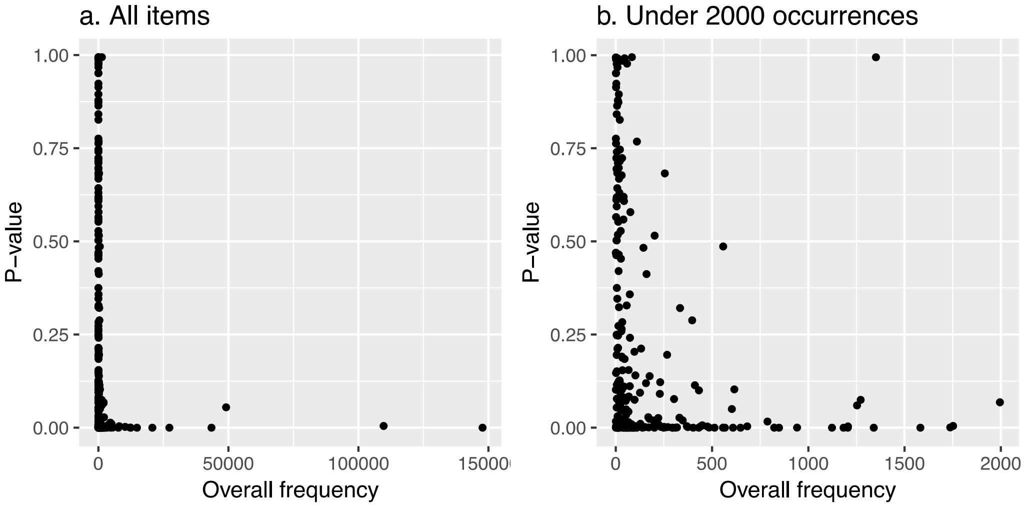

Figure 1: Raw frequency vs. p-value

The above figures illustrate the relationship between frequency and P-value. The first graph includes all variables, while the second includes only those with a raw frequency under 2000. Evidently, the most frequent variables all have a low P-value, while less frequent variables display mixed results.

3.2.2. Alternations among variables

Some variables have multiple nonstandard forms; for instance, both “thnks” and “thx” can represent “thanks.” However, in such cases there is often a disparity between the P-values for each form.

Standard form	Low-p-value form	Raw frequency	High-P-value form	Raw frequency
Got to go	gtg	11	g2g	1
Please	pls	942	plz	255
Thanks	thnks	2	thx	110
			thanku	12
			thnk u	0
Whatever	w/e	43	whatevs	12
Your, you’re	ur	2090	ure	3
			u’re	0
Something	sth	31	smthng	4

	smth	108		
Verbs				
Message	msged	3	msg	57
			msging	2
Cry	cwying	10	cwy	0
			cwied	0
Fuck	fcked	4	fck	31
			fcking	41

Table 6: Nonstandard alternations

Evidently, in many cases, the form with the lower P-value is much more frequent than the alternative, which may not appear at all in the case of “cwy/cwied.” Nevertheless, in some cases, more frequent forms have a higher P-value. Perhaps some higher-frequency items are spread out more evenly across groups, whereas items with very few instances only occur in a single group.

Looking at “msg/msged/msging” specifically, all instances of “msged” do indeed appear in a single group, and 100% of the instances of “msged/messaged” used by that group are the abbreviated form. Group differences appear for “msg” and “msging” as well, but not as strongly. “Msg” appears in sixteen out of 26 groups, and in many cases only one or two members of the group use the term. Furthermore, that “msg” may have an alternate meaning referring to a flavoring agent muddies the picture.

With regard to “fcked,” most of the members of group O used the term, while only two members of other, different groups used it, leading to its high degree of correlation. “Fck” and “fcking” are both much more common in the data, but almost all users of “fck” are outliers in their respective groups, wherein nearly no one else used “fck” at all. The distribution of “fcking” is similar, but less extreme.

The following figures are boxplots showing the distribution of "fck/fcked/fcking" for each group, shown on the y-axes. On the x-axes are the per-1000 word frequencies of each form. For each group, the first quartile, mean, and third quartile are shown, as well as any outliers. In Figure 2a, the first quarters, means, and third quarters for nearly every group are all zero, with all non-zero values as outliers. Group density is represented by color, with higher-density groups in a lighter blue and lower-density groups in a darker blue.

