Orthographic variation of (lol)*

Joshua McNeill

July 16, 2021

1 Introduction

Studies of orthographic variation in sociolinguistics have been more widespread since the adoption of the internet as computer-mediated communication (CMC) presents a wealth of easily accessible data. Additionally, CMC provides 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, leading to more casual language use. 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!

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

*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. The pronouns *they/them/their* are used in references in light of the indeterminacy of users' genders.

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, I 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. In part, the focus on (lol) stems from its appearance in the literature, where it is commonly argued to be a "phatic filler" (Baron, 2004, p. 412) or "a signal of interlocutor involvement" (Tagliamonte & Denis, 2008, p. 11). Such interpretations appear more valid on the surface than social or other pragmatic interpretations, but the latter interpretations have yet to be systematically ruled out.

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) specifically, 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, such as with studies of multi-user dungeons (MUDS) (Cherny, 1996), which are online, text-based, multiplayer games, language in computer conferencing systems (Yates, 1996), which are perhaps best thought of as progenitors of discussion forums, and chatrooms on internet relay chat (IRC) (Paolillo, 1999). 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). Text messaging on cell phones as well as language used on any communication application on a cell phone has also been examined (Schneier, 2021).

The goals of these researchers of CMC have been described as fitting into two separate "waves" (Androutsopoulos, 2008a).[2] The first wave focused on the impact of the constraints placed on users of different CMC mediums on

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

the language they produce, and the second wave focused more on analyses of what happens in CMC pragmatically and sociolinguistically (Androutsopoulos, 2008a, pp. 1-2), though one might also add that 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 understanding how results from different CMC studies relate has been a CMC medium typology developed during the first wave. In order to compare results from a study centered on Twitter, as is being done here, with results from previous work, it is important to have a framework for classifying different mediums, as CMC is not identical to face-to-face conversation Paolillo (1999, p. 1). It has also been suggested that CMC occurs in the absence of any field other than the text itself (Yates, 1996, 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.

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. Likewise, IM users have been found to multitask while connected and so did not always respond immediately (Baron, 2004, 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 also quite normal. This same difficulty in categorization is present on Twitter, where the most common sort of messages may be one-to-many, but there also exist public directed messages and private directed messages.

Again, the parameter here is still useful, but in this case, there is also an alternative: the audience design framework (Bell, 1984). Indeed, the audience design framework has already been employed to analyze Twitter (Pavalanathan & Eisenstein, 2015), yielding interesting results, as will be discussed further in the Methods section. Likewise, the audience design framework may explain why more "hip-hop slang" was observed on German hip-hop discussion forums than on German hip-hop homepages and online magazines (Androutsopoulos, 2008b, p. 293). Both of these venues are indeed one-to-many when approached from the traditional CMC 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 (Paolillo, 1999), but that it is also different from older forms of writing. For example, instant messaging has been described as a "hybrid register", somewhere between speech and older forms of writing (Tagliamonte & Denis, 2008, p. 5). Part of this hybridity perhaps stems from CMC taking place outside of the auspices of traditional social institutions, such as schools and book publishers. CMC may still be subject to certain types of overt social control on what is acceptable, such as being banned from a social media platform, but the ramifications for an individual are not as high as when schools or publishers exert overt social control. This makes CMC a fruitful area for locutors to creatively manipulate linguistic features to various ends, one of those features being orthography. It is thus useful to review three broad categories of functions studied in the literature on orthographic variation to better understand what functions and associations it has. The three categories of functions are those related to 1) audience, 2) community or subsets of a community, and 3) 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. Scores on spelling tests have been compared to participants' use of non-standard spellings in IM where no relationship could be found between the two, suggesting that non-standard spelling is not the result of a poor grasp of standard spelling (Varnhagen et al., 2010). In fact, it was 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 audience

One constraining factor for orthographic variation that has been found for the realization of (ing) as ⟨ing⟩ or ⟨in⟩ is the audience (Eisenstein, 2015). As was already discussed in section 1.1, the intended or expected audience for a message seems to have an impact on 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 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, in an examination of respellings, evidence of the importance of audience was found. On the German hip-hop website webbeatz.de, for example, more colloquial spellings (i.e., those non-standard spellings that are related to colloquial speech) were noted than in other areas of the same site (Androutsopoulos, 2008b, 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 a broader selection of passers by, so to speak.

