Marginality and orthographic variation in (lol)

Joshua McNeill

December 7, 2020

Abstract

In this study, a corpus derived from Twitter is used to analyze orthographic variation in the linguistic variable (lol). Modern social network analysis techniques are employed to detect communities in the corpus and calculate centralities, particularly PageRank, for members of those communities. The objective is to advance understanding of individual agency and how it relates to one's position in a social structure. As such, I ask whether the diversity of realizations of (lol) can be understood as a function of one's centrality in one's community. The results suggest that this is not the case, though as an early exploratory study, there is still much to be done to draw any firm conclusion. It is suggested that more linguistic variables be analyzed in a similar quantitative manner.

1 Introduction

In earlier studies of langauge variation, speakers who were peripheral to their communities in some way were often disregarded. Indeed, Labov (2006) explicitly disregarded two African-American speakers in his study of the Lower East Side in New York as they had moved to the area from Virginia when they were 10 years old (p. 118). This practice was directly critiqued in Horvath and Sankoff's (1987) study of Sydney, Australia where they suggested that excluding new arrivals and marginal members of a community might also lead to throwing out important data for better understanding language change (p. 201). Their proposed solution was what they called the linguistic grouping approach to studying language variation, wherein a researcher would first use a clustering algorithm to group speakers according to sharing similar linguistic features and then search for the social categories that those in the resulting groups might have in common, whereas variationist studies typically proceed in the opposite way.

More recently, sociolinguists have become much more interested in agentivity and hence marginal community members. When certain speakers are excluded from analysis for being recent arrivals, it now has more to do with

the type of variation being analyzed, particularly when that variation is phonological. For instance, despite her clear interest in agentivity and analysis of four marginal community members in her study of Belten High, Eckert (2000) herself excludes those who moved to the community after 8 years of age (p. 82). However, Eckert was interested in phonological variation, and by that time, there was good evidence to suggest that most speakers do not adjust their phonologies after that age[1].

The objective of this study, then, is to look again at peripheral community members and how they behave, whether they be brand new members of a given community or long-term members who are nonetheless still peripheral. Specifically, I aim to use social network analysis to detect communities on the social media platform Twitter and analyze orthographic variation in these Twitter users' use of ⟨lol⟩ as a function of their centralities in said communities.

1.1 Second Dialect Acquisition

Examining the linguistic behavior of new arrivals to a community has been an area of study in sociolinguistics at least since Payne's (1980) study of speakers in King of Prussia, a town near Philadelphia, Pennsylvania. Payne looked specifically at variation in vowels in her study, using age of arrival in King of Prussia as a extra-linguistic variable. In general, she found that 60% to 70% of the children who moved to the area before the age of 4 learned the local vowel system in its totality, whereas this percentage dropped to 40% to 67% for those who moved between the ages of 5 and 9 and 0% to 67% for those who moved between the ages of 10 and 14 (Payne, 1980, p. 155). Payne's results have since been used as evidence that, at least for phonological variables, linguistic systems are relatively stable after age 8 so that moving into a new community where a different system is dominant will not lead to the adoption of this new different system. Whether this stability is a result of the literal impossibility of fully learning a new sound system or whether sound systems too strongly represent a speaker's identity by adolescence for them to give them up was not addressed in this early study, however.

Over time, research into phenomena such as what Payne examined has come to be known as second dialect acquisition. More recently, second dialect acquisition features particularly prominently in the work of Nycz, but a number of others have also touched on this topic[2]. However, acquisition is not necessarily the topic that is being addressed in the current study. Acquisition implies that a speaker is learning to use a new linguistic form to which they do not already have access in their mental grammar or perhaps to which they have never even been exposed. The focus on dialects in this area of research, understood to be geographically defined varieties, additionally brings geography to the forefront as a conditioning factor, somewhat sidelining other social factors such as class, race, or gender. It appears then that the very name second

[1] Discussed in more detail below

[2] See Nycz (2015) for a review.

dialect acquisition suggests particular research interests which are not central to my aims with this study, though they are still important.

The aim here is to explore what peripheral members to a community do more generally. Peripherality certainly can be the result of geographic relocation but need not be. In practice, speech communities (e.g., Labov, 2006) are geographically defined, but communities can also be defined in other ways. For instance, and taken up in more detail below, communities of practice as used in Eckert (2000) and Bucholtz (1999) to situate communities primarily around shared activities. Partaking in extra-curricular school activities, in Eckert's study for instance, or playing the numbers game (Poplack, 2000, p. 216) might draw people together to form a community where perhaps geographical proximity alone would not.