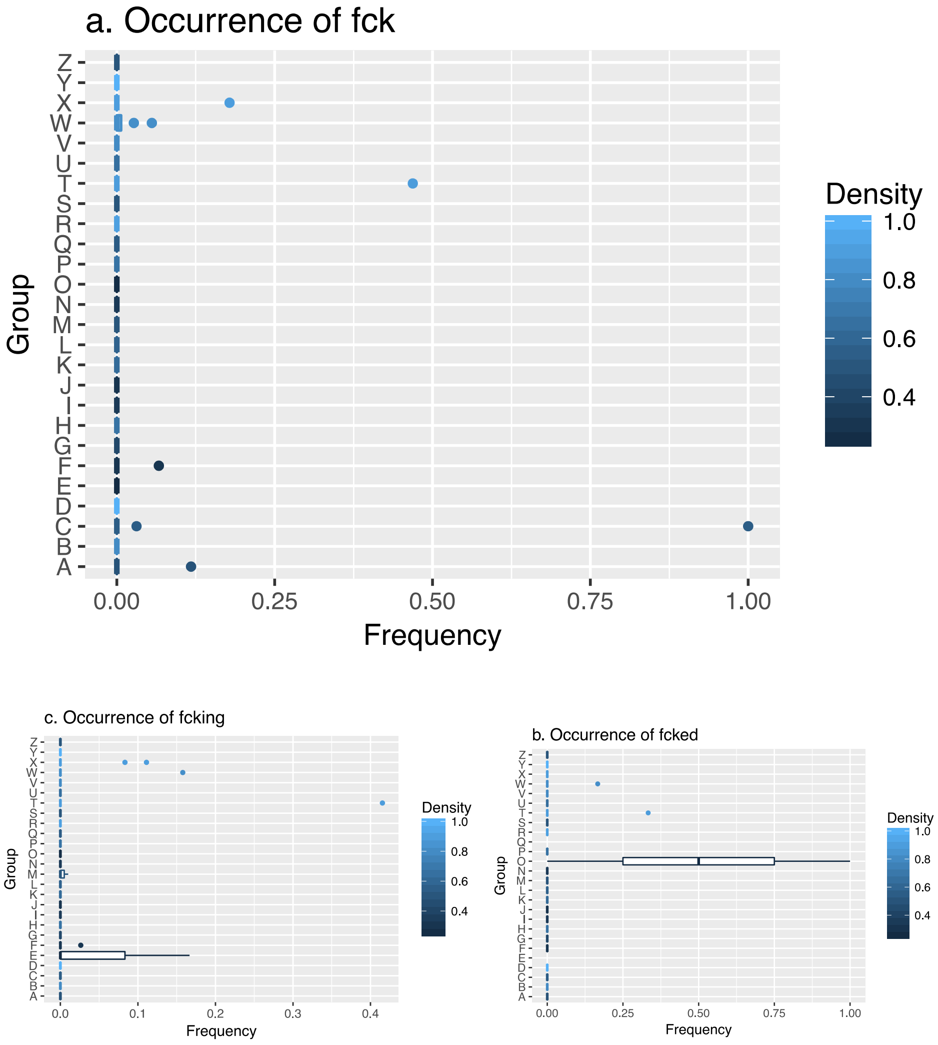

Figure 2: Occurrences of “fck,” “fcked,” and “fcking” within each group

3.2.3. *Keysmashes*

Keysmashes showed one of the lowest p-values, indicating that their usage is highly correlated with group membership. The following boxplot illustrates the range of keysmash usage for each group. As in Figure 2, the first quartile, mean, and third quartile, and outliers are shown for each group, and group density is represented by color, with higher-density groups in a lighter blue and lower-density groups in a darker blue.

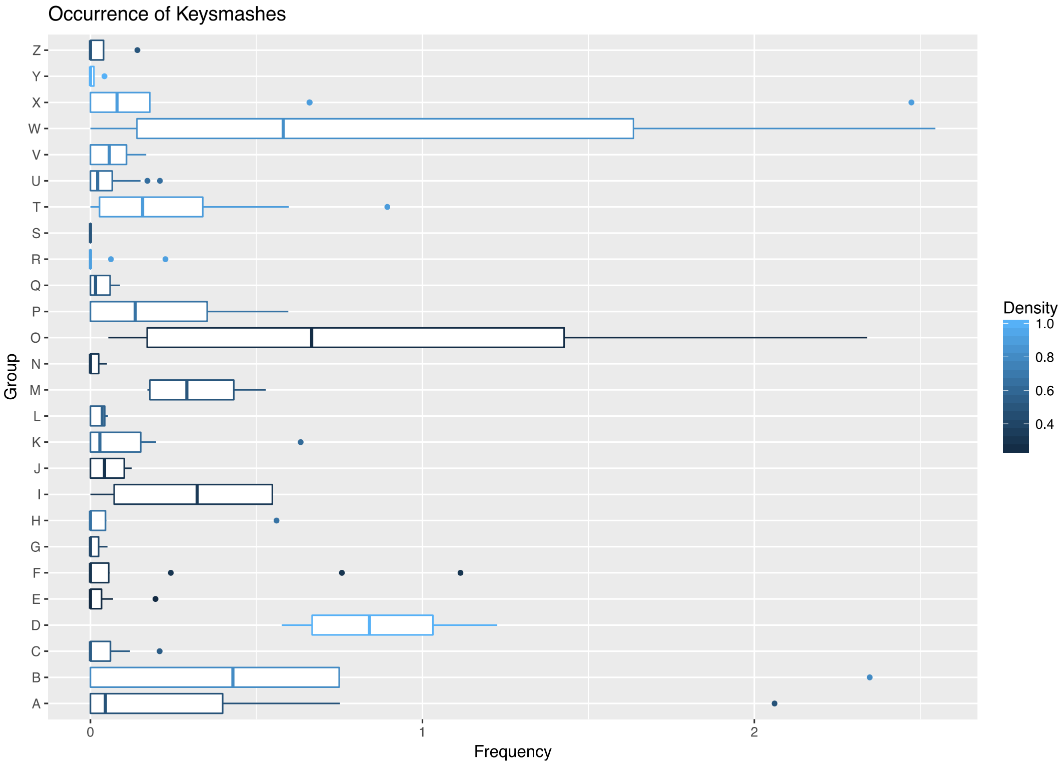

Figure 3: Keysmash usage within each group

To illustrate, I will enumerate the groups with the starkest contrasts in keysmash usage. The high-density groups R and Y barely use any keysmashes, group S uses none, and the high-density group D has the highest median keysmash usage, with every user using keysmashes.