1.2.2 Connection to community membership

While studies on the effects of audience on variation are relatively infrequent, a more common association examined for linguistic variables in variationist studies is between variants and particular communities or particular segments of communities, and indeed, such associations exist for orthographic variation, as well. At times, these associations even receives explicit commentary from those familiar with the communities. One example involves the use of ⟨z⟩ where ⟨s⟩ is expected on German hip-hop websites to signal what one informant described as extreme dedication to hip-hop (Androutsopoulos, 2008a, pp. 12-13). It appears, then, that the connection between orthographic variation and community membership can be quite salient.

Another instance where community membership is relevant also brings class conflict into the picture. It has been suggested that non-standard spellings in British Creole 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', with examples provided such as ⟨Jameka⟩ and ⟨kool⟩ where standard English orthography would have ⟨Jamaica⟩ and ⟨cool⟩ (Sebba 1998, 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 and assert

an independent identity from those speakers.[3]

Membership to a particular race, to the extent that this can be considered community membership, can also be a determining factor in orthographic variation. When examining (ing), it was 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 the claimed racial identities of these Twitter users were not known, 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 based on where they are located.

This research also highlighed the importance of geographic location. It is important not to 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 does not exist. Indeed, it has been shown for virtual communities and geographically defined communities in the Maritime Provinces of Canada, that those who live near each other tend to converge online, as well (McNeill, 2018, pp. 88-91). This lends validity to other studies that have aimed to map dialect regions using Twitter data. One such study found that the choice of non-standard spelling between ⟨nuttin⟩ and ⟨nun⟩ on Twitter, both attempts at representing *nothing* spoken with an intervocalic glottal stop, was constrained by geography. Those tweeting from the north in the US preferred ⟨nuttin⟩, and those tweeting from the south in the US preferred ⟨nun⟩ (Jones, 2015, p. 424).

Membership to a race or to a geographic location are both significant for some orthographic variation, but virtual communities may themselves be conditioning factors. A somewhat anecdotal example of this was reported in the early days of the internet. The use of ⟨u⟩ and ⟨r⟩ for ⟨you⟩ and ⟨are⟩, respectively, were analyzed in MUDs, finding that while these non-standard spellings did appear in MUDs, players considered them to be forms that originated in IRC chatrooms (Cherny 1995, as cited in Paolillo, 1999, p. 2). If MUDs and IRC chatrooms are conceptualized as separate locations, then a important determinant here was likely membership in virtual communities.

Age and gender also appear to be a determinants for the realization of orthographic linguistic variables. In an analysis of different emoticons, faces essentially drawn using alphanumeric characters, noseless emoticons were preferred over noseful emoticons (e.g., ⟨:)⟩ vs ⟨:-)⟩) by younger users of Twitter, whereas older users preferred noseful emoticons (Schnoebelen, 2012, pp. 122-124). Another study of emoticons, though this time in instant messaging, found that emoticons were almost exclusively limited to female participants (Baron, 2004, pp. 415-416), a result that was later reproduced though with the difference being less extreme (Varnhagen et al., 2010, pp. 728-729). Again, to the extent that age brackets and genders constitute communities to which one may

[3]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.

have membership, community is repeatedly found to be an important factor.