Likewise, in studies of phonological variation, acquisition may be a key concept even when there is already a familiarity with the sound system a speaker aims to adopt, as the rules of the system may be opaque, and the physiological motions used to achieve certain sounds may require practice. Morphosyntactic variables also suggest the need to learn nuanced rules in order to truly adopt a different system. Many speakers, for instance, are familiar with habitual *be* but do not use it. If such a speaker wished to begin using it, they would have to learn that it expresses the habitual aspect as opposed to being simply invariance of inflections for the copula, something that is not immediately obvious through observation alone. Indeed, Eberhardt and Freeman (2015) claimed that habitual *be* is "one of the most frequently used but least understood features of AAE [African-American English] by outsiders," hence the fact that they found it "striking" that non-African-American singer Iggy Azalea uses this feature very accurately (p. 311).

There are linguistic levels in which acquisition is arguably much simpler and so acquisition, while not absent, is not the central concern. For instance, observation alone can teach a speaker all they need to know about the meaning of perhaps most lexical items, at least for those that are not function words. I know of no argument for the stability of mental lexicons after age 8 such as is often argued for phonological systems, and the ease of acquisition is a possible explanation. Similarly, new spellings for known words, the focus of this study, can arguably be acquired easily through observation alone granted that one is literate to begin with. One could imagine speakers who are intimately familiar with spellings that they have never once in their lives used but could use at any time if they so chose, such as ⟨u⟩ for ⟨you⟩. This relative unimportance of acquisition in the sort of variation I am analyzing here is part of what leads me to distinguish this study from previous related work in second dialect acquisition.

1.2 Twitter

Twitter is a micro-blogging social media platform that provides sociolinguists with numerous interesting opportunities for studying language variation. Users

of Twitter post short messages, tweets, up to 280 characters in length[3] that can optionally be directed at a specific user or users using the @ symbol followed by target's username. Whether directed or not, the default setting is for all tweets to be publicly accessible even to people who do not have Twitter accounts. This fact is regularly understood by users, and so creating corpora from tweets does not pose as many ethical dilemmas[4] as does creating corpora from other similar online platforms where communication is expected to be generally private, such as Facebook.

Because of the text-based nature of Twitter and the ease of capturing large swaths of conversation without conducting a single interview, sociolinguistic studies based on this platform have tended to involve very large corpora and to be highly quantitative. For instance, in their study of gender identity and lexical variation, Bamman et al. (2014) constructed a corpus over a six month period that included 14,464 Twitter users and 9,212,118 tweets (p. 140). Comparing this to Labov's (2006) suggestion in 2006 that a sample of 60 to 100 speakers is relatively large and is enough to analyze stratification in a single city (p. 401), having thousands of speakers and millions of utterances is a large leap in size.

It is also notable that Bamman et al. (2014) were interested in lexical variation, whereas perhaps most variationists focus in on phonetics and phonology. There is a clear practical reason for analyzing sound systems, of course: this allows a researcher to maximize the tokens they collect and minimize the amount of recordings that must be done to obtain those tokens. Focusing on lexical items is much more difficult outside of perhaps function words but not when corpora with millions of utterances are collected. Eisenstein (2014), for example, also looked at lexical variation of regional lexical items on Twitter, such as *jawn* which originated in Philadelphia, Pennsylvania. Similarly, Pavalanathan and Eisenstein (2015) were able to test Bell's (1984) theory of audience design in how users of Twitter style shifted their lexical items according to the intended audience, using either more generic vocabulary for broad audiences or more regional vocabulary for narrower, more local audiences. This latter study also provides an example of another distinction between variationist studies of Twitter and traditional interview-based studies: the number of linguistic variables analyzed. Pavalanathan and Eisenstein (2015) looked at over 200 variables whereas a traditional variationist study may include only one to perhaps nine linguistic variables. This scale of analysis is also present in the aforementioned Twitter-based studies.

1.3 Orthographic Variation

The generally casual nature of Twitter in comparison with other forms of written communication also provides a space where it is appropriate for users to

[3]At the time of data collection for this study, the character limit was 140 characters.

[4]This should not be read as there being no ethical dilemmas involved in creating such corpora. The standards used for this study will be discussed in section 2 Methods.

eschew orthographic conventions that would be tightly controlled in other contexts, either through subtle social controls as in writing letters, for instance, or through much more explicit controls as in being graded on an essay in an English class. Indeed, several studies have shown that flouting orthographic conventions is quite common on Twitter. Tatman (2016) and Eisenstein (2015) both used orthographic forms on Twitter to examine to what extent users shape their writing to approximate what they might see as the indiosyncratic features of their spoken pronunciations. Similarly, Ilbury (2020) recently examined how persona are created on Twitter through the use of orthographic variation. While Ilbury treats orthographic variation as its own system independent of speech, to date, there appears to have been no attempts to take a quantitative variationist approach to analyzing written forms on Twitter.