The prevalence of the letters a, s, d, f, g, h, j, k, and l in keysmashes corroborates Tumblr user stanzicapparatireplayers's statement that "a deliberate keysmash will nearly always use keys only in the home row, and usually in a particular order that isn't likely to have happened purely accidentally." The following table shows the top 15 most frequent letters to appear in my list of 3,877 keysmashes. Frequency was determined by number of keysmash types (not individual tokens) a letter appears in.

	Letter	Frequency	In home row?
1	D	2990	✓
2	J	2877	✓
3	S	2594	✓
4	F	2562	✓
5	K	2246	✓
6	H	2017	✓
7	G	1630	✓
8	A	1313	✓
9	L	1223	✓
10	N	1043	
11	B	662	
12	M	450	
13	E	349	
14	W	283	
15	R	262	

Table 7: Frequent letters in keysmashes

The first nine letters indeed constitute the home row of the keyboard. Interestingly, the next three, b, m, and n, are adjacent on the keyboard, as are e, w, and r. The distribution of letters in keysmashes in my data is evidence that keysmashes are not entirely random, but are governed by certain rules or guidelines, which supports the idea that keysmash formation is influenced by the position of keys on the keyboard.

Nevertheless, keysmash results should be taken lightly, as I cannot guarantee every keysmash was counted. An avenue for future research would be to develop a more fine-tuned method of automatically identifying keysmashes, perhaps utilizing spellcheck algorithms and word frequencies to better exclude non-keysmashes, and the placement of keys on the keyboard to determine the likelihood that a string of characters could be a keysmash. Such an algorithm would enable further research into the characteristics of keysmashes (length, letter frequencies, capitalization, etc.) with regard to group membership and overall frequency.

3.2.4. Correlations between features

I would like to touch briefly on how the linguistic features I studied tend to correlate with each other. This goes hand-in-hand with the question of whether a community that tends to use certain features more will potentially use certain other features more (or less).

To accomplish this, I took the mean feature usage for each of the most significant (p<0.01) features in each group. I then computed the correlation and p-values for each pair of features using the cor.test function, as I did in section 3.2.1.

One feature in particular yielded the starkest results: "tweets beginning with uppercase letters." Of all the features I investigated, this one represents a standard variant, rather than a nonstandard one[23], as beginning textual utterances with capital letters is the prescriptive norm. Out of the 140 features I calculated correlations with this feature for, 62 correlations had a p-value under 0.05. Of these 62, all but four had a negative correlation, indicating that a higher number of tweet-initial capitals generally

[23] Tweets containing only capital letters were excluded from this category, as "all-caps" are not standard, and were placed in a different category.

predict a lower degree of nonstandard feature use. The four variables that correlated positively with tweet-initial capitalization were “bff(s),” “ty,” “wth,” and “wtf.”

One clear example of a negative correlation between tweet-initial capitalization and another feature is that of its correlation with keysmashes, as seen below.

