Centrality and Variability of (lol)

Joshua McNeill

November 28, 2020

1 Introduction

Twitter is a micro-blogging online platform that provides sociolinguists with numerous advantages for studying variation. Users of Twitter post short messages 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 as does creating corpora from other similar online platforms where communication is expected to be generally private, such as Facebook.

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 (2014) both used orthographic forms on Twitter to examine to what extent users stylize 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.

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 *social network* as used to describe platforms like Twitter – has been used in variationist research since Milroy and Milroy (1985) used it to study Belfast. However, because of the difficulty in reliably establishing ties between people, methods for employing such analyses have remained mostly lackluster 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).

Two key aspects of social network analysis are community detection and measuring centrality. Community detection is the process of using algorithms to delineate communities within a social network. In this case, a community is conceptualized as no more than a cluster of individuals who interact more with each other than with others in the network. The reasons for those more frequent interactions are then left to social theory to explain. Centrality is a measure of how integrated an individual is within such a community. Numerous approaches to measuring centrality exist, and indeed Milroy and Milroy (1985) development her own index for this, which has perhaps been the most consistent feature of social network analysis as used in sociolinguistics.

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 in which we currently find ourselves. Twitter and social network analysis give us tools to approach this

issue in a quantitative way. Because interactions on Twitter are explicit through use of the @ symbol to identify addressees, it is possible to easily construct robust social networks. With such networks constructed, modern community detection algorithms and centrality measures are also simple to implement. This allows for identifying exactly where a person is within a community when contrasting their language use with others, and thus opportunities to better understand agency in language variation arise.

Indeed, just such an approach was used in McNeill (2018). In that study, the goal was to analyze lexical variation and language contact as opposed to orthographic variation, however. In terms of community detection, it showed that Newman and Girvan's (2004) approach yielded meaningful results for analyzing language, but a lack of any users who produced a large number of tokens made it impossible to conclude anything about how a user's centrality in a community and their linguistic choices. 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 the adjunct *lol* 'laugh out loud'. Specifically, I ask the following question:

- Do individuals use fewer orthographic variants of (lol) as their centrality in a given community goes up?

2 Method

Data for this study will come from the same corpus constructed for McNeill (2018). This corpus included tweets originating from Twitter servers in the Maritime Provinces of Canada between January 11th and February 7th, 2017 and included 307,878 directed tweets containing upwards of 4,000 tokens of (lol).[1] Tokens are already coded for community, using Newman and Girvan's (2004) algorithm for detecting communities, and various centrality measures, notably PageRank (Brin & Page, 1998), as well as orthographic variant. The Simpson diversity index will be used as a measure of how stable the variability of one's spelling of (lol) is where greater stability is understood as roughly meaning that fewer variants were used and one clear mode was used. As both centrality measures and the Simpson diversity index yield continuous values, the statistical significance of centrality can be tested by testing the significance of the resulting correlation. In the case of a non-linear relationship, the data will either be transformed or a similar significance test for monotonic relationships will be used.

[1] In this case, (lol) was a lexical variable and so included variants such as *rofl* and even *mdr* from French *mort de rire*, so the exact number of orthographic variants of *lol* alone is not given.