Orthographic variation of (lol)*

Joshua McNeill

May 18, 2021

Abstract

[TBD]

1 Introduction

Studies of orthographic variation are not unheard of in sociolinguistics. This has been especially true since the widespread adoption of the internet as computer-mediated communication (CMC) presents both a wealth of relatively easily accessible data and social contexts that are less subject to overt social controls than what would have historically been the case for written forms as they had previously been largely relegated to educational and literary contexts. However, CMC studies and studies of orthographic variation in general have been focused mostly, though not always, either on situating CMC along a continuum between spoken language and older forms of written language or on the connection between phonology and spelling.

With some notable exceptions, an area of orthographic variation that has not been explored as much is how it functions on its own terms, independent of phonology. As such, the aim of the present study is to give an in depth examination of the potential social and pragmatic associations of variants of an orthographic variable that cannot reasonably be linked to either spoken language or educational and literary writing: (lol) 'laugh out loud'.

1. James Ramirez[1]: @xebeche914 Nobody knows nothing...lol
2. Josiah Eguino Silvas: @AthenaBass wave 2 me when u are on the plane over Nova Scotia LOl. Your getting closer LOL. Thx thumper Gr8 2 C U gigging
3. Alyshah Tsegay: @angew lololol that's so clever!!! Where are we flying to??? Muahaha!

*Data and code available at https://osf.io/mgdpu/.

[1]Users whose tweets appear in the corpus used here have all been anonymized without concern for gender nor social identity.

As tweets 1, 2, and 3 show, there are various ways that (lol) can be spelled, ranging from all lowercase ⟨lol⟩ to all uppercase ⟨LOL⟩ to repeated characters that defy this item's origin as an acronym, as in ⟨lololol⟩. These three tweets alone also display numerous non-standard orthographic practices such as repeated punctuation, abbreviations, acronyms, and characters meant to be understood as some homophonous word or portion of a word, such as ⟨2⟩ for *to*. Moreover, these tokens of (lol) also coincide with other non-standard features of English such as the negative concord in tweet 1. While any of these linguistic features could prove interesting to analyze, only (lol) is examined here. Specifically, we will look at (lol) as used on the social media platform Twitter, whether certain spelling variants are associated with certain Twitter communities or certain positions within communities and whether certain spelling variants are associated with particular sentiments.

The rest of this introduction will be structured as follows. Section 1.1 will cover the general nature of CMC and any special considerations that are applicable to the present study. Section 1.2 will review work that has been done on orthographic variation. Finally, a thorough review of work that has been done on (lol), whether as a lexical or orthographic variable, will be presented in section 1.3.

1.1 Computer-mediated communication

With the advent of the internet, many new mediums of communication have entered daily life, collectively referred to as CMC. Likewise, by the end of the 1990s, research focused on these new mediums began in ernest. For instance, Cherny (1996) examined multi-user dungeons (MUDS), which are online, text-based, multiplayer games, whereas Yates (1996) analyzed language in computer conferencing systems, which are perhaps best thought of as progenitors of discussion forums, and Paolillo (2001) looked at chatrooms on internet relay chat (IRC). These were all early mediums, and so studies have progressed to other mediums as they have been developed, which include instant messaging (IM) (e.g., Baron, 2004; Tagliamonte & Denis, 2008), the social media photo-sharing platform Instagram (e.g., Stewart et al., 2017), and both commonly and importantly for the present study, the social media micro-blogging platform Twitter (e.g., Bamman et al., 2014; Eisenstein, 2013; Hong et al., 2011; Ilbury, 2020; Jones, 2015; Kim et al., 2014). Schneier (2021) also looked at text messaging on cell phones, though his data included use of any communication application on a cell phone.

Androutsopoulos (2008a) described the goals of these researchers of CMC as fitting into two separate "waves",[2] the first focused on the impact of the constraints placed on users of different CMC mediums on the language they produce, and the second focused more on analyses of what happens in CMC pragmaticly and sociolinguisticly (pp. 1-2), though one might also add that

[2]This is not to be confused with Eckert's (2012) description of three "waves" of sociolinguistics, though there comparison is apt.

there has been a consistent interest in characterizing the relationship between CMC and both speech and older written mediums throughout both of these waves. A particularly useful tool for the present study – and indeed any CMC study – that was developed during the first wave is a typology for CMC mediums.

In order to compare results from a study centered on Twitter, as we are doing here, with results from previous work, it is important to have a framework for classifying different mediums. As Paolillo (1999) recognized early on CMC is not identical to face-to-face conversation (p. 1). Yates (1996) had previously gone even further, suggesting that CMC occurred in the absence of any field à la Halliday other than the text itself (pp. 45-46). Baron (2004) echoed these general sentiments, adding to them that the medium may change "the character of language produced in that medium," leading her to formulate perhaps the most prevalent typological framework in CMC for mediums, which includes two dichotomous parameters: synchronicity versus asynchronicity, and one-to-one versus one-to-many interactions (p. 398). The first parameter, synchronicity versus asynchronicity, refers to whether there is a reasonable expectation between the interlocutors that messages will be received and responses made immediately, as if in a face-to-face conversation. The second parameter, one-to-one versus one-to-many interactions, refers to whether locutors are sending messages that are meant to be received either by one person or by many people.

Baron's (2004) typological framework for CMC is useful for understanding how comparable data produced through different mediums is as there may be different limitations placed on the forms of messages. For instance, Baron (2004) suggested that chatrooms, MUDs, and IM are all synchronous CMC mediums, whereas text messaging is asynchronous as one may not expect immediate replies to text messages (p. 398). The implication is that data collected from chatrooms and IM are more comparable than data collected from chatrooms and text messages.

Indeed, there are clear examples of the synchonocity or asynchronicity of a medium having an effect on how conversations progress. Baron (2004) herself noted a phenomenon in instant messaging in which multiple topics overlap as a result of both interlocutors being able to construct messages simultaneously without interrupting each other (p. 400), which would not happen in a less synchronous medium such as e-mail. Somewhat relatedly, she observed a prevalence of multiturn sequences, where each sent message is considered a turn. Baron (2004) argued that interlocutors face pressure to break up their messages into multiple shorter turns in order to hold the floor (p. 417), the idea being that, since messages are not seen until sent, longer single turn messages provide more time for the other party to chime in before the whole statement is ready to be sent. These sorts of phenomena suggest that one must be cautious when comparing the language used in one CMC medium versus another.

However, the boundaries between what should be considered synchronous and asynchronous are somewhat fuzzy. While chatrooms would presumably yield situations where interlocutors immediately reply to each other, this is

not necessarily the case. IRC still exists today, though the way it is used may have changed over the last 20 years. For instance, at the time of this writing, the chatroom #nlp on the freenode network, which offers discussion and help with natural language processing, has a topic that states that it may take two to three hours to get a response to a question. The reason is that it is typical to be connected to an IRC chatroom without monitoring it closely, making it less synchronous than might be expected. Indeed, Baron (2004) herself acknowledged that IM users in her data were sometimes multitasking while connected and so did not always respond immediately (p. 419). Even in clearer cases of asynchronicity, such e-mail, there is still some imperfection in the classification as many people now own cell phones that receive e-mail notifications instantly wherever they are, allowing for relatively quick responses, though quick responses are not necessarily expected. This parameter is still useful, of course, since a medium can be thought of as more or less synchronous in terms of expectations – in our case, Twitter is relatively asynchronous but can be used from a cell phone that provides immediate notifications still – but there are caveats to keep in mind.

It is also important to consider the audience for messages. Baron's (2004) framework accounts for this with the one-to-one versus one-to-many parameter. Protypical examples of each might be e-mail for the former and blogs for the latter, although here again one finds some fuzziness in the categorization. While e-mail might typically be a one-to-one medium, one-to-many messages are not unknown.

This same difficulty in categorization is present on Twitter, as well, where the most common sort of messages may be one-to-many, but there exist public directed messages and private directed messages. Again, the parameter here is still useful, but in this case, there is also an alternative: Bell's (1984) audience design framework. Indeed, Pavalanathan and Eisenstein (2015) employed this framework on Twitter, which yielded interesting results, as will be discussed further in the Methods section. Likewise, the audience design framework would have been useful when Androutsopoulos (2008b) had found "hip-hop slang" to be used more on German hip-hop discussion forums than on German hip-hop homepages and online magazines (p. 293). Both of these venues are indeed one-to-many when approached from Baron's (2004) typology, but the results are not the same. One could instead argue from the audience design framework that the audience for a homepage or magazine is far broader than for a discussion forum, leading to more standard language. What is to be taken away from this is that audience appears to impact the character of the language produced on Twitter, which can inform how we interpret results in the present study.