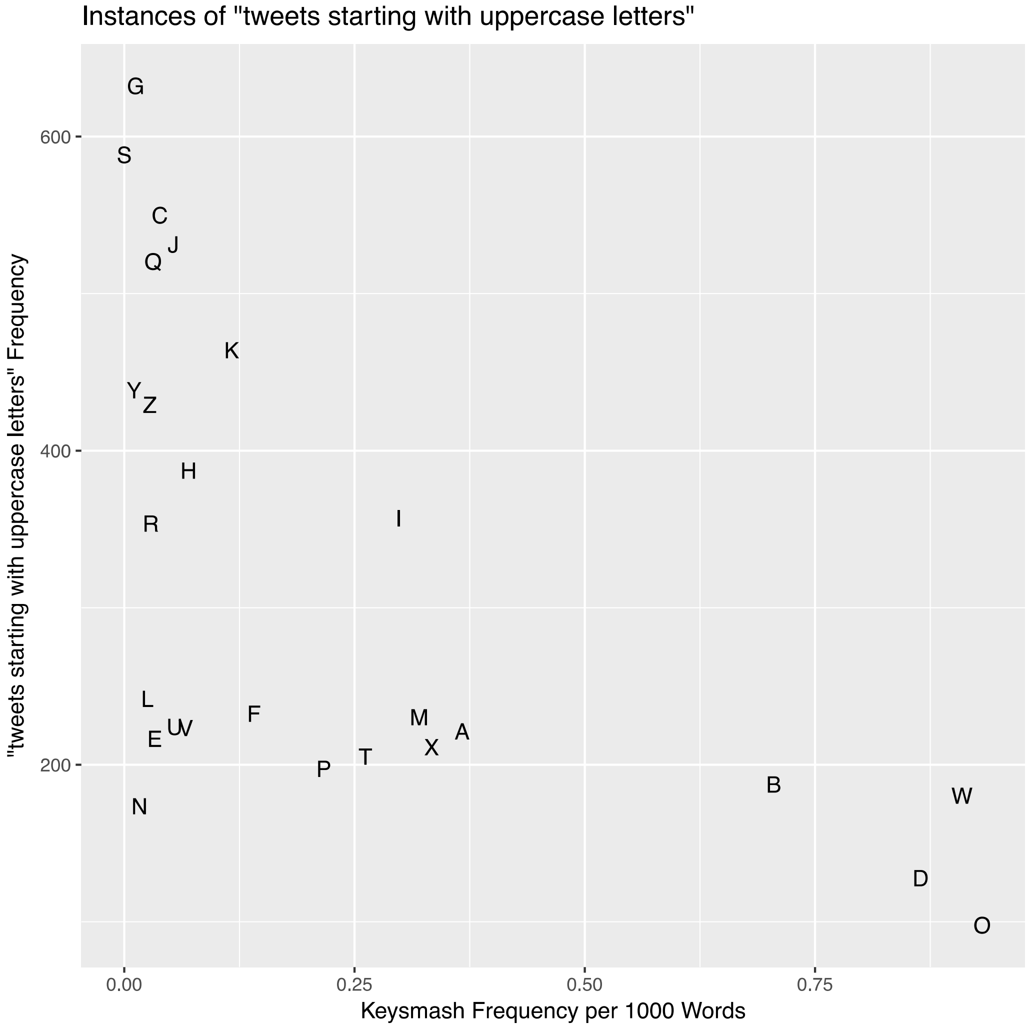

Figure 4: Average keysmash frequency vs. average frequency of tweets beginning with capital letters

Figure 4 shows the average number of keysmashes per 1000 words for each group, compared with the average number of tweets beginning with capital letters per 1000 tweets. The group with the highest average keysmash frequency, group O, has the lowest average frequency of tweets beginning with capitals, and groups W and D follow suit. Meanwhile, groups G and S have the highest average frequency of tweets beginning with capitals, and among the lowest keysmash frequencies, with no member of group S using any keysmashes in their tweets.

Interestingly, groups G and S each have the lowest average frequencies for many variables, with G having the lowest average for 62 variables and S for 67. Meanwhile, group O had the highest average frequency for 37 variables, more than any other group by far. Groups D and W have the highest averages for only a few variables each (four and seven respectively). The following table shows a selection of variables for which these five groups have either among the highest or lowest possible average frequencies. Items in italics do not have the highest or lowest average, but have a comparatively high or low average.

	Group D	Group G	Group O	Group S	Group W
Keysmashes	*High*	*Low*	High	Low	*High*
Tweets beginning with capitals	*Low*	High	Low	*High*	*Low*
STFU	Low	Low	Low	High	Low
IDGAF	Low	Low	Low	High	Low
WYD	*Low*	Low	Low	High	*Low*
Borger	Low	Low	Low	Low	High
Cronch	Low	Low	Low	Low	High
Gorl	*Low*	Low	Low	Low	High

Table 8: A comparison of eight variables in five groups

Each of these five groups have either a high frequency of keysmashes and a low frequency of tweet-initial capitals, or a low frequency of keysmashes and a high frequency of tweet-initial capitals, but each has its own distinct character. Of these groups, only S uses the acronyms “stfu,” “idgaf,” and “wyd,” and only W uses the alternative spellings “borger,” “cronch,” and “girl.” In contrast, groups D and O have relatively high frequencies for many other nonstandard variables, but not those used by S or W. G, on the other hand, hardly uses any nonstandard variables. Thus, groups D, O, S, and W all seem to use informal, though not identical, registers, while G appears to use a more formal register. Furthermore, it seems that the acronyms used by group S (but not the other groups) form a bundle of co-occurrent features, as do the alternate spellings used by group W. In the case of the spellings, this association is all the more apparent, as all three spellings replace what would be pronounced as an unrounded mid-vowel with an “o,” indicating a rounder pronunciation produced farther back in the mouth.

4. Conclusion

The results of my research strongly support my hypothesis that Twitter users' usage of linguistic features associated with CMC is affected by the feature-usage of those with whom they share social bonds. In other words, as in traditional communities, Twitter users in immediate contact with each other tend to share linguistic features. It is evident that there are many such features potentially affected by community membership; however, the effects of group size and density on variability remain inconclusive.