1.2.3 Connection to pragmatics

It has long been acknowledged that orthographic variants can perform pragmatic work. For instance, Androutsopoulos (2000) argued that they "signal certain attitudes or evoke certain frames of interpretation by establishing a contrast to the text's spelling regularities or to the default spelling of a linguistic item" (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. One case involved the typical spelling of the term *fanzine* as ⟨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 for 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 can be known about Twitter users without directly interacting with them, I 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. For instance, it has been documented as 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 (Liénard, 2014, p. 158). Likewise, significant use of *lol* has been documented in what would otherwise be viewed as French-language tweets on Twitter (McNeill, 2018). 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.

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, meaning it was typed quickly, led him to this conclusion as one would want to rapidly 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 in CMC (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 former has no clear prosodic dimension nor a gestural dimension 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 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

There is thus reason to suspect that spelling variation in (lol) may be pragmatically constrained, as was just discussed, or socially constrained, as discussed in sections 1.2.1 and 1.2.2. 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)?

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 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, some statistics used 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 (Rosen, 2017).

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 (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 to obtain tweets:

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

The first parameter of the search string uses latitude, longitude, and radius 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 (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 as has been found in the language use of gay White men from England (Ilbury, 2020), the presence of ⟨g⟩ in the (ing) orthographic variable (Eisenstein, 2015), and in lexical choices generally (Pavalanathan & Eisenstein, 2015), making this an important factor to control.

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 alternate lexical 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 originally included. Filtering out these 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 and auto-correct 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 (Schneier, 2021, p. 9), though this was not done for the present study.

While this lack of distinction between what is auto-completed or -corrected 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 and -correct systems are sometimes generated by the typing habits of the user. If a user always manually types ⟨LOL⟩, their auto-correcting 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-correction is ⟨Lol⟩ at the beginning of a sentence as this would be in adhering to the capitalization rules of standard English orthography, 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, it has been found that Twitter users accommodate their language styles relative to their interlocutors, both symmetrically and asymmetrically (Danescu-Niculescu-Mizil et al., 2011, 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 only ten users as opposed to the 1,139 users in the corpus used here. 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) faced the challenge of determining if the users of Twitter that he examined were indeed native speakers of African-American Language (AAL)[4] 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 AAL even if not African-American.

Another creative technique for gathering social details of users on Twitter came from a study of gender. 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 (1966/2006) 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

[4]Ilbury uses the term African-American Vernacular English (AAVE), but I use AAL to match current practices.

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.

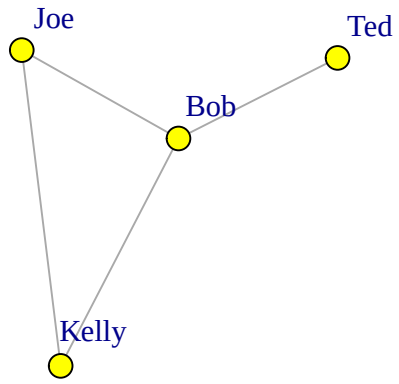
What is at least implicit in all these concepts is the idea of interaction, which is how I define communities for the present study. Those who interact with each other more than with others may be considered as members of the same community. Formally, this can be quantified using community detection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

Figure 1: Simple example network



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 originally included. These unwanted lexical items were filtered out of the corpus for the present analysis, 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 (1980/1987). Just as Milroy (1980/1987) 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 to respond.

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 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 AAL 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 is 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 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* in general is that it is 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, 2015/2019) was used to calculate the polarity sentiment of each turn. No preprosessing of the corpus was done before sending it to this classifier, though the dictionary used by the classifier was checked to ensure that it did not contain *lol* itself as a lexical item.

2.3 Statistics

The data analyzed in this study is all categorical, and the typical descriptive statistics for categorical data were 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 is a rather novel use for D in variationist studies where it otherwise shows up as a measure of language ecology (Greenberg, 1956) or as a measure of the diversity of interactions one has (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 fewer 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 as having a very low D.

3 Results

The research question for this study asked what the social and/or pragmatic constraints are for the realization of the orthographic variable (lol). In terms of social constraints, section 3.1 specifically looks at how detected communities differ, how provinces differ, and how a Twitter user's centrality within their detected community impacts their realizations of (lol). For pragmatic constraints, section 3.2 looks into possible relationships between (lol) and the sentiment of a tweet.