1.4 Social Network Analysis

Another advantage of using Twitter for variationist research is that it provides ample opportunity to employ robust social network analysis techniques without many of the difficulties that normally come with sketching out ties between people. In social network analysis, researchers map out which people in their studies are connected to which people, ultimately generating a sort of web of connections that provides an opportunity to perform quantitative analyses that are not as easily performed when identifying communities and relationships in less concrete terms.

Social network analysis – not to be confused with terms like *social network* or *social media* as used to describe platforms like Twitter – has been used in variationist research since Milroy (1987) used it to study Belfast. Her motivation for employing social network analysis was the inadequacy of the then current variationist methods for explaining why two people who can be placed into all the same broad social categories still end up speaking differently, as did Hannah and Paula, two co-workers that she describes in her study (Milroy, 1987, pp. 131-134). Indeed, social network was able to explain such anomalies by analyzing how integrated speakers were into different social networks in Belfast, suggesting that one's interactions may be more important than social characteristics that can be ascribed to them.

However, because of the difficulty in reliably establishing ties between people, methods for employing such analyses have remained mostly fairly simple and have not kept up with methodological developments that have occurred in sociology, though there are notable exceptions more recently such as Dodsworth and Benton (2017). Milroy (1987) herself did not actually sketch out a network but instead assumed that different neighborhoods in Belfast each constituted separate networks. She then created an index from that ranged from zero to five to quantify speakers' integration into those networks where such factors as working with people from the same neighborhood or having kin in the neighborhood would increase one's score (Milroy, 1987, pp. 139-142). Similar indices have been used since in studies such as Li et al.'s (2000)

study of Chinese immigrants in Tyneside, England and Sharma's (2011) study of Indians in London.

Community Detection One important area of development in modern social network analysis techniques that has been 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 mostly all interact with each other. For instance, if a network consists of Joe, Kelly, Bob, and Ted, and Joe, Kelly, and Bob all know each other but the only one that knows Ted is Bob, then 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.

At the core of any tie, though, 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.

A community conceptualized as a group of people who interact with each other does not necessarily accord with other conceptualizations of communities. For instance, speech communities are geographically chosen and then linguistically validated. There is no need to establish that the members of such communities interact with each other because interaction is not to be found in the traditional definition. Labov's participants in the Lower East Side simply need to live in the same geographic area and share speech patterns and speech evaluations to be part of the same speech community. Whether they all actually interact with each other is immaterial. This is not to say that there communities of practice or communities as defined in social network analysis are better or more valid than speech communities, but each conceptualization lends itself better to answer different sorts of questions.

Centrality Measures A more recent concern for sociolinguists has been analyzing language variation with special attention to individual agency, which is at least implicit in Eckert's (2012) view of the "third wave" of sociolinguistics. Social network analysis give researchers tools to approach this issue in a quantitative way through the use of centrality measures. In this case, centrality refers to how integrated a person is into a given community.

There are many different centrality measures, and in fact Milroy's index and the similar indices used by subsequent researchers are all effectively centrality measures. What all centrality measures have in common is quantifying the number of people in a community with whom a given person interacts and

quantifying the frequency or importance of those interactions. A very basic way to measure this is with what is by measuring a person's degree in the community, which is just the sum of their ties in the community. For instance, going back to the hypothetical network made up of Joe, Kelly, Bob, and Ted, Ted would have a degree of one since his only tie links him to Bob, whereas Joe and Kelly would have degrees of two because they interact with each other and Bob, and finally Bob would have a degree of three because he interacts with all the other people in the network. If we were just interested in a person's degree within the community, however, Ted would have a degree of zero since he is not part of the community, and the others would each have a degree of two.

The above describes a simple case not only because of the small size of the network but also because ties can be directional and there can be multiple ties between the same two people. Because interactions on Twitter are made explicit through use of the @ symbol to identify addressees, and because this adds directionality and the potential for counting multiple ties (i.e., interactions) between two users, it is possible to easily construct robust social networks from data collected from the platform, implement performant community detection algorithms, and use more nuanced centrality measures.