Although this study supports my core hypothesis, other potential variables affecting language use on Twitter must be studied further. For instance, the effect of social categories remains difficult to evaluate. Because my study focuses on community membership, and because I opted to ignore personal information (gender, age, location, etc.) so as to better gain a representative sample of Twitter users, most of whom are at least partially anonymous, I necessarily excluded identity-related factors from my analysis. This is not to say that factors such as gender, race, and class do not affect language use on Twitter; simply that they are beyond the scope of my research. Due to this anonymity, it may be impossible to fully disentangle the effects of community membership from those of categorical factors, especially given that "A person's position in a social network could reflect that person's social choices" and both choice of community and language "may simply reflect other, as yet unidentified factors, such as underlying attitudes to the local group" (Marshall, 2004, p. 19). In other words, people may choose their community such that it is made up of people like themselves. However, given the amount and variety of variables that correlate with group membership, it seems unlikely that the Twitter users studied joined groups made up of people who happened to already share these features, and more likely that they share these features due to exposure to them from within the group.

Another issue that I cannot account for is the effect of other social networks; many Twitter users surely belong to multiple social networking websites, all of which may affect their speech. Since my data consisted of tweets from over 200 individuals, it was not possible for me to perform such in-depth analysis. A potential future line of research could be the close investigation of language use on multiple websites by a small group of individuals. Another study that would benefit from a small data pool would be the study of how online communities affect the usage of syntactic features.

Other avenues for future research concern the specific variables under study. While I have shown that a selection of variables may be affected by group membership, I cannot yet account for why those particular variables correlate, while others do not. What kind of formal or functional characteristics, if any, affect whether a linguistic feature's usage correlates with group membership? What characteristics affect the frequency of features, or the likelihood that their frequency will increase over time? Perhaps the latter question could be answered by a longitudinal study.

The field of internet sociolinguistics, like the internet itself, is young and ever-expanding. Many questions remain for future researchers to investigate. Nevertheless, I hope that the present study can provide another piece of the puzzle, and that it may elucidate the similarity between the effects of traditional and online communities on language usage.

Appendix A: Regular Expressions

<table>
<tr><th>Expression</th><th>Meaning</th></tr>
<tr><td>.</td><td>Any character.</td></tr>
<tr><td>\\b</td><td>Any word-boundary, such as the boundary between a letter and a space.</td></tr>
<tr><td>\\s</td><td>Any empty space.</td></tr>
<tr><td>*</td><td>Indicates the previous expression may appear any number of times, including none.</td></tr>
<tr><td>+</td><td>Indicates the previous expression may appear one or more times.</td></tr>
<tr><td>?</td><td>Indicates the previous expression may or may not appear.</td></tr>
<tr><td>|</td><td>Indicates that either or any expression out of a list may appear. For instance, a|i|e|o|u looks for instances of a, i, e, o, and u.</td></tr>
<tr><td>( )</td><td>Indicates that the characters inside should be treated as a singled expression. For example, (ha)+ catches ha, haha, hahaha, and so on, while ha+ catches ha, haa, haaa, and so on.</td></tr>
<tr><td>{ }</td><td>Contains a number or range of numbers indicating the number of times an expression may appear. {,2} means up to two, {2,} means two or more, and {2,4} means two to four.</td></tr>
<tr><td>[ ]</td><td>Indicates a range of possible ASCII values. [a-z] would be any alphabetical character from a to z, lowercase.</td></tr>
<tr><td>^</td><td>Indicates the beginning of a string.</td></tr>
<tr><td>$</td><td>Indicates the end of a string.</td></tr>
<tr><td>\\ or \</td><td>Indicates that a character like . or ? is acting as a character and not a regular expression. \\. will catch all periods, while . will catch all characters.</td></tr>
</table>

Table AA1: Regular expressions

Appendix B: Variable Lists

Variable	Standard variant	Regular expression
abt	about	abt
neone	anyone	neone
bc	because	bc
buncha	bunch of	buncha
convo(s)	conversation(s)	convos?
coulda	could have	coulda
every1	everyone	every1
fav(s)	favorite(s)	favs?
fave(s)	favorite(s)	faves?
foreal	for real	foreal
gonna	going to	gonna
gnight	good night	gnight
g’night	good night	g’night
gr8	great	gr8
gotta	got to	gotta
kno	know	kno
l8r	later	l8r
luv	love	luv
mb	maybe	mb
msg	message	msg
msged	messaged	msged
msging	messaging	msging
nbhd(s)	neighborhood(s)	nbhds?
nvm	nevermind, never mind	nvm
notif(s)	notification(s)	notifs?
obv	obviously	obv
ofc	of course	ofc
ppl	people	ppl
prsnl	personal	prsnl
prsnly	personally	prsnly
pls	please	pls
plz	please	plz
prob	probably	prob
protag(s)	protagonist(s)	protags?
rly	really	rly
srsly	seriously	srsly
shoulda	should have	shoulda
some1	someone	some1
sum1	someone	sum1
sth	something	sth
smth	something	smth
smthng	something	smthng
sry	sorry	sry
thx	thanks	thx
thnks	thanks	thnks
thanku	thank you	thanku
thnk u	thank you	thnk u
tho	though	tho
tmrw	tomorrow	tmrw
ttly	totally	ttly
w8	wait	w8
w/e	whatever	w/e
whatevs	whatever	whatevs
w/	with	w/
w/o	without	w/o
woulda	would have	woulda
u’re	you’re	u’re
ure	youre	ure
ur	your	ur
urself	yourself	urself