1.2 Orthographic variation

From early on in CMC research, it has been acknowledged that CMC is not only different from face-to-face language, as Paolillo (1999) recognized, but that it is also different from older forms of writing. Indeed, Tagliamonte and

Denis (2008), speaking of instant messaging, described it as a "hybrid register", somewhere between speech and older forms of writing (p. 5). Part of this hybridity perhaps stems from CMC taking place outside of the auspices of social institutions, such as schools and book publishers, that extert overt social control on what is acceptable. This makes CMC a fruitful area for locutors to creatively manipulate linguistic features to various ends, one of those features being the focus of the present study: orthography. It is thus useful to review the literature on orthographic variation to better understand what functions and associations it has, as that is what we would like to analyze for ⟨lol⟩. We group these under four broad categories: those related to 1) grammar, 2) audience, 3) community or subsets of a community, and 4) pragmatics.

Before moving on to the possible functions and associations of orthographic variation, it should first be noted that orthographic variation is unlikely to simply be the result of poor spelling capability. Varnhagen et al. (2010), in their examination of instant messaging, compared participants scores on spelling tests to their use of non-standard spellings in IM. They found no relationship between the two, suggesting that non-standard spelling is not the result of a poor grasp of standard spelling. In fact, they found that non-standard spelling norms were acquired quite readily. For example, ⟨shoulda⟩ 'should have' was found but never forms like ⟨shulda⟩ (Varnhagen et al., 2010, p. 731). The implication is that non-standard spellers may actually be very good spellers if the qualification of "good" in this case is defined as 'accurately follows norms'.

1.2.1 Connection to grammar

It is not unusual for variant spellings to be constrained by grammar. A very standard and overt example of this comes from French verbs. For most verbs, the singular first, second, and third person present forms of verbs are phonetically identical, yet the written forms vary anyway to agree with the subject. *Parler* 'to speak' is pronounced [paʀl] for all three singular subjects in perhaps most varieties of Hexagonal French,[3] but is written ⟨parle⟩ for first and third person and ⟨parles⟩ for second person. It is thus useful to consider other potential grammatical constraints on orthographic variation.

One such case can be found in Hinrichs and White-Sustaíta's (2011) work on Jamaican Creole, which has English as its superstrate. They found that speakers of the language living outside of Jamaica would vary between spellings such as ⟨mi⟩ and ⟨me⟩, both pronounced [mi], based on the syntactic function within the sentence, using the former as the subject pronun and the latter in all other cases. Hinrichs and White-Sustaíta (2011) argued that ⟨mi⟩ was limited to the subject pronoun function because this was a stereotypical feature of Jamaican Creole as English would have *I* in these cases. However, they note that it could also be said that this variation indicates a complex relationship to English and perhaps non-Creole-speakers (Hinrichs and White-Sustaíta 2011, as cited in Eisenstein, 2015, p. 165), making it not only grammatical constrained but

[3] French spoken in France.

possibly socially motivated.

Hinrichs and White-Sustaíta's example from Jamaican Creole is somewhat similar to Tatman's (2016) report of variation between ⟨work⟩ and ⟨werk⟩ among drag queens online. She compared collocations for the two and found that ⟨werk⟩ was typically used to express approval and ⟨work⟩ to express the more typical notion of doing work, leading her to conclude that these are in fact different lexical items (Tatman, 2016, pp. 163-164). Determing whether this involved two lexical items versus one polysemous lexical item is well beyond the scope of the present study, and it would be admittedly a weak argument to claim that this variation is purely constrained by syntax, but there is enough in this example to make the more grammatical interpretation possible, wherein ⟨werk⟩ is limited to particular syntactic positions.

Clearer examples of grammatically constrained orthographic variation come from Eisenstein's (2015) work on Twitter. He examined two orthographic variables that, on the surface, would appear to have a spoken phonological connection: (ing) as ⟨ing⟩ or ⟨in⟩[4] and (th) fortition, which he refers to as *th*-stopping. In both cases, grammatical constraints were found. The variable (ing) was more likely to be realized as ⟨in⟩ for verbs (p. 176), and the variable (th) was more likely to be realized as a glyph ⟨d⟩ when the spoken form would typically be voiced as opposed to having it realized as ⟨t⟩ when the spoken form would be voiceless (Eisenstein, 2015, pp. 170-171). There were other social factors found to be constraining these variables, but there is a strong argument here that linguistics factors were important for both.

1.2.2 Connection to audience

Another constraining factor that Eisenstein (2015) found for the realization of (ing) was the audience. As was already discussed in section 1.1, the intended or expected audience for a message seems to have an impact or the character of language used on CMC just as it does in spoken language, but the examples from Pavalanathan and Eisenstein (2015) and Androutsopoulos (2008b) were related to lexical variation instead of orthographic variation. Eisenstein (2015), on the other hand, found this some relationship for orthographic variation where ⟨in⟩ was the more frequent variant for @-messages on Twitter, meaning messages that were directed at a particular user instead of posted as general public statements (p. 176).

Additionally, Androutsopoulos (2008b) did generally look for respellings in his data, as well, also finding evidence of the importance of audience. On the German hip-hop website webbeatz.de, for example, he noted more colloquial spellings (i.e., those non-standard spellings that are related to colloquial speech) than in other areas of the same site (p. 297). The implication is that discussion forums are expected to be read by the more engaged members of the community whereas general web pages are more likely to be viewed by

[4]This presumably would include ⟨in'⟩, though Eisenstein (2015) does not explicitly give that variant.

broader passers by, so to speak.

1.2.3 Connection to community membership

Perhaps the most common associations examined for linguistic variables in variationist studies are associations between variants and particular communities or particular segments of communities, and indeed, these associations exist for orthographic variation. For instance, Sebba (1998) examined spelling in British Creole and suggested that non-standard spellings are used both because they distance the language from its English superstrate and because there is no orthographic norm for British Creole. The very idea of non-standard in this case, then, is 'not as would be done for English'. Sebba (1998) provided examples such as ⟨Jameka⟩ and ⟨kool⟩ where standard English orthography would have ⟨Jamaica⟩ and ⟨cool⟩ (as cited in Androutsopoulos, 2000, p. 515). The idea is not simply to buck the superstrate but to make a statement against speakers of the superstrate,[5] to assert an independent identity from those speakers.

Androutsopoulos (2008a) also presented an example of using a particular orthographic variant on German hip-hop websites to signal membership in the particular community. In this case, ⟨z⟩ would be written where ⟨s⟩ was expected. What makes this example's importance particularly clear is that Androutsopoulos (2008a) quoted one of his informants as explicitly claiming that such ⟨z⟩ signaled an extreme dedication to hip-hop (pp. 12-13). It appears, then, that orthographic variation can be quite salient.

Eisenstein's (2015) once again proves useful in also showing community membership can be a determining factor in orthographic variation. When examining (ing), he found that tweets emanating from US counties with high population density and/or high Black populations were more likely to have tokens of ⟨in⟩ than those emanating from other counties (Eisenstein, 2015, p. 176). It is important to note that he did not know the claimed racial identities of these Twitter users, but the presumption is that they are likely to identify as Black or to at least be highly exposed to those who identify as Black in their daily lives.

Eisenstein's (2015) finding also begins to highlight the important of geographic location. Indeed, one should not confound virtual communities, those formed through consistent interaction online around shared interests (Castells 2000, as cited in Androutsopoulos, 2008b, p. 283), and physically centered communities, but this does not mean that a connection between the two is impossible. As McNeill (2018) showed for virtual communities and geographically defined communities in the Maritime Provinces of Canada, those who live near each other tend to converge online, as well (pp. 88-91).

As such, Jones (2015) found that the choice of non-standard spelling between ⟨nuttin⟩ and ⟨nun⟩ on Twitter, both attempts to represent *nothing* spoken with an intervocalic glottal stop, was constrained by geographic. Those

[5] Of course, it is possible that many British Creole speakers are also English speakers, but the presumption is that speakers identities are bound up in their language preferences.

tweeting from the north in the US preferred ⟨nuttin⟩, and those tweeting from the south in the US preferred ⟨nun⟩ (p. 424). It might thus be expected that northerners and southerners form separate virtual communities, as well, where different norms are established, as the spoken pronunciations in both areas would be the same.

There are also cases, though, where one might say that a virtual location is the constraining factor. Cherny (1995) offered somewhat anecdotal evidence of this in the early days of the internet. She analyzed the use of ⟨u⟩ and ⟨r⟩ for ⟨you⟩ and ⟨are⟩, respectively, finding that while these non-standard spellings did appear in MUDs, players of MUDs considered them to be forms that originated in IRC chatrooms (as cited in Paolillo, 1999, p. 2). If MUDs and IRC chatrooms are conceptualized as separate locations, then a important determinant here was likely virtual location.