Just such a community detection algorithm and centrality measures were used in McNeill (2018), the study that pre-empted the present work. Although this previous study was focused on analyzing lexical variation and language contact as opposed to orthographic variation, the object of analysis included *lol* as a lexical item, and the communities detected and the centrality measures used are still perfectly valid for further analyses. The findings in McNeill (2018) suggest Newman and Girvan's (2004) approach to communities yields meaningful results for analyzing language in that the community one belongs to is a predictor of the lexical variant one will use for (lol), even in what would be considered bilingual conversation. However, a lack of any users who produced a large number of lexical tokens made it impossible to conclude anything about the relationship between a user's centrality in a community and their linguistic behavior. As such, this study aims to look at a broader swath of data by ignoring the issue of language contact and focusing on orthographic variation in the spelling of *lol* 'laugh out loud'. Specifically, I ask the following question:

- Do individuals show more or less diversity of orthographic variants of (lol) as their centrality in a given community goes up?

2 Methods

2.1 Data Collection

Data for this study comes from the same corpus constructed for McNeill (2018). This corpus includes tweets originating from Twitter servers in the Maritime Provinces of Canada between January 11th and February 7th, 2017 and includes 307,878 directed tweets. All tweets were used when applying Newman

and Girvan's (2004) community detection algorithm, even those that would not ultimately be analyzed. This initially yielded 8,945 communities, but as the goal at the time 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 4,732 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).

2.2 Data Coding

As mentioned above, all 307,878 directed tweets were used for detecting communities. The implementation of Newman and Girvan's (2004) algorithm provided in the software package Gephi (Bastian et al., 2009) was used for this task. Newman and Girvan's algorithm aims to maximize modularity in a network, which is the quality of a given division of the network. Essentially, the algorithm attempts to divide the network into different potential arrangements of communities and calculates the modularity for each. It ultimately chooses the arrangement that returns the highest modularity value. Every token was thus coded for the community of the user who sent the tweet given the results of the community detection algorithm.

Each token was also coded for various centrality measures, all of which were calculated in Gephi. Many basic measures were included, such as the degree of each user, and also some slightly more complex measures, such as the in-degree and out-degree, which are the degrees for just incoming and just outgoing tweets respectively. The centrality measure focused on here, though, is PageRank (Brin & Page, 1998).

PageRank was originally developed for ordering search engine results and ultimately led to the creation of Google. Page A's PageRank PR, or in this case person A's PR, is calculated as follows:

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

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.

Finally, several other factors were coded for each token that are not of direct importance in the present study. For instance, geographic location was coded based on that which was entered in the profile of each user. This information is optional and manually entered by users, meaning it is not always reliable. It was also found to be less important statistically and subsumed by community in McNeill (2018) in that those who were from the same area were likely to be part of the same Twitter community. Additionally, the language of the text surrounding each token was coded, as well as the device tweeted from (e.g., mobile or web), and the number of follow and followers of the user.

2.3 Statistical Analyses

As the data analyzed in this study is all categorical, the descriptive statistics used are fairly basic with a notable exception. The measure of central tendency for (lol) is the mode. This can be calculated for individuals as well as for communities as wholes for comparison. What is key, however, is having a way to quantify the dispersion of variants of (lol) for either an individual or a community as my research question is specifically interested in how stable one's variation in spelling becomes or does not become relative to their centrality in a community. The Simpson diversity index (Simpson, 1949) is thus used as a measure of this stability, which is rarely found in variationist studies (see Greenberg 1956 and Sharma 2011 for noteworthy examples).

The Simpson diversity index D is common in ecological studies, and it is calculated as follows:

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

Here, 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.

As both centrality measures and the Simpson diversity index yield continuous values, they can be plotted against each other to search for a correlation or monotonic relationship. This is done for both the entirety of individuals included as well as for those produced at least 10 tokens of (lol). The latter is done to account for the possibility of many users with very low diversities of (lol) only having low diversities because they only produced very few tokens of (lol), perhaps as few as one token. Where appropriate, a test for statistical significance of a correlation is used.

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 1 and 2 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.