Table AB1: Abbreviations

Variable	Meaning	Regular expression
af	as fuck	af
afaik	as fair as I know	afaik
afk	away from keyboard	afk
brb	be right back	brb
bff(s)	best friend(s) forever	bffs?
bf	boyfriend	bf
btw	by the way	btw
fml	fuck my life	fml
ftw	for the win	ftw
fyi	for your information	fyi
gf	girlfriend	gf
gtfo	get the fuck out	gtfo
gtg	got to go	gtg
g2g	got to go	g2g
gdi	god damn it	gdi
icymi	in case you missed it	icymi
idc	I don’t care	idc
idgaf	I don’t give a fuck	idgaf
idk	I don’t know	idk
idr	I don’t remember	idr
iirc	if I recall correctly	iirc
ikr	I know right	ikr
ily	i love you	ily
ilysm	i love you so much	ilysm
imho	in my humble opinion	imho
imo	in my opinion	imo
inb4	in before	inb4
jfc	jesus fucking christ	jfc
jic	just in case	jic
jk	just kidding	jk
jw	just wondering	jw
kms	kill myself	kms
kys	kill yourself	kys
lmao	laughing my ass off	lmao
lmgdao	laughing my god damn ass off	lmgdao
lmk	let me know	lmk
lms	like my status	lms
lol	laugh out loud	lol
nbd	no big deal	nbd
np	no problem	np
omfg	oh my fucking god	omfg
omg	oh my god/gosh	omg
omw	on my way	omw
otoh	on the other hand	otoh
rn	right now	rn
rofl	rolling on (the) floor laughing	rofl
cu	see you	cu
smh	shakes my head	smh
stfu	shut the fuck up	stfu
s2g	swear to god	s2g
ttyl	talk to you later	ttyl

tbf	to be fair	tbf
tbh	to be honest	tbh
tbt	throwback thursday	tbt
tf	the fuck	tf
tfw	that feel(ing) when	tfw
tmi	too much information	tmi
ty	thank you	ty
tysm	thank you so much	tysm
wtf	what the fuck	wtf
wth	what the hell/heck	wth
wwyd	what would you do	wwyd
wyd	what you do	wyd
yw	you’re welcome	yw

Table AB2: Acronyms

Variable	Standard variant	Regular expression
im	i’m	im
ive	i’ve	ive
youre	you’re	youre
youve	you’ve	youve
youll	you’ll	youll
yall	y’all	yall
youd	you’d	youd
weve	we’ve	weve
hes	he’s	hes
hed	he’d	hed
shes	she’s	shes
theyre	they’re	theyre
theyve	they’ve	theyve
theyll	they’ll	theyll
theyd	they’d	theyd
isnt	isn’t	isnt
aint	ain’t	aint
arent	aren’t	arent
wasnt	wasn’t	wasnt
werent	weren’t	werent
dont	don’t	dont
didnt	didn’t	didnt
doesnt	doesn’t	doesnt
wont	won’t	wont
cant	can’t	cant
havent	haven’t	havent
hasnt	hasn’t	hasnt
hadnt	hadn’t	hadnt
wouldnt	wouldn’t	wouldnt
wouldve	would’ve	wouldve
wouldntve	wouldn’t’ve	wouldntve
couldnt	couldn’t	couldnt
couldve	could’ve	couldve
couldntve	couldn’t’ve	couldntve
shouldnt	shouldn’t	shouldnt
shouldve	should’ve	shouldve
shouldntve	shouldn’t’ve	shouldntve
mustnt	mustn’t	mustnt
mustve	must’ve	mustve
mustntve	mustn’t’ve	mustntve
heres	here’s	heres
herell	here’ll	herell
hered	here’d	hered
theres	there’s	theres
therell	there’ll	therell
thered	there’d	thered
itll	it’ll	itll
itd	it’d	itd

thats	that’s	thats
thatll	that’ll	thatll
thatd	that’d	thatd
whens	when’s	whens
whenll	when’ll	whenll
whend	when’d	whend
whats	what’s	whats
whatll	what’ll	whatll
whatd	what’d	whatd
whys	why’s	whys
whyll	why’ll	whyll
whyd	why’d	whyd
everythings	everything’s	everythings
somethings	something’s	somethings
nothings	nothing’s	nothings
anythings	anything’s	anythings
everyones	everyone’s	everyones
anyones	anyone’s	anyones
no ones	no one’s	no ones
someones	someone’s	someones
everythingd	everything’d	everythingd
somethingd	something’d	somethingd
nothingd	nothing’d	nothingd
anythingd	anything’d	anythingd
everyoned	everyone’d	everyoned
anyoned	anyone’d	anyoned
no oned	no one’d	no oned
someoned	someone’d	someoned
everythingll	everything’ll	everythingll
somethingll	something’ll	somethingll
nothingll	nothing’ll	nothingll
anythingll	anything’ll	anythingll
everyonell	everyone’ll	everyonell
anyonell	anyone’ll	anyonell
no onell	no one’ll	no onell
someonell	someone’ll	someonell
maam	ma’am	maam