Age also appears to be a determinant for the realization of orthographic linguistic variables. Schnoebelen (2012) analyzed the use of different emoticons, faces essentially drawn using alphanumeric characters. He looked particularly closely at those that included noses, such as ⟨:-)⟩, versus those that did not, such as ⟨:)⟩. He found that noseless emoticons were preferred more by older users of Twitter more so than younger users, though noseless emoticons were also associated with users who disregarded various orthographic standards by doing things such as repeating letters or leaving out apostophes (pp. 122-124).

Baron (2004) also analyzed emoticons in instant messaging with the addition of analyzing abbreviations and acronyms that she considered unique to CMC. She found that emoticons were almost exclusively limited to female participants in her data (pp. 415-416). This result was reproduced in Varnhagen et al.'s (2010) work on instant messaging, though the results were not as quite as lopsided as in the earlier study (pp. 728-729).

1.2.4 Connection to pragmatics

Androutsopoulos (2000) long ago acknowledged that orthographic variants can perform pragmatic work. He argued that they "signal certain attitudes or evoke certain frame of interpretation by establishing a contrast to the text's spelling regularities or to the default spelling of a linguistic item" (Androutsopoulos, 2000, p. 517). An example of such a contrast can be found in his study of German punk fanzines, essentially low budget magazines made by enthusiasts. Androutsopoulos (2000) noted that the typical spelling of the term *fanzine* was ⟨fanzine⟩, which in fact bucks against standard German spelling in which all nouns are capitalized. Nevertheless, this was the established norm within fanzines themselves. As a result, those familiar with the medium would sometimes produce hypothetical quotations from Germans who were not familiar that included the spelling ⟨Fähnziehn⟩, much more in line with standard German orthography, with the result being mockery of the latter's ignorance (Androutsopoulos, 2000, p. 526).

The example of a pragmatic factor in the realization of orthographic variables given in this section as well as the examples of other factors from the

preceding sections provide some insight into what the possible determinants variables are. Unsurprisingly, the range is as broad as that which can be found for linguistic variables in spoken language. While it will not be possible to look at each and every factor in analyzing (lol) due to the limitations of what we know about Twitter users without directly interacting with them, we will endeavor to include as many factors as possible in line with the exploratory nature of this study.

1.3 Previous work on (lol)

The orthographic variable being analyzed in the present study is (lol), which at least originally was an acronym that stood for 'laugh out loud'. This acronym is thought to have originated in English language chatrooms in the 1980s (McCulloch 2019, as cited in Schneier, 2021, p. 4). However, it has found its way into other languages, as well. Liénard (2014) documented (lol) being used by early adopters of the internet in Mayotte, an island nation near Reunion in Africa. What is notable about this case is that the internet had only effectively been accessible starting in 2012 (p. 154), and English was not a local language nor an official language in Mayotte, those being Shimaroe and Kibushi for the local languages and French for the official language (p. 158). Likewise, McNeill (2018) documented significant use of *lol* in what would otherwise be viewed as French-language tweets on Twitter. It appears possible then that (lol) has become something of an internet-language acronym rather than an English-language acronym, though almost all the work done on it has been focused on English.

Despite its penetration even into other languages, *lol* is not overall a frequent lexical item when compared to other lexical items. In Baron's (2004) IM data, *lol* made up only 0.6% of the total words. Similarly, in Tagliamonte and Denis's (2008) IM data, it made up 0.41% of the words, and 0.35% in Schneier's (2021) cell phone data. Schneier (2021) also found *lol* to be most frequent turn-initially and turn-finally (p. 14). This low overall frequency might be damning for quantitative analyses, but it is perfectly in line with Zipf's law, which claims that the most frequent lexical items in any corpus will be exponentially more frequent than those ranked even only slightly lower in frequency, and indeed *lol* is highly frequent when compared to other as far as content words go (Baron, 2004, p. 412; McNeill, 2018, p. 60), rendering it analyzable in a quantitative fashion.

1.3.1 As a lexical variable

Previous analyses of *lol* have by and large treated it as a lexical variable as opposed to an orthographic variable. While the current study aims to present an orthographic analysis, the results from lexical analyses provide some context for better understanding variant spellings. What those results repeatedly suggested was that *lol* is employed primarily for pragmatic purposes.

Baron (2004) included *lol* in her study of gender in instant messaging, though she did not draw any conclusions about it being constrained by gender. Instead, she aruged that it functioned as a "phatic filler" used to show engagement in the same way as lexical items such as "*OK*, *cool*, or *yeah*" (Baron, 2004, p. 412). Likewise, Tagliamonte and Denis (2008) described *lol* "as a signal of interlocutor involvement," specifically contrasting this interpretation with items such as *haha* and *hehe*, which they described as simply forms of laughter (p. 11). A non-pragmatic constraint was also found, however, in that *lol* was more common among young user of instant messaging than older users (Tagliamonte & Denis, 2008, p. 13).

Schneier (2021) also gave a lexical treatment of *lol*, though he did not wash *lol* of its connection to laughter as others had done. He argued instead that laughter in spoken conversation has pragmatic functions itself, specifically that it helps mitigate face threatening acts(Schneier, 2021, p. 5). Shorter keybursts for *lol* at the beginning of turns led him to this conclusion as one would want to very quickly respond when attempting to help one's interlocutor save face (Schneier, 2021, pp. 17-18). It was also argued that *lol* goes beyond the functions of laughter in spoken language in that it also helps to coordinate turn taking (Schneier, 2021, p. 5).

1.3.2 As an orthographic variable

While (lol) shows up in analyses as a lexical variable, to date, no research has analyzed what the constraints on and functions of orthographic variants of it are. That said, some work on orthographic variation in CMC has been done for other variables the results of which can be suggestive for how (lol) works.

A rather obvious difference between the written language mode and spoken language mode is that the latter has a clear prosodic dimension nor a gestural dimension in the former that can be used for pragmatic purposes such as turning an utterance into a question or expressing disbelief. However, there are arguably orthographic tools available to cover these functions such as capitalization and reduplication. This phenomenon of using the unique tools that a medium offers to cover the functions of the tools that it lacks can be referred to as paralinguistic restitution (Thurlow and Brown 2003, as cited in Schneier, 2021, p. 3).

Along the same lines, the idea of affective lengthening has also been proposed, which is essentially the reduplication of individual typed characters for pragmatic effect (Schnoebelen, 2012, pp. 117-118). Indeed, just such lengthening is readily apparent in the tweets collected for the present study as can be seen in tweet 4, though an argument for what exact function this has or whether it even represents a pragmatic function has yet to be made.

4. Thaabita el-Kamel: @jasethebell @SkyFootball looool I'll have to pass but thanks for the offer

Not all studies of orthographic variation in CMC propose that variants are linked to pragmatics. Stewart et al. (2017), for example, analyzed variation

in the spelling of a banned hashtag on Instagram that was used for posts that treated anorexia as a positive condition. Instagram users developed variants to get around the ban, and these variants became more and more disconnected from the original as measured using Levenshtein distance (Stewart et al., 2017, p. 4). It was found that newcomers to this Instagram community preferred the variants with the greatest distance from the original (Stewart et al., 2017, pp. 5-6), suggesting perhaps that the community was stratified by something akin to age and that these variants were not in contrast serving any pragmatic functions.

There is thus reason to suspect that spelling variation in (lol) may be either socially constrained or serving pragmatic functions. As a result, the primary research question of the present study is the following:

RQ What are the social constraints and/or pragmatic function of the orthographic variable (lol)?

This is naturally a rather open-ended exploratory question given the lack of research on this particular variable, meaning results will have as much precision as possible, but more research will likely be needed.

2 Methods

The corpus used for this study is a collection of tweets from Twitter. As such, this section will briefly discuss some characteristics of Twitter along with how the tweets were scraped in section 2.1. This will be followed by discussion of the information information that was coded for each token in section 2.2. As some of the social network analysis techniques involved in the coding are fairly novel in sociolinguistics, a greater amount of detail will be provided in the coding section. Finally, the statistics to be used to analyze the detail will be discussed in section 2.3.

2.1 Data collection

The social media platform Twitter offers great opportunities for the study of language but also presents its own special challenges for research. In Yates's (1996) early work on CMC, he decided to collect data from a computer conferencing system as these systems involved "open" discussions that carried fewer ethical concerns than CMC mediums that users expect to be private (p. 31). This is also true of Twitter, which allows users to make their accounts private but is by and large used as a form of open public discourse. Messages, referred to as tweets, are posted to users' timelines to be read by other users who explicitly follow the posters. At the time of the data collection for this study, tweets were limited to 140 charaters, though that number has since been increased to 280.

