The Prevalence of Regionalist Vocabulary on Twitter

Declan Golsen

December 10, 2021

Abstract

This project sought to measure the degree to which regional divisions in vocabulary were evident on Twitter, in the hopes of better understanding the overall relationship between online speech and regional vocabulary. Five concepts with established regional terms were studied, and their geographic distributions were measured. The hypothesis being tested was that these five concepts would see significant differences in distribution across the four Census Bureau designated regions of the United States. The results were mixed; two of the five concepts had significantly regional distributions, while the other three were too evenly distributed to be considered regional. These findings indicate that at least some elements of America's regionalist vocabulary shine through on the internet, though it may be considerably diluted by travel, formal writing, etc. Confounding factors may also have obscured regional variety, such as the use of some regionalisms for other concepts altogether, and the presence of non-human accounts on Twitter.

Contents

1 Introduction

Distinctions between spoken English and the English used on social media are relatively new and still in the process of developing, and a possible distinction worthy of analysis is the prevalence of regionalist vocabulary. Regionalisms often stem from the geographic isolation between speakers of the same language, and the internet has left speakers of English less isolated than ever, strengthening the assimilating effects first brought about by the radio and television.

1.1 Background

Studying the extent to which a few terms vary amongst Twitter users from across the United States can provide insight into how regionalist vocabulary holds up online; if the distribution of terms such as "pop,"

"coke," and "soda" are highly concentrated in various regions of the country (as they are known to be in spoken English), this would indicate that regionalist vocabulary asserts itself online as it does in spoken English. Academic articles have compared twitter corpora to traditionally compiled corpora before, as Grieve did for the UK in 2019. This particular study found that trends on twitter often matched those found in traditionally compiled surveys, but proportions were often vastly different, and accuracy could not be guaranteed in smaller regions. The accuracy of general trends, however, means that Twitter could provide insight into American regionalist vocabulary as seen online. This possibility led to this project's research question: does Twitter data support the persistence of regional differences in vocabulary in English used online?

1.2 Hypothesis

My hypothesis is that, across the four Census Bureau designated regions of the United States, there are statistically significant differences in the distributions of regional terms for the five chosen concepts, all of which were taken from the 2003 Harvard Dialect Survey. These five concepts are: second person plurals (the varieties being "y'all" and "you guys"), outdoor sales (the varieties being "yard sale" and "garage sale"), gym shoes (the varieties being "sneakers" and "tennis shoes"), soft drinks (the varieties being "soda," "coke," and "pop"), and finally major roads (the varieties being "freeway" and "highway").

2 Methods

Data gathering, manipulation, and statistical analysis were conducted with the R Statistical Software (R Core Team 2016). An API was used to gather information from both Twitter and the Census Bureau, with tweets being taken from Twitter and information on state boundaries taken from the Census Bureau. Each group of search terms (e.g. "y'all" and "you guys") were searched for amongst tweets sent in the prior 6-9 days within the United States, with a random sample being taken to form a new CSV file.

2.1 Data source

The CSV files mentioned in the previous section, which were automatically created by Twitter upon searching for the regionalist terms in question, were then appended with information regarding the state and Census Bureau region of origin for each tweet. These updated files, which contained a list of tweets containing the search terms and their region of origin, were used for all statistical analysis in this project.

2.2 Statistical approach

Maps of the distribution of opposing terms, as well as bar charts demonstrating regional distributions of these terms, were created for exploratory analysis. For inferential analysis, which is at the center of this project, the chi-squared test was used to analyze the relationship between the four regions and the groups of regionalist terms. Each test determined whether there was statistically significant evidence of regional concentrations, or whether a completely random distribution of terms could not be ruled out.

3 Results

Of the five groups of terms tested, only two were found to vary in a statistically significant way, per the chi-squared test. Exploratory analysis also revealed that the distribution of some groups of terms stands in stark contrast to the documented in Vaux's and Katz's dialect surveys. For detailed analysis, these results will be broken down on a group-by-group basis.

3.1 Second Person Plurals

Seen above are bar charts showing the distribution of second person plurals, a feature of English known to vary across the English speaking world, and certainly within the United States. Within the US two phrases

are most commonly used to form the second person plural: "y'all" and "you guys" (since "y'all" is often informally spelt "yall," both were included in the Twitter search). Some other terms such as "you all," "yinze" and simply "you" are also used in various parts of the country, according to the 2003 Harvard Dialect Survey, but in such small frequencies that adequate data could not be gathered. In both dialect surveys consulted in this project, "you guys" reigns supreme in the Northeastern, Midwestern and Western regions of the country, while "y'all" is most popular in the South. This sample from Twitter corroborated these findings somewhat, although "you guys" and "y'all" were tied for the most popular term found in the Western United States. An equal number of tweets containing each term were gathered, rather than a quantity proportional to the relative popularity of each term, and this may contribute to such disparities. Ultimately, we do see that "you guys" is more likely to occur than average in the West, Midwest and Northeast, and less likely to occur than average in the South. Despite the accuracy of these trends, the chi-squared test performed on this data did not establish statistically significant regional variation, instead producing a p-value of 0.2796.