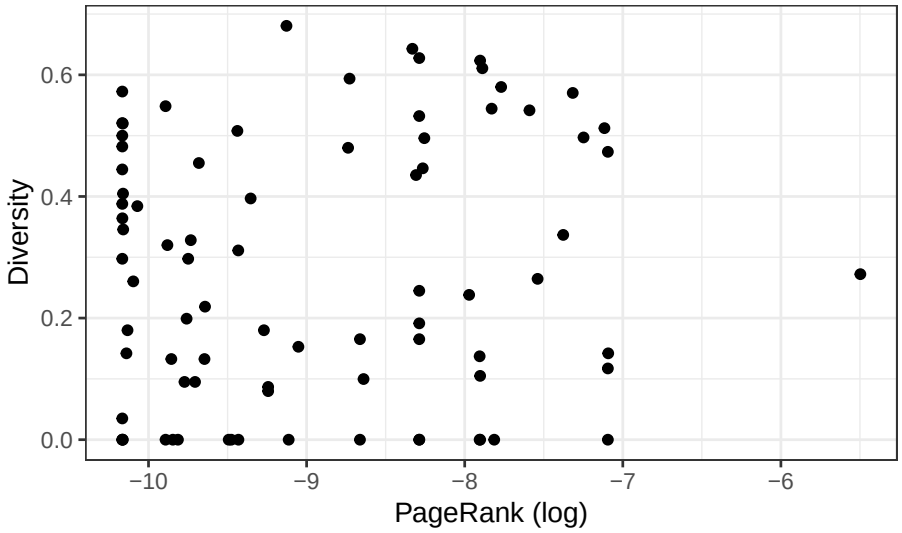
Table 1: Summary statistics for each community

Community	Mode	Diversity	Members who used (lol)
173	lol	0.45	46
302	lol	0.42	316
572	lol	0.45	90
756	lol	0.41	25
799	lol	0	1
1032	lol	0.61	358
1097	lol	0.53	57
1227	lol	0.57	74
1291	lol	0.49	15
1917	lol	0.44	138
2067	lol	0	2
2265	LOL	0.71	8
6817	lol	0.52	10

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 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 1, 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 2 provides summary statistics for each user in this community.

Figure 1: Diversity according to PageRank for users who produced at least 10 tokens of (lol)



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 0.27. Rithanya's place in this community is fairly central, as well, with a PageRank of 0.0041, which despite the narrow range of PageRanks, is one of the highest among any user analyzed in this corpus for any community.

Table 2: Summary statistics for community 2265

User	PageRank	Mode	Diversity	Tokens
Calisse	0.0039	lol	0	4
Rithanya	0.0041	LOL	0.27	13
Kenlyn	0.004	lol	0.75	4
Ayedan	0.0038	lol	0	4
Bertile	0.0038	Lool	0.5	2
Noorain	0.0041	LOL	0.44	3
Condy	0.0058	lol	0	1
Shelbea	0.0038	lol	0	1

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 3 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 variants of the linguistic variable than any individual member of that community, but overall, Figure 2 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 3 outdue the diversity of their community.

Table 3: 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						
Rithanya	0.0041	LOL	0.27	13	2265	0.71
Amair	0.00083	lol	0.14	26	1291	0.49
Saadiya	0.00083	LOL	0	10	1291	0.49
Seprina	0.00083	lol	0.12	16	1291	0.49
Rheya	0.00083	lol	0.47	13	1291	0.49
Lowest PageRanks						
Leyann	0.00004	LOL	0.03	113	1032	0.61
Dellanira	0.00004	lol	0.44	21	1032	0.61
Jocques	0.00004	Lol	0.57	36	1032	0.61
Kentoria	0.00004	lol	0	11	1032	0.61
Yogi	0.00004	LOL	0	10	1032	0.61

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 3, 4, and 5. Other than community 2265 where Rithanya made up the bulk of the total tokens, the 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

Figure 2: Diversity of each user's realization of (lol) by the diversity of their community as a whole among those with at least 10 tokens

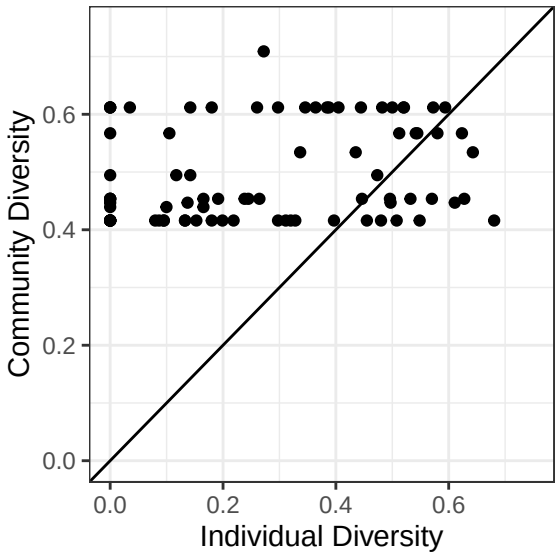


Figure 3: The distribution of (lol) for the communities of those with the highest PageRanks and lowest