Table AB3: Apostrophes

Variable	Regular expression
all uppercase	^[A-Z0-9]+$
all lowercase	^[a-z0-9]+$
tweet starts with uppercase	^[A-Z]
tweet starts with lowercase	^[a-z]
UPPercase	[A-Z]{2,}[a-z]
uppercase	[a-z][A-Z]
mixed pattern 1	([a-z][A-Z]){2,}
mixed pattern 2	[a-z]{2,}[A-Z]{2,}[a-z]{2,}[A-Z]{2,}
mixed pattern 3	[a-z]+[A-Z]+[a-z]+[A-Z]+
mixed pattern 4	[a-z]+[A-Z]+[a-z]+
mixed pattern 5	[A-Z]+[a-z]+[A-Z]+

Table AB4: Capitalization

Variable	Regular expression
Oh	Oh
Ohh	oh{2,}
Oohhh	o{2,}h{2,}
ugh	Ugh
ughh	ugh{2,}
nooo	no{2,}
nono	(no)+
wow	Wow
woww	wow{2,}
woowww	w{2,}w{2,}
wowowow	w(ow)+
ya	Ya
yaaa	ya{2,}
yaya	(ya){2,}
yayaaa	(ya){2,}a+
yay	Yay
yayyy	yay{2,}
yaayyy	ya{2,}y{2,}
ay	Ay
ayyy	ay{2,}
aayyy	a{2,}y{2,}
ey	ey
eyyy	ey{2,}
eeyyy	e{2,}y{2,}
yea	yea
yeaaa	yea{2,}
yeah	yeah
yeahhh	yeah{2,}
ack	ack
aaack	a{2,}ck
aacckk	a{2,}c+k{2,}
grr	grr
grrr	gr{3,}
pft	pft
pffft	pf{2,}t
pfffft	pf{2,}t{2,}
pfh	pfh
pfffh	pf{2,}h
pfffhh	pf{2,}h{2,}
pff	pff
pfff	pf{3,}
ffh	ffh
fffh	f{3,}h

Table AB5: Exclamations

Feature	Code
uh	Uh
uhh	uh{2,}
um	Um
umm	um{2,}
eh	Eh
ehh	eh{2,}
ah	Ah
ahh	ah{2,}
erm	erm
ermm	erm{2,}
huh	huh
huhh	huh{2,}
hm	hm
hmm	hmm{2,}

Table AB6: Filler

Variable	Regular expression
Haha	(ha){2}
Hahaha	(ha){3,}
Ahahahah	(b|w)?a?(ha){2,}h

Bahahaha	ba(ha){2,}h?
Wahahaha	wa(ha){2,}h?
Hehe	(he){2}
Hehehe	(he){3,}
Heheheh	e?(he){2,}h
Eheheheh	e(he){2,}h?
Heehee	(hee){2}
heeheehee	(hee){3,}
Teehee	tee(hee)+
Hihihi	(hi){2,}
Tihihi	ti(hi){2,}
Huhuhu	(hu){2,}
Huehue	(hue){2,}
Hohoho	(ho){3,}
Nyahaha	nya(ha)+
Nyehehe	nye(he)+
nyohoho	nyo(ho)+

Table AB7: Laughter

Variable	Regular expression
ellipses (just 3)	\\w\\.{3}[^.]
extra ellipses (4 or more)	\\.{4,}
extra ellipses (5 or more)	\\.{5,}
extra ellipses (6 or more)	\\.{6,}
extra exclamations (2 or more)	!{2,}
extra exclamations (3 or more)	!{3,}
extra exclamations (4 or more)	!{4,}
extra question marks (2 or more)	\?{2,}
extra question marks (3 or more)	\?{3,}
extra question marks (4 or more)	\?{4,}
interrobang	(\?!|!\?)+
extra commas (2)	,{2}[^,]
extra commas (3 or more)	,{3,}
extra commas (4 or more)	,{4,}
extra semicolons (2 or more)	;{2,}
extra semicolons (3 or more)	;{3,}
extra semicolons (4 or more)	;{4,}
extra slashes (2 or more)	/{2,}
extra slashes (3 or more)	/{3,}
extra slashes (4 or more)	/{4,}
tildes (2 or more)	~{2,}
tildes (3 or more)	~{3,}
tildes (4 or more)	~{4,}
right shift	>>>
sarcasm text	~\*.\~
emphasis text	\\S\\s{4,}
apostrophes around text	\*.\
script format	\\bme:\\s
tm	™

Table AB8: Punctuation

Variable	Standard variant	Regular expression
gdo	God	gdo
birfday	Birthday	birfday
amirite	am i right	amirite
blease	please	blease
pwease	please	pwease
hewwo	hello	hewwo
cwy	Cry	cwy
cwied	cried	cwied
cwying	crying	cwying
fck	fuck	fck
fcked	fucked	fcked
fcking	fucking	fcking
gorl	girl	gorl
smol	small	smol
tol	tall	tol
lorge	large	lorge
borger	burger	borger
cronch	crunch	cronch

Table AB9: Nonstandard spelling

Variable	Standard variant	Regular expression
-ign	-ing	ing$
-in'	-ing	ing$
-in	-ing	ing$
-inf	-ing	ing$
-en	-ing	ing$
-cc	-ck	ck$
-kc	-ck	ck$
-ie	-y	y$

Table AB10: Nonstandard suffixes

Bibliography

a6, anemotionallyunstablecreature, and stanzicapparatireplayers. (2017, December 2). Retrieved from http://stanzicapparatireplayers.tumblr.com/post/168103393903/anemotionallyunstablecreature-a6-u-kno-when-u.

