Marginality and orthographic variation in (lol)

Joshua McNeill

December 1, 2020

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 look use social network analysis to detect communities on

[1] Discussed in more detail below

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 (see Nycz, 2015, for a review). 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 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) 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 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 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[2] 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

[2]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 **??**.

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, 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 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 ex-

plicit 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 use more or fewer 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 3,938 tokens of the orthographic variable (lol), which includes 83 spelling variants.

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:

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

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 whose degrees were at or above the median degree for the community. 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

4 Discussion

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

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

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

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

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).