3.1 Social constraints

The summary of the characteristics for each community, shown in Table 1 reveal a general lack of variation in the mode of each community but internal variation in the variants used. The mode for every community is ⟨lol⟩ except community 2265 with all uppercase ⟨LOL⟩ as the mode. The diversity measures, however, suggest that ⟨lol⟩ is far from being the exclusive variants used, with a median diversity of 0.45 among the communities. The only communities to be highly consistent in using a single variant are communities 799 and 2067 which each use the ⟨lol⟩ spelling at all times. However, (lol) was rarely used in these two communities to begin with as only 1 and 2 members produced (lol) at all in each, respectively. Communities 799 and 2067 can therefore not be taken as evidence that (lol) would be invariant even within said communities save for the unlikely scenario where a larger sample would not uncover more tokens.

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

The only community to not have ⟨lol⟩ as its mode was community 2265, which had ⟨LOL⟩ as its mode instead. Table 2 provides summary statistics

for each user in this community. What is immediately apparent is that the mode for this community stems mostly from Jesus Ibarra'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. It is not clear what this community's linguistic behavior relative to (lol) would look like given a larger sample size. As it stands, it is difficult to take results for community 2265 as evidence for or against a difference in norms for (lol) between communities.

Table 2: 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

It is quite possible that the distributions for (lol) in each community are not significantly different. To test this null hypothesis, Fisher's exact test was used. A χ^2-test was not appropriate given the low expected counts for some cells in the relevant contingency table. The null hypothesis that the distribution of (lol) for each community was the same was rejected ($P < 0.05$). Using Cramér's V to measure the effect size returned a value of 0.18, a generally small effect size suggesting that the differences were not great. The large number of variants and communities makes displaying a contingency table of the results here both difficult and uninformative, but the proportions for the ⟨lol⟩ variant range from 75.8% in community 756 to 52.5% in community 1032, though most were had proportions near 70.0%. Thus, while the dominance of the ⟨lol⟩ variant in almost every community does vary, and while the distributions for (lol) are not identical, the differences are not great. It would be difficult to argue that the spelling of (lol) varies enough from community to community to make it maleable as a marker of one's social identity.

Similarly, Table 3 presents a summary of (lol) usage relative to the province[5] of the Twitter user. While modes other than ⟨lol⟩ appear, this happens only where the number of residents using (lol) is limited to one to three people and so hardly represents what would be produced if a larger sample were taken. In all other provinces, ⟨lol⟩ was always the most frequent variant. Also, as with community, the null hypothesis that the distribution of (lol) for each province was the same was rejected ($P < 0.05$), though in this case the effect size was even smaller than for community ($V \approx 0.16$). Examining provinces thus returns

[5]In some cases, these are actually states as parts of the US and of New Zealand were captured during the data mining process, presumably because these tweets were still processed through servers in the Maritime Provinces.

results that are even less promising for the idea that (lol) is socially meaningful. There are indeed some differences between the provinces, but nothing great enough to indicate that Twitter users express their provincial identities through the spelling of (lol).

Table 3: Summary statistics for each province

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

Another way to approach the question of whether (lol) is socially conditioned is to look at individuals relative to their communities. While communities may have clear norms that differ little from each other, this does not mean that all speakers follow these norms. Indeed, out of the 82 active producers of (lol) in the data, defined as those producing at least 10 tokens, 18 who belong to ⟨lol⟩ dominant communities use another variant most frequently, namely ⟨LOL⟩ or ⟨Lol⟩. A sample of those included in this group who produced at least 20 tokens of (lol) are summarized in Table 4. All of these users other than Ambria Howard are above the 64th percentile for PageRanks in their communities, meaning they are fairly central members of their communities. Ambria is an outlier with a PageRank percentile of 0.37, making this user rather peripheral to their community. This suggests that individuals do in fact go against the norms of their community and that those individuals who do so are generally those who have a central position in their community.