Tweets can also be directed at specific users, in which case this form of CMC begins to resemble face-to-face conversation in some ways, as has been

noted (e.g., Danescu-Niculescu-Mizil et al., 2011, p. 31). There are some key differences, however. First, only a quarter of Twitter users have been found to hold conversations (Java et al. 2007, as cited in Danescu-Niculescu-Mizil et al., 2011, p. 1), defined as sending directed tweets back and forth, though this proportion varies depending on the language being used (Hong et al., 2011). Second, and relatedly, conversations may or may not be synchronous, particularly in today's world where Twitter is commonly accessed via smartphones that can provide instant notifications of incoming messages wherever a user may be. Lastly, and most obviously, Twitter involves only written communication, which both removes access to some methods of expression such as gestures and prosody but also introduces new methods such as punctuation, spelling, and images.

The particular Twitter corpus used for this study was originally collected for a study looking at (lol) as a lexical variable as used in French tweets originating in the Maritime Provinces of Canada (McNeill, 2018). Tweets were collected continuously between January 8th, 2017 and February 8th, 2017 using what was at the time referred to as the spritzer level of access to Twitter's API, which allows samples to be taken from 1% of all public tweets (as opposed to gardenhose access at 10% and firehose at 100%). Collecting samples of tweets is essentially identical to using Twitter's search bar, which allows users to specify geographic regions, languages, dates, and so on. The following search string was used for the McNeill (2018) corpus:

geocode:46.0878,-64.7782,200mi exclude:retweets exclude:links

The first parameter of the search string uses latitude and longitude to target the Maritime Provinces. The second parameter excludes retweets, which involve one user repeating another users tweet in the former's own timeline and so does not count as language use by the former. The last parameter excludes tweets that contain links as these are more likely to be posts from commercial accounts as opposed to language use from regular users. Both of these exclusions are common practice when working with Twitter data (e.g., Pavalanathan & Eisenstein, 2015, p. 199).

This method of scraping Twitter for data resulted in a corpus made up of 1,274,233 tweets, by and large in English, though not all of these were used in the original study nor will they all be used in the present study. In order to limit the effects of variable audiences for tweets, this initial set of tweets was filtered so as to include only directed tweets. This increases the likelihood of the message being part of a conversational interaction and ensures that it was intended for the community to which the poster belonged.

Indeed, audience is a constraining factor in language variation on Twitter, so including tweets with different types of audiences (i.e., individuals versus the public in general) is an important step. African-American English features used by gay White men from England on Twitter were more likely to occur in tweets directed at other such users rather than in broadcast tweets (Ilbury, 2020, p. 256). The fact of being directed or not has also been found to constrain

the presence of ⟨g⟩ in the (ing) orthographic variable (Eisenstein, 2015, p. 176) as well as lexical choices (Pavalanathan & Eisenstein, 2015).

Filtering the data down to only directed tweets resulted in a corpus of 307,878 tweets, but as the focus was only on French language tweets in the original study, this number was further filtered down to include only those Twitter communities that contained French tweets, resulting in 19 communities, each with a three- or four-digit ID, and each still containing far more English tweets than French.

Community detection will be discussed in greater detail below, though what is important for the moment is that these communities contained 4,733 tokens of (lol). In this case, these were tokens of (lol) as a lexical variable and so included variants such as *rofl* 'roll on the floor laughing' and even *mdr* from French *mort de rire*, the rough equivalent of *lol*. In other words, lexical items that are not of interest in the current study were included and spelling variants of ⟨lol⟩ such as ⟨LOL⟩ or ⟨lolol⟩ were collapsed into tokens of one lexical item: *lol*. Fortunately, the original spellings were stored in the corpus, and by far the most common lexical item was *lol*, so filtering out unwanted lexical items resulted in a final corpus with 13 communities, 3,938 tokens of the orthographic variable (lol), and 83 spelling variants for (lol).

2.2 Coding

In order to perform a quantitative analysis on the use of (lol) in the present Twitter corpus, each token was coded for several different variables. The linguistic variable in this case, (lol), is made up of orthographic variants of what was originally an acronym standing for 'laugh out loud'. Despite there being 36 variants in the data, only 3 occurred more than 5 times, namely ⟨lol⟩, ⟨Lol⟩, and ⟨LOL⟩. Of special note when interpreting why particular orthographic variants appear is the presence of auto-complete systems on both smartphones and computers. For instance, the initial capital in ⟨Lol⟩ could sometimes be forced by the typing device when at the beginning of a sentence as opposed to being something the user purposely did. The use of keylogging software that tracks key presses can identify whether auto-complete was used (e.g., Schneier, 2021, p. 9), though this was not done for the present study.

While this lack is distinction between what is auto-completed and what is not could introduce a potentially significant methodological issue, there are two important reasons to believe that the issue is not a large one. First, auto-complete systems are sometimes generated by the typing habits of the user. If a user always manually types ⟨LOL⟩, their auto-complete system may very well start correcting any spelling of (lol) to the fully capitalized variant. While this means the user's likelihood of varying their spelling is somewhat diminished, the spelling that is ultimately produced is at least representative of their own personal norm. Second, the most likely candidate for auto-completion is ⟨Lol⟩ at the beginning of a sentence, but this variant is easily the least frequent of the three top variants, as will be evident in the Results section below.

Beyond the linguistic variable, there are both social and pragmatic variables for which each token of (lol) was coded. The former will be covered in section 2.2.1 and the latter in section 2.2.2.

2.2.1 Social variables

It has been well established that social variables are meaningful for language variation in CMC. For instance, Danescu-Niculescu-Mizil et al. (2011) found that Twitter users accommodate their language styles relative to their interlocutors, both symmetrically and asymmetrically (pp. 6-8). As the literature on accommodation theory suggests that accommodation is triggered not just by a need for "communicational efficiency", but also to gain social approval and maintain one's identity (Danescu-Niculescu-Mizil et al., 2011, p. 3), this suggests that social factors are unsurprisingly at work on social media.

CMC presents challenges for obtaining social descriptors for locutors in a corpus that are not present in data that was obtained through traditional sociolinguistic interviews as researchers do not always interact directly with those producing data in CMC studies, as is the case in the current study. This is particularly true when large corpora are collected, which is often the case in CMC research. For instance, Ilbury (2020) targeted openly gay men from the south of England in his study of Twitter, but he analyzed a only ten users as opposed to the 1139 users in my corpous. This small number of users allowed Ilbury to conduct a sort of virtual ethnography to establish the identities of those he analyzed even though he never interacted with them directly.

Similarly, Jones (2015) also faced the challenge of determining if the users of Twitter that he examined were indeed native speakers of African-American Vernacular English (AAVE) or if they were simply performing an identity. He dealt with this problem much like Ilbury did: through ethnography (Jones, 2015, p. 412). However, and perhaps because his data was much more large scale than Ilbury's, Jones's ethnography included experience with African-Americans in person, and his goal was not to identify African-Americans on Twitter but to identify native speakers of AAVE who may not necessarily be African-American.

Gender has also found its way into CMC research on Twitter. Bamman et al. (2014) deduced users' genders by establishing gender associations for given names using US census data in which the majority gender for a name in the census data would be taken as the typical gender for people with that name (p. 140). While most names had a clear association using this method, and while most people did not have names that were highly ambiguous, there is still a level of uncertainty with classifying users this way, just as there is for race, ethnicity, and sexual orientation. For this reason, I focus on social variables that are more directly observable on Twitter, namely community membership and centrality in a community as calculated using well established social network analysis techniques as well as geographic location as given by the user.

Concepts of community Perhaps the two most common ways to conceptualize communities in variationist research is through the concepts of speech communities and communities of practice. The former are defined primarily according to shared linguistic norms and linguistic evaluations among people in a relatively clearly delineated geographic area. This concept of community has been employed since Labov (2006) proposed it in his foundational work in sociolinguistics The latter came to prominence in the work of Eckert and gives precedence to what draws people together, that being a shared activity. With communities of practice, smaller scale communities are typically identified and so the importance of geography is not necessarily as present as it typically is for speech communities. The key is simply that there is a shared activity, which implies sharing physical space, as well, but that is no longer required as people can partake in shared activities virtually with the advent of the internet.