3.1.1 Outdoor Sales

The terms "garage sale" and "yard sale," while well known and used with some frequency in spoken English, barely appeared in Tweets sent in the week prior to the search. While the academic dialect surveys found "garage sale" to be more popular in the Midwest and "yard sale" to be more popular in the rest of the country, the Twitter search found a roughly even distribution across the four regions, with the exception of "garage sale" being more popular in the Northeast. Not surprisingly, the chi-squared test did not support the idea that these two terms were regionally distributed, providing a p-value of 0.8977, meaning it is much more probable than not that the two terms are randomly distributed. This might be attributable to both the inadequate sample size and the fact that the distribution of these terms in spoken English, while very regional, does not fit inside the Census Bureau regions very neatly.

3.1.2 Gym Shoes

Although these findings don't quite line up with what past dialect surveys have found, they do get one thing right: the term "tennis shoes" sees almost no use in the Northeast. According to the surveys consulted in this project, "tennis shoes" should be more common in the South and the West than "sneakers," and the term "tennis shoes" should be more common overall; in this sample from Twitter, however, fewer than 1,000 tweets were found to contain the term "tennis shoes," while at least 1,000 contained the term "sneakers." This may stem in part from the fact that "sneakers" is only one word and thus more efficient to use on Twitter, where tweets can only be so long. In any case, the chi-squared test performed on this data found statistically significant regional variety, with a p-value of 0.002869.

3.1.3 Soft Drinks

Terms for Soft Drinks likely suffered more than any other group from confounding words. Since "coke" is both the term used for all soft drinks in parts of the country and a brand, the distribution of this term likely does not reflect the distribution of speakers who use "coke" to mean any soft drink; the same goes for "pop," which has a variety of possible meanings other than "a soft drink." Thus it may not be surprising that the distribution of these terms in the findings from Twitter bears little resemblance to what has been recorded in past dialect surveys. These findings correctly identify "coke" as most popular in the South and "soda" as more popular in the West, but also found "coke" to be most popular in the Northeast and Midwest, where "soda" and "pop" have been found to be most popular, respectively; these discrepancies are likely attributable to the various meanings of the words "coke" and "pop." It is not surprising, then, that the chi-squared test performed on these findings produced a p-value of 0.2374, which is to say there is no significant difference in the frequency of these terms across the four regions.

3.1.4 Major Roads

The distribution of terms for major roads, "highway" and "freeway" followed previously established geographic trends rather accurately. As attested to in the 2003 Harvard Survey and Katz's 2013 survey, the term "freeway" is more common in the West, while the term "highway" is more common in the North and the

Midwest. There is a discrepancy in the South, however; while the academic surveys find the term "highway" to be more common there, the sample from Twitter finds "freeway" to be more common. There are a number of reasons why this might have occurred, most notably the fact that the word "freeway" has several meanings and is used throughout the country, just not always as a synonym for "highway." Whatever the reason for this distribution, the chi-squared test performed on the data found that it was, in fact regional, with a p-value of 9.317e-05. That is to say, it's extremely unlikely that the terms "highway" and "freeway" are evenly distributed across the country.

4 Discussion

As the results above demonstrate, regional vocabulary as seen on Twitter was not conclusively found to emulate the distribution noted in academic dialect surveys. Because of some potential concerns with the data, such as small sample sizes and terms with multiple meanings, it cannot be said that regional terms are less frequently used on Twitter, however; and because terms for major roads and gym shoes have been found to vary significantly across the different regions of the country, it can be said that, at least in some cases, regional terms do persist in English used online, and are not restricted to spoken English. The persistence of regionalisms may point to English as used on Twitter bearing greater resemblance to spoken English than formal written English, which usually loses almost all traces of regional variety.

5 Conclusion

The only empirical conclusion that can be drawn from these findings is that terms for major roads and gym shoes as seen in tweets vary significantly across the four Census Bureau designated regions of the United States, findings which match regional variations attested to in several past dialect surveys. Whether other regional varieties are less or more pronounced on Twitter is impossible to say, but from these findings it seems as though at least some elements of regional distinctions make their way onto Twitter. Thus, the hypothesis for this project was affirmed in part, though not entirely, and it can safely be said that twitter users do not completely assimilate their vocabulary to General American English when writing tweets.

6 References

Vaux, Bert. "Harvard Dialect Survey." Harvard Dialect Survey, 2003, http://dialect.redlog.net.

Katz, Josh. "Dialect Survey." Josh Katz, 2013, http://joshkatz.co/dialect.html.

Grieve, Jack, et al. "Mapping Lexical Dialect Variation in British English Using Twitter." Frontiers in Artificial Intelligence, vol. 2, 2019, p. 11, doi:10.3389/frai.2019.00011.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.