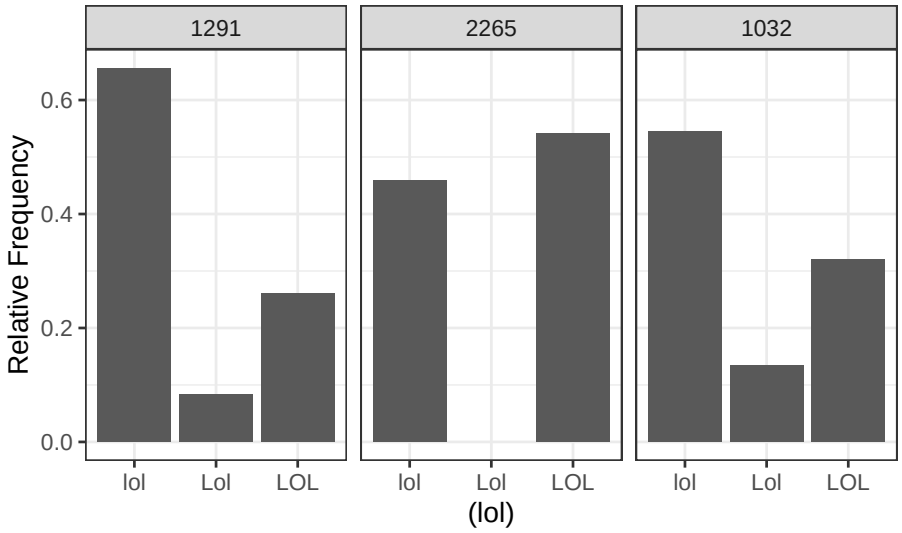


Figure 4: The distribution of (lol) for users with the highest PageRanks

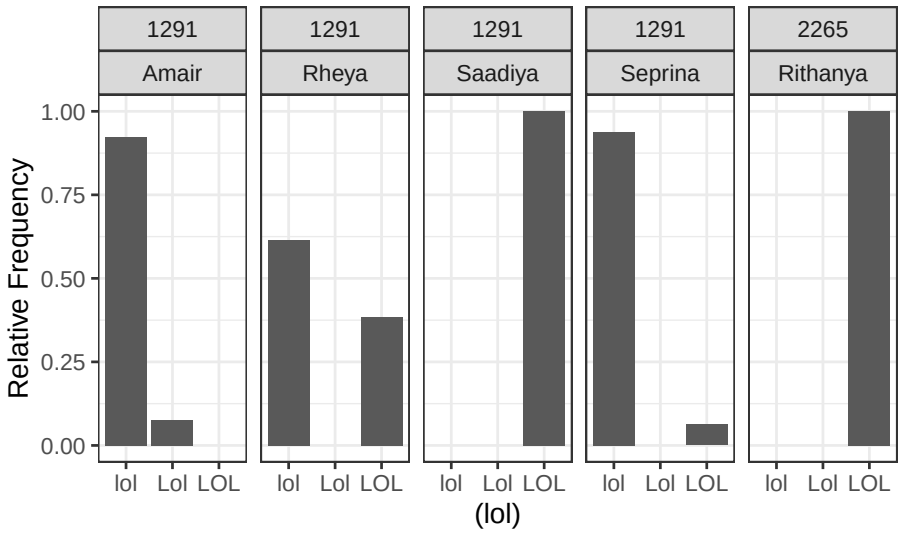
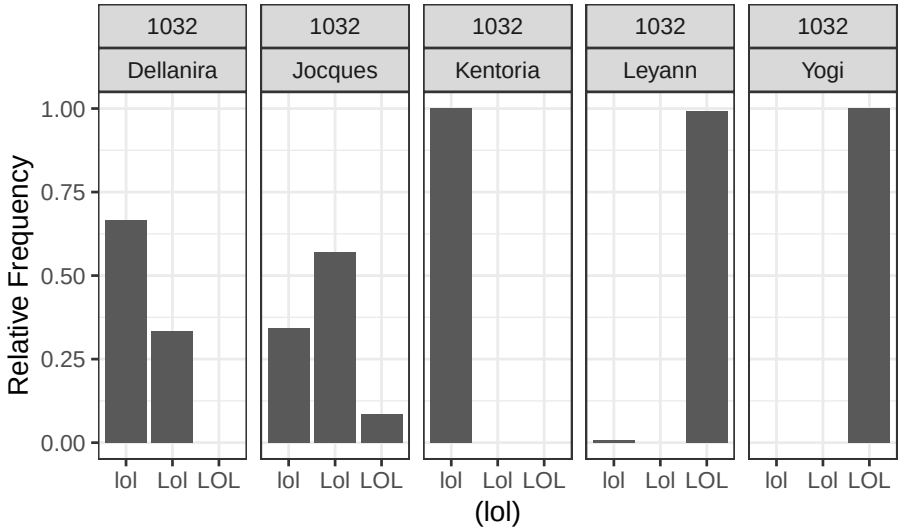