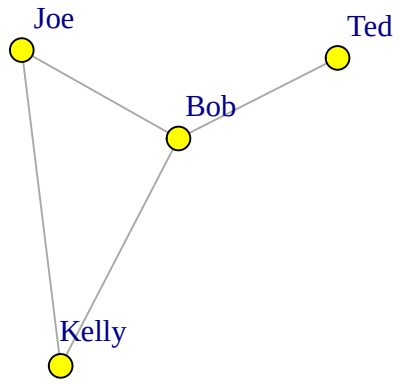
Various concepts of community have thus been used in CMC research. Castells (2000) defined communities along similar lines as communities of practice where they are "organized around a shared interest or purpose," but he did not think these were the same as face-to-face communities because of the differences in how interaction occurs. Castells (2000) referred to these communities as virtual communities (as cited in Androutsopoulos, 2008b, p. 283). Androutsopoulos (2008b), for his part, recognized that the connections between members of these communities could be quite literal through the user of hyperlinks (pp. 283-284), which is indeed a feature that is not possible in face-to-face communities.

What is at least implicit in all these concepts is the idea of interaction, which is how I define communities for the present study. Those who interact with each other more than with others may be considered as members of the same community. Formally, this can be quantified using community detection methods from social network analysis, one of which I will now describe.

Community detection One important area of development in modern social network analysis techniques that has been generally missing from sociolinguistics is community detection. Community detection is the process of using algorithms to delineate communities within a social network. In this case, a community is conceptualized as a cluster of individuals who interact with each other more than they do with others. For instance, Figure 1 shows a very simple hypothetical network consisting of Joe, Kelly, Bob, and Ted. In the network, Joe, Kelly, and Bob all know each other, but the only person who knows Ted is Bob. As a result, Joe, Kelly, and Bob likely form community of which Bob is not a part. Each of these connections is called a tie, but what may constitute a tie or how to quantify the strength of a tie is decided by the researcher. In the case of the present study, I define a tie existing betweeen two users if there is any directed tweet sent between them and the strength of that tie as the number of directed tweets between them.

At the core of any tie is the idea that two people who are tied together interact with each other on some level so that ultimately any community in

Figure 1: Simple example network



a social network is based on mutual interactions. This conceptualization of what a community is, a group of people who interact with each other, accords with conceptions of communities such as communities of practice. To share an activity together is inextricably about interacting with one another. For this reason, Schenkel et al. (2002) argued that social network analysis can be used to quantify the characteristics of communities of practice as the latter are more often identified through qualitative means. I take no stance on what draws people into their communities here, though, as is at least implicitly done with communities of practice as they suggest that the activity is what binds the community together.

There are a number of algorithms for community detection, but the one used for this study comes from Blondel et al. (2008) and is commonly referred to as the Louvain method. The general mechanics of the Louvain method involve trying out different possible community divisions and calculating the modularity Q for each pass in order to find the division that yields the highest Q. Q itself is "a measure of the quality of a particular division of a network" (Newman & Girvan, 2004).

A standard test for the reliability of community detection algorithms is to apply them to Zachary's (1977) karate club data. This data includes a karate club that dissolved into two separate clubs, thus two explicitly different communities. An algorithm which takes the same network and divides it into the same two communities is thought to be valid and reliable. Newman and Girvan (2004) evaluated Q itself for this using several Q maximization algorithms and obtained good results, suggesting that this is indeed a useful measure. For the Louvain method specifically, Waltman and Eck (2013) tested it on the karate

club data and found it to be almost perfect at finding the maximum Q possible (p. 471), suggesting that it is a reliable modularity optimization algorithm.

As for the application of the Louvain method for community detection to sociolinguistic data, this has only been done once, to my knowledge. In McNeill (2018), which used the same corpus as the present study, the lexical variable (lol) was found to be significantly constrained by the community to which one belonged. The variants in that study were English-origin and French-origin equivalents of *lol* as used within what could generally be considered French-language tweets. While there have not been other applications of community detection in sociolinguistic research, this result combined with its history of evaluation in sociology and computer science point to its validity.

The implementation of the Louvain method used here comes from the social network analysis software Gephi (Bastian et al., 2009). Gephi uses a slightly modified version of the algorithm made to handle directed networks as opposed to the original version, which handled only undirected networks. This was more appropriate for the Twitter corpus used here as users do not always respond to directed tweets, making some ties asymmetric. The resolution option for the algorithm was kept at the default of 1.

Initially, 8,945 communities were detected in the data, but as the goal at the time of the original study involved looking at those who might be considered French speakers specifically, only communities which contained tweets with French in them were analyzed, resulting in 19 communities, each with a three- or four-digit ID.

These 19 communities contained 4,733 tokens of (lol). In this case, these were tokens of (lol) as a lexical variable and so included variants such as *rofl* 'roll on the floor laughing' and even *mdr* from French *mort de rire*, the rough equivalent of *lol*. In other words, lexical items that are not of interest in the current study were included and spelling variants of ⟨lol⟩ such as ⟨LOL⟩ or ⟨lolol⟩ were collapsed into tokens of one lexical item: *lol*. Fortunately, the original spellings were stored in the corpus, and by far the most common lexical item was *lol*, so filtering out unwanted lexical items resulting in a corpus with 13 communities, 3,938 tokens of the orthographic variable (lol), and 83 spelling variants for (lol).

Centrality measures In social network analysis, a centrality measure is a measure of a person's position within a given community. In the perhaps more familiar terms of communities of practice, this is somewhat similar to deciding which members are core members and which are peripheral members. However, centrality measures are always quantitative, as the name implies.

CMC research has included centrality measures at least as far back as Paolillo (1999) in his study of an IRC community. They have also been used in analyzing language on German hip-hop web sites (Androutsopoulos, 2008b) and on Twitter using follower count (Danescu-Niculescu-Mizil et al., 2011). Centrality has often, though not always, been found to be significant in these studies.

Likewise, centrality measures have been used in language variation studies since Milroy (1987). Just as Milroy (1987) did, the implementation in sociolinguistics tends to involve the use of an index with a relatively small scale of possible values, such as 5. Part of the reason for this approach is that it can be exceedingly difficult to track face-to-face interactions, which is conversely not a problem at all with Twitter data as directed tweets are explicit.

There are many other centrality measures, as well, though the one used in this study is PageRank (Brin & Page, 1998), which was calculated for each user in the data relative to the community of which they were a member. PageRank was originally developed for ordering search engine results and ultimately led to the creation of Google. Equation 1 shows how page A's PageRank PR, or in this case person A's PR, is calculated.

$$PR(A) = (1 - d) + d\left(\frac{PR(T1)}{C(T1)} + ... + \frac{PR(Tn)}{C(Tn)}\right) \quad (1)$$

Here, d is a damping factor between zero and one, Tn is a page that links to page A, and $C(Tn)$ is the total pages linked to by page Tn. This effectively makes PageRank a function of the number of pages that link to the page of interest as well as the PageRanks of those pages. As a result, a user who directs many tweets to other members of their community but who receives very few tweets from other members will not have a particularly high PageRank. The intuition is that it is not difficult to talk a lot, but it is difficult to get people to care enough about what a person thinks to bother talking to that person.

Geographic location Geographic location has also proved to be a meaningful social variable for language variation on Twitter, despite Twitter not be regionally segregated. It is possible to for users to search for tweets that are emanating from a particular physical area, but this is not the default setting nor do a user's followed accounts necessarily come from their own region. However, geographic variation has been found at least for AAVE features (Eisenstein, 2013; Jones, 2015) and lexical variables (Huang et al., 2016). Part of this importance may be due to a propensity for Twitter users to form virtual communities with those from the same regions despite this not being necessary, as there as at least some evidence that this sort of agglomeration happens (McNeill, 2018, pp. 88-91).

Geographic location is also a social variable that is fairly easy to obtain for users of Twitter, though there are some caveats. The simplest way to get this information, and what is was done for the present study, is to use the location that each user entered manually in their profile. This is often available and also returned by the Twitter API when tweets are collected. The downside to this is naturally that users can enter any location that they want regardless of accuracy, thus an assumption that only a small number of users enter inaccurate locations is required.

An alternative approach to finding the geographic locations of users is to use geotags. Twitter users can turn on this feature so that, when they send a

tweet, the exact location they sent it from will be stored as metadata with said tweet. However, this feature is rarely used. Jones (2015) found only 150 to 800 geotagged tweets per lexical item in his study, which accounted for between 2.5% and 7.0% of the tweets containing those lexical items (p. 407). Using this method calls for an API access level that makes amassing extremely large corpora possible, which was out of reach for the current study. For example, Huang et al. (2016) used geotagging but were also able to collect 924 million such tweets over the course of a year with the access level that they had (p. 244). As a result, manually entered geographic locations are used for the present study.