In this sense, (lol) is indeed socially variable, though it is clearly not indexing any particular community. The motivation for why these central community members buck the norms of their communities would likely require a discourse analysis that is beyond the scope of the present study but would perhaps be an interesting avenue for future research.

Table 4: Summary statistics for a sample of outlier individuals

User	Mode	Diversity	PR Percentile	Community
Briana Gray	Lol	0.57	0.76	1032
Cherokee Martin	Lol	0.59	1	1032
Aliyya el-Mowad	LOL	0.45	0.7	302
Erin Robinson	LOL	0.57	0.99	572
Jasmine Boller	LOL	0.4	0.78	1032
Josiah Eguino Silvas	LOL	0.03	0.76	1032
Tony Nauth	LOL	0.63	0.64	572

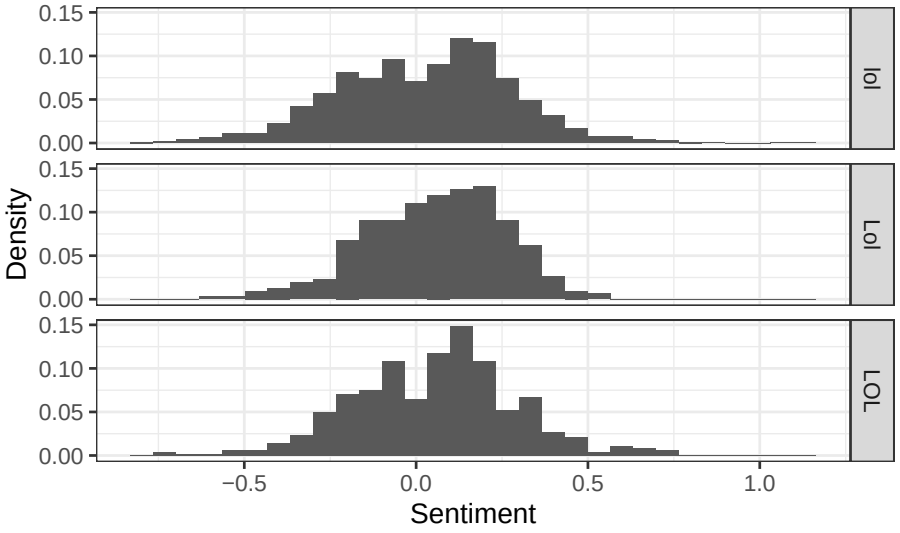
3.2 Pragmatic constraints

While some small amount of the variation in (lol) may be explained by a handful of users going against the norms of their communities, it would be difficult to argue that (lol) is primarily conditioned by how users express their social identities. Another possibility is that variations in the spelling of (lol) are used for pragmatic effect. One such possibility explored here is the relationship between the variant of (lol) used and the sentiment of the tweet. The assumption is that, if different spellings are associated with different sentiments, there is some sort of pragmatic work being done by those particular spellings. This work could be boosting the sentiment in the same positive or negative direction that the rest of the tweet suggested. Alternatively, the spelling may be doing the opposite, softening a negative tweet or adding mockery to a positive tweet. These nuances would be best explored through a thorough discourse analysis, which is beyond the scope of this study, but the quantitative results here are indicative of whether that is a worthwhile avenue for future research.

Among all the variants of (lol), three were exponentially more frequent: ⟨lol⟩, ⟨LOL⟩, and ⟨Lol⟩. The sentiments for tweets containing only these three variants were thus analyzed. Additionally, tweets that received sentiments of zero were excluded. Figure 2 shows the distribution of sentiments for each of the major variants of (lol) where ⟨lol⟩ has a mean of 0.027, ⟨Lol⟩ a mean of 0.056, and ⟨LOL⟩ a mean of 0.055.