Figure 5: The distribution of (lol) for users with the lowest PageRanks



pattern. Saadiya produced only ⟨LOL⟩ despite being central to this community. Likewise, Leyann and Yogi among the most marginal members in community 1032[5] 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 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 2 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

Recall that the research question for this study asked if the diversity of the realization of (lol) increased or decreased relative to a user's centrality in their community. There is no clear relationship betwen this linguistic variable and diversity, and there are a couple reasons why that might be outside of simply concluding that social position is not an important factor in language variation. Indeed, such a conclusion would at least implicitly be at odds with long-held assumptions in the variationist literature that social position is important as in, for example, Labov's analysis of central community members in Philadelphia in Celeste S. and Carol Meyers (as cited in Eckert & Labov, 2017), and Eckert (2000) proposed concept of sociolinguistic icons, community members who make extreme use of advanced variants and are also central to their communities (pp. 216-219).

One reason for the lack of a relationship could be that there is no variation in mode from community to community and so there is no new norm for

[5]The five marginal members listed in Figure 5 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.

marginal members to learn. They of course do not automatically follow the norms of the community for (lol), nor do central members for that matter, but they are perhaps behaving linguistically in the same way as they would in the communities in which they are central. To disentangle this potential explanation would require a different community detection algorithm that allows for placing individuals in multiple communities as opposed to Newman and Girvan's (2004) which places every individual into exactly one community. While the algorithm used here has been well validated, it is also unreasonable to think that any individual belongs to only one community *a priori*.

Another reason for the lack of a relationship could be that (lol) happens to be used primarily to express one's personal identity. It is important to remember that (lol) is one linguistic variable out of very many possible variables, all of which may be utilized towards different ends. A marginal member may, for example, attempt to employ lexical items that feel more appropriate to a community to which they are trying to become more central while maintaining their own distinctiveness through their use of (lol). It is important to remember that the goal of participating in a community is not always to conform entirely but is perhaps more often to find the threshold of conformity required to be marked as a member without going any further.

4.1 Limitations

Both of the potential explanations above function admit some limitations to the present study. Because of the community detection algorithm used, how any given individual behaves from community to community cannot be analyzed. Without closer ethnographic work or interviews, the level of conformity aimed for by marginal members cannot be identified. Two other limitations are worth noting, as well: the age range being represented by this Twitter corpus and the type of linguistic variable that was analyzed.

People of many ages use Twitter but not to the same extent. For one, it must be assumed that there is no one in the corpus below perhaps 6 to 8 years old as using Twitter require the ability to read and write. This sets research into orthographic variation apart from research into speech where any age group can potentially be included given that ethical concerns are resolved. Even in the case of abundant precaution to protect those included in a corpus, it is simply not possible to study the orthographic variation of those who are not yet literate. Writing in this sense is another language that they have not yet learned.

Likewise, it is unlikely that there is adequate representation of those over 35 years old in this or any Twitter corpus. Sloan et al. (2015) have shown evidence that the bulk of Twitter users are under 35 years old. This should be kept in mind when comparing the results of studies based on Twitter as while they may be comparable to results from studies like Payne's (1980) as she did not use older speakers, there is little to no comparability possible when looking at studies focused on older speakers, which is no small point as there is still more to learn about age grading (see Wagner, 2012, for a review).

Finally, as has already been mentioned, (lol) is yet one of many possible linguistic variables at play in these communities. In particular, (lol) was most often an adjunct in the corpus used here, which may make it more susceptible to purely pragmatic uses as opposed to, for example, signaling social affiliation since adjuncts are likely to be speaker-oriented adverbs (Ernst, 2009). This is not to say that a variable cannot do both types of work at once, but disentangling the two is made more difficult by the part-of-speech.

4.2 Future Research

As this study is quite exploratory, there are still plenty of avenues to go down and topics to address and clarify. Expanding on the idea of what other linguistic variables would be appropriate to look at in the same way as was done with (lol) here, pronouns are likely a good candidate. At least in English, pronouns are often obligatory, making them frequent enough for robust quantitative analysis. This is particularly true when exceptionally large corpora can be constructed with relative ease as is possible when mining data from Twitter. One pronoun that appears even in the current corpus in various spellings is the second person pronoun, which takes at least two spellings: ⟨you⟩ and ⟨u⟩. It is hard to imagine such a pronoun as performing pragmatic work or as having slightly different references for each spelling, so what is left is its social meaning.