2.2.2 Pragmatic variables

As was discussed in section 1.3, a repeated argument for (lol) as a lexical variable and orthographic variables in general is that they are used for pragmatic effects. One of the difficulties for performing the sort of discourse analyses that could uncover such effects is that long and/or repeated conversations between pairs of individuals on Twitter are not easily obtained. Danescu-Niculescu-Mizil et al. (2011) solved this problem by identifying pairs of users who were likely to converse often and mining each of their entire Twitter histories. With both histories at hand, it was possible to reconstruct repeated, long conversations between the pairs (Danescu-Niculescu-Mizil et al., 2011, p. 3). Such a solution was not achievable given the resources of the current study, so I chose not to perform an exhaustive discourse analysis for pragmatic factors.

What can be analyzed in a quantitative fashion that could also shed light on some pragmatic factors that are linked to the orthographic variation of (lol) is the sentiment of each turn. In this case, a turn is conceived of as a single tweet, which may or may not be multiple sentences but is always limited to 140 characters. The sentiment classifier R package sentimentr (Rinker, 2019) was used to calculate the polarity sentiment of each turn. No preprosessing of the corpus was done before sending it to this classifier other than removing *lol* from the dictionary that the classifier uses.

2.3 Statistics

The data analyzed in this study is all categorical, and the typical descriptive statistics for categorical data are used. One such statistic that is worth discussing as it does not appear in sociolinguistic research a great deal is the measure of dispersion for variants of (lol) for either an individual or a community. The Simpson diversity index D (Simpson, 1949), as described in equation 2, is used as a measure of stability in the sense that a small dispersion can be interpreted as a consistent preference for a particular variant whereas a large dispersion can be interpreted as a lack of clear preference for any particular variant. This ia rather novel use for D in variationist studies where it otherwise shows up as a measure of language ecology (e.g., Greenberg, 1956) or as

a measure of the diversity of interactions one has (e.g., Sharma, 2011).

$$D = 1 - \sum_{i=1}^{R} p_i^2 \tag{2}$$

In equation 2, p is the relative frequency of a variant i of the variable in question. Essentially, the few variants included and the greater the frequency of the mode relative to the other variants, the lower D will be. One can imagine a uniform distribution as having a very high D and a strongly unimodal distribution having a very low D.

3 Results

The summary of the characteristics for each community, shown in Table 1, do not reveal much. The mode for every community is ⟨lol⟩ except community 2265 with all uppercase ⟨LOL⟩ as the mode. What this does suggest is that even if a user is new to one of these communities, there is a good chance that they came from a community where the orthographic norm for (lol) was also ⟨lol⟩, just as in most of these communities.

As for diversity, the only communities to be particularly consistent are communities 799 and 2067 which each use the ⟨lol⟩ spelling at all times. However, (lol) was rarely used in these two communities as only 33 and 44 members produced (lol) at all in each, respectively. Overall, there was a median of 0.45 for the diversity of variants of (lol) used by the communities.

Table 1: Summary statistics for each community

Community	Mode	Diversity	Members
173	lol	0.45	2480
302	lol	0.42	17279
572	lol	0.45	3601
756	lol	0.41	980
799	lol	0	33
1032	lol	0.61	22531
1097	lol	0.53	2955
1227	lol	0.57	2214
1291	lol	0.49	1073
1917	lol	0.44	4432
2067	lol	0	44
2265	LOL	0.71	242
6817	lol	0.52	592

Taking a look at how the diversity for any given user relates to their PageRank within their community, there appears to be no pattern at all. The result is not even monotonic let alone linear. One might ask, though, if there are

Table 2: Summary statistics for each province

Province	Mode	Diversity	Members using (lol)
England	lol	0	1
Auckland	lol	0	2
North Brabant	LOL	0	1
California	lol	0	4
Prince Edward Island	lol	0.68	30
US Virgin Islands	LOL	0	1
Maine	lol	0.51	7
New Jersey	lol	0	2
New Brunswick	lol	0.51	195
Nova Scotia	lol	0.56	403
Ontario	lol	0.67	2
Provence-Alps-French Riviera	lol	0	3
Quebec	lol	0.56	4
Wairarapa	LOL	0.59	3

a large number of users who produced very few tokens and so only have low diversity due to this fact. The mean number of tokens of (lol) produced by users is 3.46, and the median is 1, suggesting that indeed most users produced very few tokens. Filtering out users who produced fewer than 10 tokens yields 1,769 from 82 users. The resulting plot can be seen in Figure 2, though just as when all users were included, it appears that there is no relationship between the diversity of (lol) and PageRank.

The lack of any clear relationship overall suggests that at least for the spelling of (lol), users are not reacting to the norms of their communities, or possibly there is simply nothing to which they could react given that the mode of each was ⟨lol⟩ except for community 2265, which had ⟨LOL⟩ as its mode instead. Table 3 provides summary statistics for each user in this community. What is immediately apparent is that the mode for this community stems mostly from Rithanya's linguistic behavior in that they produced far more tokens of (lol) than anyone else and were also quite consistent in their spelling with a diversity of 0. Rithanya's place in this community is fairly central, as well, with a PageRank of 0.0039, which despite the narrow range of PageRanks, is one of the highest among any user analyzed in this corpus for any community.

Digging deeper into the results for individual users such as Rithanya who have relatively high PageRanks as well as those who have particularly low PageRanks provides some interesting nuances. Table 4 gives the summary statistics for the users with the highest and lowest PageRanks among those who produced a reasonable number of tokens of (lol). In each of these cases, each user, regardless of their position in the community, has less diversity in their realizations of (lol) than their community does as a whole. This is not surprising as a community as will necessarily always have realized more vari-

Figure 2: Diversity according to PageRank for users who produced at least 10 tokens of (lol)

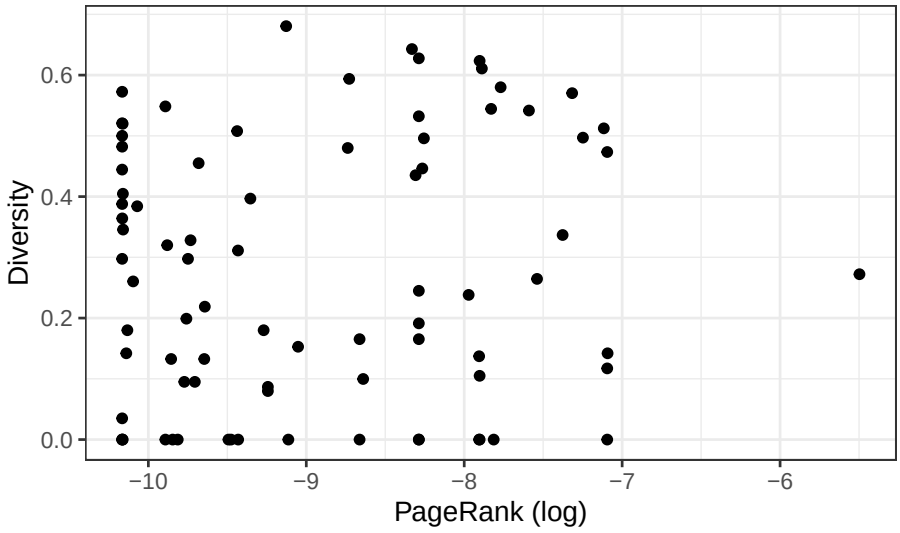


Figure 3: Diversity according to PageRank percentile within their communities for users who produced at least 10 tokens of (lol)

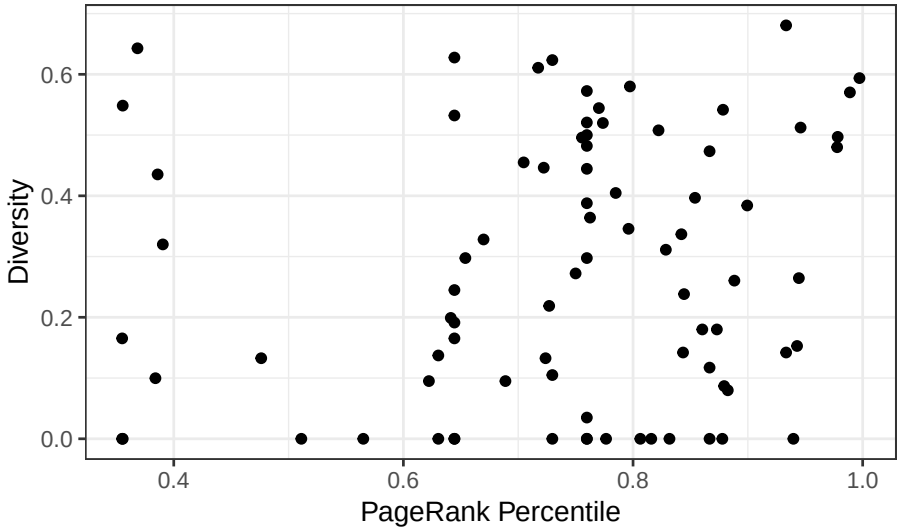


Table 3: Summary statistics for community 2265

User	PageRank	Mode	Diversity	Tokens
Aarti Nguyen	0.0038	lol	0	4
David Johnson	0.0039	lol	0	4
Fernando Long	0.0041	LOL	0.44	3
Hasana al-Irani	0.0038	Lool	0.5	2
Jesus Ibarra	0.0041	LOL	0.27	13
Kevin Rae	0.004	trololol	0.75	4
Naaif al-Shaheed	0.0038	lol	0	1
Randa el-Uddin	0.0058	lol	0	1