All three variants have positive sentiment means, though they are not much above the neutral sentiment of zero. Relative to each other, the two version containing capitalization do appear to be different from ⟨lol⟩ in that they are more positive. Indeed, a one-way ANOVA for the difference in means shows this difference to be statistically significant ($P < 0.05$). While this does suggest that some amount of pragmatic work is being done by the different variants of (lol) in relation to sentiment, the amount is seemingly very small. If (lol) is in fact pragmatically conditioned to a great extent, it is unlikely that this will be found by looking at sentiment. Indeed, one of the sentiment dictionaries available in the sentimentr classifier package includes *lol* with a corresponding sentiment of 0.111 (Cambria et al., 2016). Out of the 23,626 lexical items in this dictionary, 8,786 have higher sentiments, most of which are at least twice

Figure 2: Density plots for the sentiments of tweets for each of the three major variants of (lol)



as high as *lol*, meaning that a large proportion of lexical items are assumed to have a much larger impact on sentiment than *lol*. A more fruitful direction may thus be to look into other pragmatic functions, perhaps using discourse analysis instead of a quantitative analysis.

4 Discussion

The goal of this study has been to search for possible social and/or pragmatic constraints on the realization of the orthographic variable (lol). The methods used have been quantitative and so work well for uncovering aggregate group patterns. However, while there is some evidence for such patterns, they are undoubtedly small. The distributions of realizations of (lol) do differ from community to community and province to province, yet the norm of lowercase ⟨lol⟩ is shared by all groups. Variants of (lol) are also associated with positive sentiment tweets, but the sentiment polarities of said tweets are by and large only slightly above neutral. In both cases, however, there are clues that a more qualitative analysis of individual behavior would uncover both social and discursive functions for (lol).

4.1 Outlier individuals

As discussed in section 3.1, some Twitter users show personal norms for (lol) that do not adhere to the norms of their communities. Those personal norms

do not match any outgroup norms captured in the data, so it cannot be said that they were indexing other analyzed communities through their spelling of (lol). Other possibilities remain to be analyzed, such as whether these users were indexing general personal qualitites, broad outgroup membership in the vein of imagined communities (Anderson, 1983/1991), or simply outgroup communities that were not captured here. In any case, while it is worthwhile to discuss the behavior of these outlier individuals, a proper interpretation is beyond the scope of the present study.

The tweets below provide form a small sample of those produced by users whose mode for (lol) was all uppercase ⟨LOL⟩, and in each case, there are some capitalization practices in their writing styles that are worth noting. In tweet 5, Jack Encina uses *BTW* 'by the way' with the fully capitalized spelling ⟨BTW⟩, just as they do with (lol) in the same tweet as well as all others. The same user also capitalizes ⟨Madam⟩ in tweet 6 as one might do for kinship terms when used as names in formal writing styles. In the first case, using all uppercase for acronyms matches prescriptive writing norms for English, but the second case does not unless it is an instance of hypercorrection, which cannot be determined from the analysis in this study. Jasmine Boller also uses a mix of standard capitalization practices along with non-standard practices, as in tweet 7. This example contains the acronym ⟨USA⟩ in all uppercase but also the non-standard ⟨Presidents⟩ in place of ⟨presidents'⟩. Again, the latter may be a case of hypercorrection, but there are many other possibilities, as well, and sorting these possibilities out is beyond the scope of the present study.

5. Jack Encina: @thejasonmack BTW, of course I can take a joke. I laughed at you didn't I? LOL! No offence, pal , keep smiling!

6. Jack Encina: @Lesqueenb LOL!! The image did cross my mind also Madam!

7. Jasmine Boller: @pharris830 LOL Sounds like Trump. I bet Ivanka will write his speech and use the past USA Presidents speeches to make it sound authentic.