The overall objective of this study was to shed more light on how social position relates to language variation. While individual agency has become much more prominent in sociolinguistic research, precise measurements of social position are rarely used. For instance, Eckert (2000) suggests that Judy is a sociolinguistic icon (p. 217) but seems to have come to that conclusion through qualitative or linguistic means. She does in fact provide a sociograms for the students at Belten High (pp. 173-174), but she does not employ any centrality measures. Having done just that here has allowed for posing a question that is both testable and falsifiable. Qualitative work is of course highly useful, but we are perhaps at a point where that qualitative works needs to be operationalized quantitatively more, as has been done here, to move towards more robust theories of agency.

With the advent of particularly casual medium of written communication on the internet, the dearth of studies of orthographic variation is surprising. Just as phonological variation provides for very frequent linguistic variables that make for effective quantitative analyses, so does orthographic variation for the written language. Lexical variation variation is of course also easily analyzed through a variationist framework, but such studies are generally limited to looking into the diffusion of given lexical items (Grieve et al., 2018, e.g.,) as opposed to the way the realization of lexical linguistic variables might change from group to group and context to context. Such lexical diffusion studies are indeed exciting, but further work on orthographic variation may help fill in some gaps.

Finally, written language needs to be analyzed more in and of itself. To state that written language is separate from spoken language is not at all controversial, yet it is rarely touched on in variationist studies as a distinct entity. While Tatman (2016) and Eisenstein's (2015) work on orthographic variation is elucidating and important, both approach writing through the lens of its relationship to speech. In the case of (lol) here, I have attempted to treat writing as its own distinct language, and in fact it would have been difficult to do otherwise as this lexical item originated in writing and is only marginally used in speech. As exploratory studies such as this lead develop into something larger, it will be important to also develop a conceptual framework that does not confuse written and spoken language.

References

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

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

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

Bucholtz, M. (1999). "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, *28*(2), 203–223. https://doi.org/10.1017/s0047404599002043

Dodsworth, R., & Benton, R. A. (2017). Social network cohesion and the retreat from Southern vowels in Raleigh. *Language in Society*, *46*(3), 371–405. https://doi.org/10.1017/S0047404517000185

Eberhardt, M., & Freeman, K. (2015). 'First things first, I'm the realest': Linguistic appropriation, white privilege, and the hip-hop persona of Iggy Azalea. *Journal of Sociolinguistics*, *19*(3), 303–327. https://doi.org/10.1111/josl.12128

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Blackwell Publishers, Inc.

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

Eckert, P., & Labov, W. (2017). Phonetics, phonology and social meaning. *Journal of Sociolinguistics*, *21*(4), 467–496. https://doi.org/10.1111/josl.12244

Eisenstein, J. (2014). Identifying regional dialects in online social media. *Preprint*, 1–15. https://doi.org/10.1002/9781118827628.ch21

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

Ernst, T. (2009). Speaker-Oriented Adverbs [Publisher: Springer]. *Natural Language & Linguistic Theory*, *27*(3), 497–544.

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

Grieve, J., Nini, A., & Guo, D. (2018). Mapping Lexical Innovation on American Social Media [Publisher: SAGE Publications Inc]. *Journal of English Linguistics*, *46*(4), 293–319. https://doi.org/10.1177/0075424218793191

Horvath, B., & Sankoff, D. (1987). Delimiting the Sydney speech community. *Language in Society*, *16*(2), 179–204. https://doi.org/10.1017/s0047404500012252

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

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

Li, W., Milroy, L., & Sin Ching, P. (2000). A two-step sociolinguistic analysis of code-switching and language choice: The example of a bilingual Chinese community in Britain [original-date: 1992]. In W. Li (Ed.), *The Bilingualism Reader* (pp. 175–197). Routledge.

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) [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

Nycz, J. (2015). Second Dialect Acquisition: A Sociophonetic Perspective. *Language and Linguistics Compass*, *9*(11), 469–482. https://doi.org/10.1111/lnc3.12163

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

Payne, A. (1980). Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. In W. Labov (Ed.), *Locating Language in Time and Space*. Academic Press.

Poplack, S. (2000). Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching [original-date: 1979/1980]. In W. Li (Ed.), *The Bilingualism Reader* (pp. 205–240). Routledge.

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.

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

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data [Publisher: Public Library of Science]. *PLOS ONE*, *10*(3), e0115545. https://doi.org/10.1371/journal.pone.0115545

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).

Wagner, S. E. (2012). Age Grading in Sociolinguistic Theory. *Language and Linguistics Compass*, *6*(6), 371–382. https://doi.org/10.1002/lnc3.343