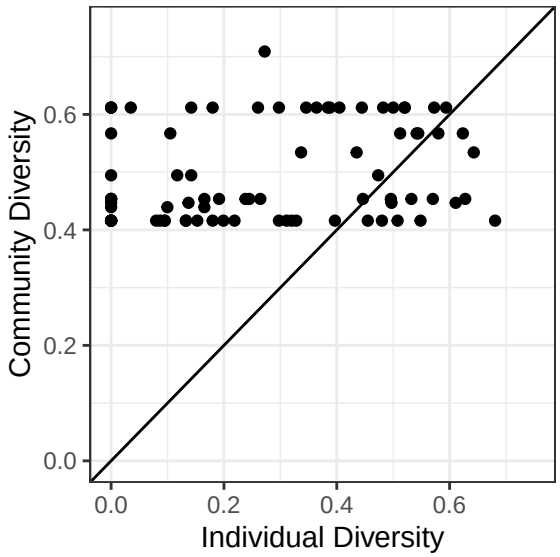
ants of the linguistic variable than any individual member of that community, but overall, Figure 4 demonstrates that some individuals do in fact show much more diversity than their community as a whole. If that were not the case, all points would be above the line going through the center of the graph. This must come down to the lack of a clear mode in among their realizations as they will never have produced more variants than the community. More on point, this fact seems to have nothing to do with their position in the community as none of the users in Table 4 outdue the diversity of their community.

Table 4: Summary statistics for users with the highest and lowest PageRanks among those with at least 10 tokens of (lol)

User	PageRank	Mode	D	Tokens	Community	D
Highest PageRanks						
Yadira Sandoval	0.00025	lol	0	1	572	0.45
Savannah Ortiz	0.00037	lol	0	1	1227	0.57
Sasha Tram	0.00043	lol	0.44	3	1227	0.57
Savanah Clark	0.00183	lol	0	5	6817	0.52
Shadhaa al-Salahuddin	0.00083	lol	0.12	16	1291	0.49
Lowest PageRanks						
Kristie Forsythe	0.00022	lol	0	1	302	0.42
Manaara el-Mansur	0.0002	LOL	0	1	1917	0.44
Michaela Auyeung	0.00008	lol	0	1	302	0.42
Michelle Gallegos	0.00025	lol	0	1	572	0.45
Miqdaam el-Salaam	0.00005	lol	0	1	302	0.42

In the case of the user with the highest PageRank, Rithanya, they basically make up their whole community in terms of (lol) usage, but for the others who do not have ⟨lol⟩ as their mode, a comparison of the distribution of (lol) between them and their communities can be seen in Figures 5, 6, and 7. Other than community 2265 where Rithanya made up the bulk of the total tokens, the

Figure 4: Diversity of each user's realization of (lol) by the diversity of their community as a whole among those with at least 10 tokens



distribution for communities 1291 and 1032 are quite similar: ⟨lol⟩ is the clear mode, ⟨LOL⟩ is still pretty frequent, and ⟨Lol⟩ has an appreciable frequency but is also the least frequent. There were, of course, many other variants produced, but following Zipf's law, no other variant had more than five tokens out of all tokens by users who produced more than 10 tokens of (lol), and most others had only one, so they were excluded from these charts to make the charts more readable.

In community 1291, Amair, Rheya, and Seprina look much like the community as a whole in their modes for (lol), but Saadiya does not follow the same pattern. Saadiya produced only ⟨LOL⟩ despite being central to this community. Likewise, Leyann and Yogi among the most marginal members in community 1032[6] also clearly favored ⟨LOL⟩ unlike the community as a whole. Finally, Jocques was the only shown here who had ⟨Lol⟩ for a mode. However, Jocque only used this variant at the very beginning of a tweet or after punctuation, meaning that it was always the first word of what might be orthographically considered the beginning of a sentence. In the case of typing on a mobile device, it is possible that many of the cases of ⟨Lol⟩ come from typical auto-correct systems that might capitalize the first word of every new sentence, though this is not the case for Jocques as they tweeted from the web. Clearly, it is just as possible for central members to have modes that differ from their community

[6]The five marginal members listed in Figure 7 are not in fact the most marginal in the network as there were 28 users with the same PageRank. They instead form a sample of the most marginal members.

Figure 5: The distribution of (lol) for the communities of those with the highest PageRanks and lowest

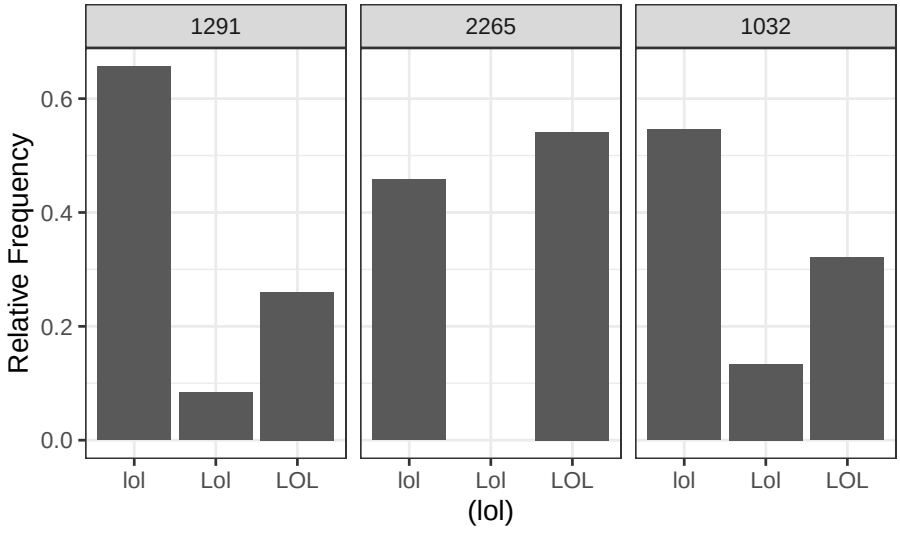


Figure 6: The distribution of (lol) for users with the highest PageRanks

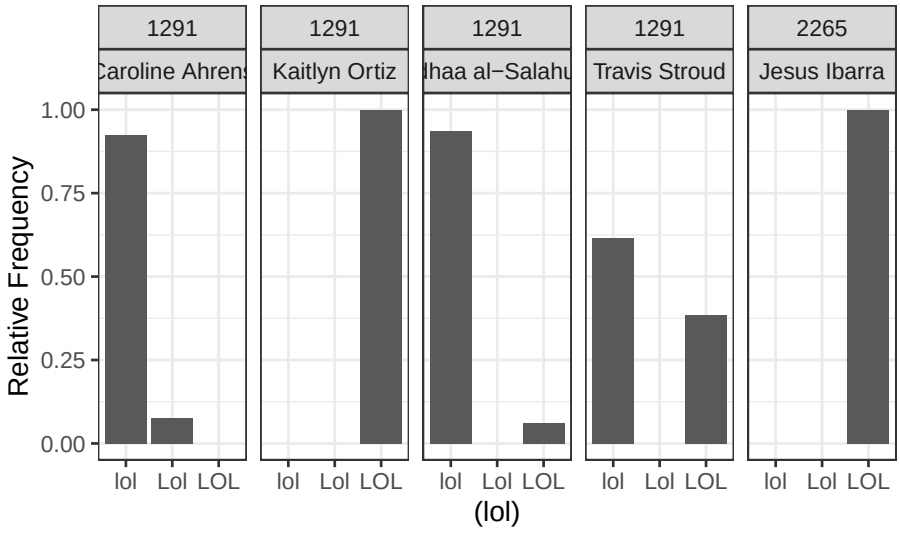
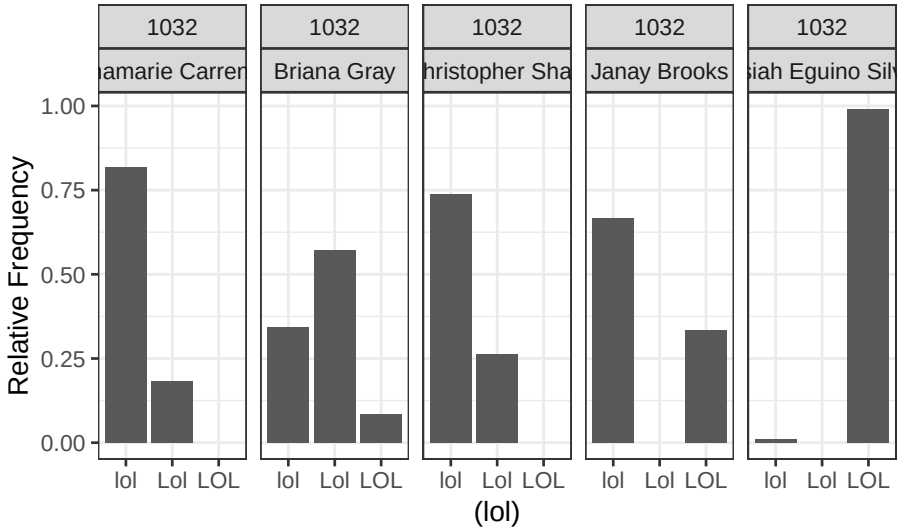


Figure 7: The distribution of (lol) for users with the lowest PageRanks



as it is for marginal members.