Also present in tweet 7 was discussion of then United States president Donald Trump. As these tweets were collected during the time when the highly controversial Trump was entering office, he was naturally the focal point of many online discussions. Likewise, the tweets below all touched on Trump or politics in general. Tawfeeqa al-Khalifa in tweet 8 capitalizes ⟨President⟩ counter to formal writing styles, but also capitalizes other words that are not always capitalized in CMC, such as ⟨I'm⟩ and the sentence-initial ⟨You're⟩. This is in contrast to Josiah Eguino Silvas in tweets 9 and 10 who does not capitalize the first words in sentences nor the pronoun *I* nor the proper name *Trump*. What Josiah does capitalize, however, is ⟨Global Warming⟩. The reason for this token of capitalization is not clear, though it is also not unusual for Josiah. They also capitalize words like ⟨Death Penalty⟩ and ⟨Dickheads⟩, for example.

8. Tawfeeqa al-Khalifa: @Vicki170266671 @shogunator1 @wikileaks You're the people that voted Donald Trump as President, and you say I'm insane? LOL

9. Josiah Eguino Silvas: @iowa_trump funny way the leftists fight Global Warming LOL

10. Josiah Eguino Silvas: @BlissTabitha @BreitbartNews well i didnt see that coming, actors being critical of trump LOL

These outlier users who do not follow the norms of their communities for (lol) may be indexing personal qualities or perhaps imagined communities, but it is also possible that they are indexing other communities that were not analyzed here. Indeed, only 13 communities were analyzed in this study out of the 8,945 initially detected. However, it is not clear that any one of those 8,945 communities is salient to users in any other community. One limitation of community detection is that defining communities by interactions alone does not necessarily mean those communities are salient in the minds of other nor even in the minds of members of those communities. Combining community detection with ethnography would be helpful for uncovering such saliencies as well as what is behind the behavior of individuals who buck community norms.

4.2 Individual variation

Out of the 82 active users of (lol) in the data analyzed here, only 17 had diversities of zero, meaning over 75% of these users employed at least two variants. The majority of these users were naturally not outliers and instead conformed to the norms of their communities. Their use of spellings of (lol) other than ⟨lol⟩ could of course have the same motivation as for outlier users who do not fit their communities' norms. Unlike outliers, however, a norm-adhering user is not habitually producing forms such as ⟨LOL⟩, and so whereas outliers may arguably be doing so in order to establish some sort of identity, non-outliers may be more likely doing so to create some discursive effect via unexpected contrasts to their norms, just as Androutsopoulos (2000) argued that non-standard spellings can signal attitudes or frames of interpretation.

Following the concepts of paralinguistic restitution (Thurlow and Brown 2003, as cited in Schneier, 2021, p. 3) and affective lengthening (Schnoebelen, 2012, pp. 117-118), as discussed in section 1.3.2, a natural function to consider is emphasis. Specifically, capitalized or reduplicated variants of (lol) may be encoding a more emphatic reaction than the typical ⟨lol⟩ variant to make up for lack of prosody in text. If this is indeed the case, one might expect to find capitalized and reduplicated variants co-occurring with exclamation marks or perhaps other capitalization/reduplication that is more clearly meant to show stress. Tweets 11 to 16 below are examples from normative users in the data that fit this criteria.

11. Carlton Brink: @sahilshiraz99 lolololol liar!!! U were sexually harassing her and when she shot ur ugly ass down u stalked her....ice cold killer

12. Carlton Brink: @RugbyBrainStorm @magnhusdmitriy lolol I know!! Wales should've won that game but bombed just about every scoring opportunity they had!!