Androutsopoulos, J. (2006). Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics*, *10*(4), 419-438.

averagefairy, feynites, and runawaymarbles. (2017, December 16). Retrieved from https://feynites.tumblr.com/post/168623069349/runawaymarbles-averagefairy-old-people-really.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135-160.

bec (oldyelIer). (2016, September 4, 12:19 UTC). me typing. Retrieved from https://twitter.com/oldyelIer/status/772408669296209925.

Beißwenger, M., & Storrer, A. (2008). Corpora of Computer-Mediated Communication. *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

boyd, d., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210-230.

Cheshire, J. Linguistic variation and social function. Romaine, 153-166.

Csardi, G., & Nepusz, T. (2006) The igraph software package for complex network research. InterJournal, Complex Systems 1695. Retrieved from http://igraph.org.

crtter, anexperimentallife, & tertiusiii. (2017, October 5). Retrieved from http://anexperimentallife.tumblr.com/post/166088872849/tertiusiii-crtter-intentional-misspellings.

Crystal, D. (2005, February). The scope of Internet linguistics. In *Proceedings of American Association for the Advancement of Science Conference; American Association for the Advancement of Science Conference, Washington, DC, USA* (pp. 17-21).

Eckert, P. (1989). *Jocks and Burnouts: Social Categories and Identity in the High School*. Teachers College Press.

Gentry, J. (2015). twitteR: R Based Twitter Client. R package version 1.1.9. Retrieved from https://CRAN.R-project.org/package=twitteR.

Hu, Y., Wood, J. F., Smith, V., & Westbrook, N. (2004). Friendships through IM: Examining the relationship between instant messaging and intimacy. *Journal of Computer-Mediated Communication*, *10*(1), JCMC10111.

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication*, *10*(2), JCMC10211.

Jones, T. (2015). Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter." *American Speech*, *90*(4), 403-440.

Kapidzic, S., & Herring, S. C. (2011). Gender, communication, and self-presentation in teen chatrooms revisited: Have patterns changed? *Journal of Computer-Mediated Communication*, *17*(1), 39-59.

Labov, W. (1972). *Language in the Inner City: Studies in the Black English Vernacular* (Vol. 3). University of Pennsylvania Press.

Labov, W. (1990). The Intersection of Sex and Social Class in the Course of Linguistic Change. *Language Variation and Change*, *2*(2), 205-254.

Lyddy, F., Farina, F., Hanney, J., Farrell, L., & Kelly O'Neill, N. (2014). An analysis of language in university students' text messages. *Journal of Computer-Mediated Communication*, *19*(3), 546-561.

Mallinson, C., Childs, B., & Van Herk, G. (2013). *Data Collection in Sociolinguistics: Methods and Applications*. Routledge.

Marshall, J. (2004). *Language Change and Sociolinguistics: Rethinking Social Networks*. Springer.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(2), 339-384.

Milroy, L. Social network and linguistic focusing. Romaine, 141-152.

Murray, S. O. (1993). Network determination of linguistic variables? *American Speech*, *68*(2), 161-177.

Paolillo, J. (1999). The virtual speech community: Social network and language variation on IRC. *Journal of Computer-Mediated Communication*, *4*(4), JCMC446.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Raclaw, J. (2006, October). Punctuation as Social Action: The Ellipsis as a Discourse Marker in Computer-Mediated Communication. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 32, No. 1, pp. 299-306).

Romaine, S. (Ed.). (1982). *Sociolinguistic Variation in Speech Communities*. Arnold.

Romaine, S. What is a speech community? Romaine, 13-24.

sleepy bitch nisha (corpsejaw). (2017, July 22, 2:08 UTC). if youve ever talked to me You Know. Retrieved from https://twitter.com/sunsetstarboy/status/888581379373625344.

steveogers, cappuccinohowell, elenorekarat, fandom-scarein, fihli, honey-stick, marquis-d-la-baguette, open-plan-infinity, oxime-anime, popcorn-fox, pepperminthowell, poseidhn, starlight-sanders, studyandlush, twentyonelizards, & watermellens. (2018, February 11). Retrieved from https://elenorekarat.tumblr.com/post/170775118441/oxime-anime-fandom-scarerin-popcorn-fox.

Tamburrini, N., Cinnirella, M., Jansen, V. A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. *Social Networks*, *40*, 84-89.

Thurlow, C. (2006). From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication*, *11*(3), 667-701.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (Vol. 8). Cambridge University Press.

Wickham, Hadley. (2018). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.3.0. Retrieved from https://CRAN.R-project.org/package=stringr.

□the wiggler□ (_lps666). (2018, March 7, 10:10 UTC). things the gays(tm) say. Retrieved from https://twitter.com/_lps666/status/971447901288189954.