Outside of the modes for these users, the general shapes of the distributions for the central as well as for the marginal members were never just like that of the community except in the case of Rithanya. For instance, Dellanira and Kentoria, who had the same mode as the community in having ⟨lol⟩, did not use ⟨LOL⟩ at all let alone as their second most frequent variant. Likewise, among central members, Amair produced no tokens of ⟨LOL⟩ but did produce some ⟨Lol⟩, and Seprina use ⟨lol⟩ almost to the complete exclusion of ⟨LOL⟩ and ⟨Lol⟩, which both had higher relative frequencies for community 1291. Where Figure 4 showed that not all individual users have less diversity for (lol) than their community. Some do show less diversity, of course, but regardless, this comparison of distributions for central and marginal members suggests the shapes of those distributions do may not often line up with the community regardless of social position.

4 Discussion

References

Androutsopoulos, J. (2000). Non-standard spellings in media texts: The case of German fanzines [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9481.00128]. *Journal of Sociolinguistics*, *4*(4), 514–533. https://doi.org/https://doi.org/10.1111/1467-9481.00128

Androutsopoulos, J. (2008a). Potentials and Limitations of Discourse-Centred Online Ethnography. *Language@Internet*, *5*(8), 1–20. http://www.languageatinternet.org/articles/2008/1610

Androutsopoulos, J. (2008b). Style online: Doing hip-hop on the German-speaking Web. In P. Auer (Ed.), *Style and Social Identities: Alternative Approaches to Linguistic Heterogeneity* (pp. 279–317). De Gruyter, Inc. Retrieved February 7, 2019, from http://ebookcentral.proquest.com/lib/ugalib/detail.action?docID=364724

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160. https://doi.org/10.1111/josl.12080

Baron, N. S. (2004). See You Online: Gender Issues in College Student Use of Instant Messaging. *Journal of Language and Social Psychology*, *23*(4), 397–423. https://doi.org/10.1177/0261927X04269585

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the Third International ICWSM Conference*, 361–362. Retrieved November 2, 2017, from https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204. https://doi.org/10.1017/s004740450001037x

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks [arXiv: 0803.0476]. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Cherny, L. (1996). The MUD register: Conversational modes of action in a text-based virtual reality., 1. Retrieved February 12, 2021, from https://www.elibrary.ru/item.asp?id=5621106

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: Linguistic style accommodation in social media, 745–754. https://doi.org/10.1145/1963405.1963509

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, *41*(1), 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828

Eisenstein, J. (2013). Phonological Factors in Social Media Writing. *Proceedings of the Workshop on Language in Social Media (LASM 2013)*, 11–19.

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, *19*(2), 161–188. https://doi.org/10.1111/josl.12119

Greenberg, J. H. (1956). The Measurement of Linguistic Diversity. *Language*, *32*(1), 109–115. https://doi.org/10.2307/410659

Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters in Twitter: A Large Scale Study. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 4.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, *59*, 244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Ilbury, C. (2020). "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, *24*(2), 245–264. https://doi.org/10.1111/josl.12366

Jones, T. (2015). Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech*, *90*(4), 403–440. https://doi.org/10.1215/00031283-3442117

Kim, S., Weber, I., Wei, L., & Oh, A. (2014). Sociolinguistic Analysis of Twitter in Multilingual Societies. *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, 243–248. https://doi.org/10.1145/2631775.2631824

Labov, W. (2006). *The Social Stratification of English in New York City* (2nd) [original-date: 1966]. Cambridge University Press.

Liénard, F. (2014). Les communautés sociolinguistiques virtuelles. Le cas des pratiques scripturales numériques synchrones et asynchrones mahoraises. *Studii de lingvistică*, *4*, 145–163.

McNeill, J. (2018). *LOL sur Twitter: Une approche du contact de langues et de la variation par l'analyse des réseaux sociaux* (Master's thesis). Université du Québec à Montréal. Montreal, QC. Retrieved December 8, 2018, from https://archipel.uqam.ca/11948/

Milroy, L. (1987). *Language and Social Networks* (2nd) [original-date: 1980]. Blackwell.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks [arXiv: cond-mat/0308217]. *Physical Review E*, *69*(2), 1–16. https://doi.org/10.1103/PhysRevE.69.026113

Paolillo, J. C. (1999). The virtual speech community: Social network and language variation on IRC. *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, *Track2*, 10 pp.–-. https://doi.org/10.1109/HICSS.1999.772680

Paolillo, J. C. (2001). Language variation on Internet Relay Chat: A social network approach. *Journal of Sociolinguistics*, *5*(2), 180–213. https://doi.org/10.1111/1467-9481.00147

Pavalanathan, U., & Eisenstein, J. (2015). Audience-Modulated Variation in Online Social Media. *American Speech*, *90*(2), 187–213. https://doi.org/10.1215/00031283-3130324

Rinker, T. (2019). Sentimentr: Calculate Text Polarity Sentiment [original-date: 2015-08-16T00:55:55Z]. Retrieved May 17, 2021, from https://github.com/trinker/sentimentr

Schenkel, A., Teigland, R., & Borgatti, S. P. (2002). Theorizing Structural Properties of Communities of Practice: A Social Network Approach. *Communities of Practice or Communities of Discipline: Managing Deviations at the Oresund Bridge* (pp. 1–31). The Economics Research Institute, Stockholm School of Economics.

Schneier, J. (2021). Digital Articulation: Examining Text-Based Linguistic Performances in Mobile Communication Through Keystroke-Logging Analysis [Publisher: Frontiers]. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.539920

Schnoebelen, T. (2012). Do You Smile with Your Nose? Stylistic Variation in Twitter Emoticons. *University of Pennsylvania Working Papers in Linguistics*, *18*(2), 116–125. https://repository.upenn.edu/pwpl/vol18/iss2/14

Sharma, D. (2011). Style repertoire and social change in British Asian English. *Journal of Sociolinguistics*, *15*(4), 464–492. https://doi.org/10.1111/j.1467-9841.2011.00503.x

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, *163*(4148), 688. https://doi.org/10.1038/163688a0

Stewart, I., Chancellor, S., De Choudhury, M., & Eisenstein, J. (2017). #Anorexia, #anarexia, #anarexyia: Characterizing online community practices with orthographic variation. *2017 IEEE International Conference on Big Data (Big Data)*, 1–9. https://doi.org/10.1109/BigData.2017.8258465

Tagliamonte, S. A., & Denis, D. (2008). Linguistic Ruin? Lol! Instant Messaging and Teen Language. *American Speech*, *83*(1), 3–34. https://doi.org/10.1215/00031283-2008-001

Tatman, R. (2016). "I'm a spawts guay": Comparing the Use of Sociophonetic Variables in Speech and Twitter. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 161–170. http://repository.upenn.edu/pwpl/vol22/iss2/18

Varnhagen, C. K., McFall, G. P., Pugh, N., Routledge, L., Sumida-MacDonald, H., & Kwong, T. E. (2010). Lol: New language and spelling in instant messaging. *Reading and Writing*, *23*(6), 719–733. https://doi.org/10.1007/s11145-009-9181-y

Waltman, L., & Eck, N. J. v. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 471. https://doi.org/10.1140/epjb/e2013-40829-0

Yates, S. J. (1996). Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. In S. C. Herring (Ed.), *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives* (pp. 29–46). John Benjamins Publishing Company. Retrieved February 11, 2021, from http://ebookcentral.proquest.com/lib/ugalib/detail.action?docID=680383

Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, *33*(4), 452–473. https://doi.org/10.1086/jar.33.4.3629752