13. Ryan Johnson: @CharlesMBlow lolol @CharlesMBlow ..I'm tired and cranky and that instantly brought me around! Thanks!

14. Adamina Dee-Hamilton: @prvsmatic OH! LOL! Sorry

15. Daijha Tinsley: @danielhorton42 @gabrielledoug LOL!!! So funny how people turn into little bitches when you call them out on here... pathetic bud

16. Fat'hi el-Abdalla: @CohenTisha Really?! LOL!! I got one of those big LOUD mouths too LOL

In tweets 11 and 12, Carlton Brink's use of ⟨lolololol⟩ co-occurs with three exclamation marks and their use of ⟨lolol⟩ with two exclamation marks, matching the number of reduplications. Admittedly, there may not be a connection for this user between the number of exclamations and the level of exclaiming, but these examples at least show reduplicated (lol) going along with what are overt symbols meant to represent emphasis. Similary, Ryan Johnson in tweet 13 used the reduplicated variant ⟨lolol⟩ with sentences that ended in exclamation marks. Exclamation marks also co-occur with the tokens of ⟨LOL⟩ in tweets 14, 15, and 16, though in these cases, the exclamation marks are directly adjacent to the tokens of (lol), which are treated as independent sentences. This suggests that what is being exclaimed is the laughter itself. Additionally, ⟨LOL⟩ in tweets 14 and 16 also co-occur with other fully capitalized words, ⟨OH⟩ and ⟨LOUD⟩, respectively. The latter is particularly noteworthy as ⟨LOUD⟩ is found within a sentence featuring otherwise standard capitalization patterns, making it more contrastive, and it also overtly describes the volume of the user's speech. It would be difficult to argue that Fat'hi el-Abdalla is not using capitalization in both ⟨LOUD⟩ and ⟨LOL⟩ to emphasize these words, though this cannot be established concretely just from this anecdotal evidence.

Some variants of (lol) seem as though they may have very specific meanings, as well, though the available data for the most likely candidates in the corpus used here is far too sparse to even provide good examples. For instance, ⟨yolololololo⟩ and ⟨TROLOLOLOLOLOL⟩ are each used once, but in both cases, they make up the entirety of the content of those tweets, and so no hints for future research on these are available. The former seems to be a blend with *YOLO* 'you only live once' and the latter a blend with *troll*, understood as an internet user who purposely attempts to get a rise out of their interlocutors purely for their own entertainment. To get beyond this very basic understanding of such variants necessarily requires a more focused discourse analysis than what the current study can include.

5 Conclusion

The aim of this study was to search for possible social and pragmatic conditioning for the realization of the orthographic linguistic variable (lol) on Twitter. The methods used were squarely quantitative, employing the modern social network analysis technique of community detection in particular. The results suggested that communities do differ slightly in their distributions for the variants of (lol), but all communities had the same norm of ⟨lol⟩. It was also found that both ⟨lol⟩ and capitalized variants ⟨Lol⟩ and ⟨LOL⟩ were associated with positive sentiments, the latter two being associated with slightly more positive sentiments, but that the effect size was not great. A more qualitative look at the data revealed that approaching (lol), and perhaps orthographic variation in general, from a more individualized angle before looking for aggregate patterns may be more fruitful. Some individuals do pattern contrary to their communities, and those that pattern with their communities are still motivated to occasionally use non-normative spellings, but possible explanations are not easily explored through quantatitive analyses.

References

Anderson, B. (1991). *Imagined Communities: Reflections on the Origin and Spread of Nationalism* (2nd). London Verso. (Original work published 1983)

Androutsopoulos, J. (2000). Non-standard spellings in media texts: The case of German fanzines. *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.

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.

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:0803.0476v2*. 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

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2666–2677.

Cherny, L. (1996). *The MUD register: Conversational modes of action in a text-based virtual reality* (PhD). Stanford University.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words! Linguistic style accommodation in social media. *WWW 2011*, 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*, 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*, 518–521.

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). Cambridge University Press. (Original work published 1966)

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.

Milroy, L. (1987). *Language and Social Networks* (2nd). Blackwell. (Original work published 1980)

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *arXiv: cond-mat/0308217*. 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*. https://doi.org/10.1109/HICSS.1999.772680

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.

Rosen, A. (2017). Tweeting Made Easier. Retrieved June 30, 2021, from https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier

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. *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.

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

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.

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