Linguistics Independent Project

Clara Duffy

April 2019

Abstract:

Research Question How does the kind of restaurant on online review websites change the linguistic tendencies of the reviewer, especially those regarding word choice and frequency?

Possible Answers It is possible that the kind of restaurant being reviewed bears no effect on the text of the reviews. It is also possible that lower priced, lower reviewed restaurants have reviews more similar to those that are higher priced but lower reviewed restaurants, which would indicate that the kind of review a person writes has more impact on their language than the kind of restaurant that is being reviewed. A third possible outcome is that the lower priced, lower reviewed restaurants have more similar textual reviews to lower priced higher reviewed restaurants, which would indicate that the text of reviews has more correlation with the kind (in terms of pricey-ness) of restaurant that is being reviewed rather than the perceived quality (average review quality) that the restaurant has.

Introduction:

This study will primarily be an observational one meant to facilitate further research. This study began as a way to explore the field of computational linguistics more deeply, and understand how factors about who is writing or what they are writing about effects the text that is written. In order to do this, I chose to explore restaurant reviews, as there are discrete measures of what the text is about, namely the pricey-ness rating and the total number of stars, or average review rating, of the restaurant.

Methods:

A data set of the Tripadvisor reviews of 147 restaurants around San Francisco will be used as the sample data to explore this question. Each restaurant has 100 reviews, and this data also includes all of the tripadvisor classifications, but only the ”$” rating, the review text, and the average * ratings out of 5 will be used.

The two independent variables are the ”$” rating and the average * rating out of 5. The dependent variable that will be measured is the frequency of non essential words such as ”the”, ”a”, pronouns, and other short words in the text.

In order to analyze the frequency of these words, I will use python and the nltk library to sort through the corpus, and then I will use R to create graphs of frequency in each category of review (both by number of stars and by number of ”dollar signs” or pricey-ness of the restaurant that is being reviewed).

The four most frequent words in any kind of review will be compared to each other as a percent of total words used. For example, the percent of the word ”the” out of all words used in the text will be compared across review types.

Although correlation is not causation, if there is a correlation between percent of a word’s usage and the pricey-ness or the rating of a restaurant, I will make hypothesis on why this could be, and then leave these hypotheses open for testing in the future.

Data/Results:

The following graphs display the correlation (or lack of it) between word, percent that word is used, and type of restaurant (measured in two dimensions). The dependant variable is the percentage, always measured on the y-axis.

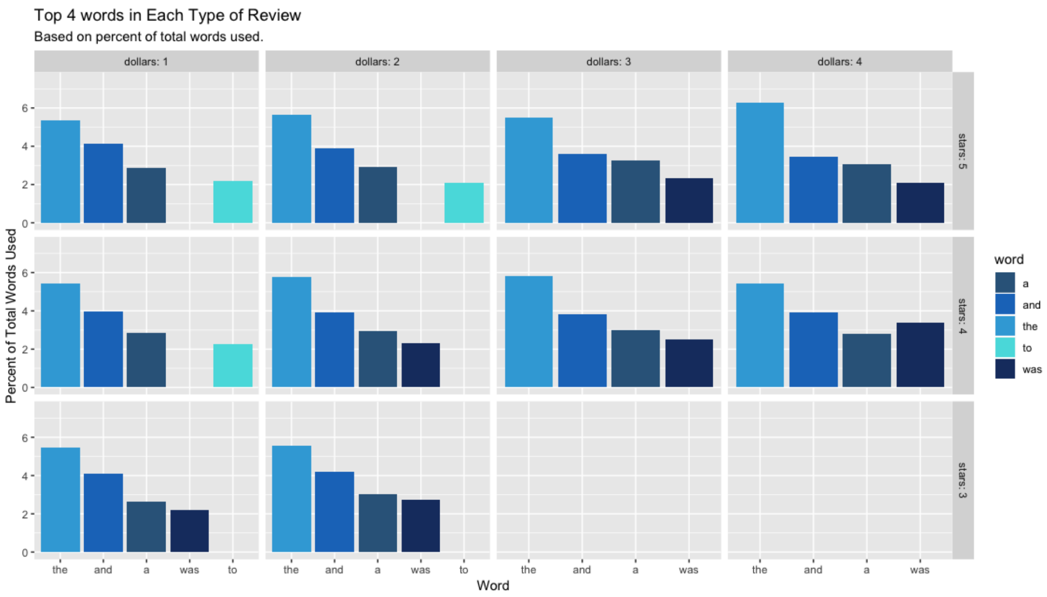

Figure 1: Bar Graph

From this graph, the conclusion that the same helping words are used in similar percentages across all types of review can be established. The downward slope of percent use for these most frequent words is similar in all of these different categories of restaurant. One outlier here is the category of the ”best” restaurants, where the prices are the highest and the rating is the highest, and the word ”the” is used the highest percent of the time, and the other most frequent words are much less frequent. One reason for this might be that these reviews tend to describe the food or restaurant in superlatives, such as ”the best ...” or ”the most ...”, or because these reviews use loftier language to describe an elevated lifestyle, where everything is uncommon, ie ”the [insert unique kind of food]...”. This could be an interesting place to explore what linguistic patterns this type of review tends to follow, and why it is more present in the reviews of the best and pricey-est restaurants.

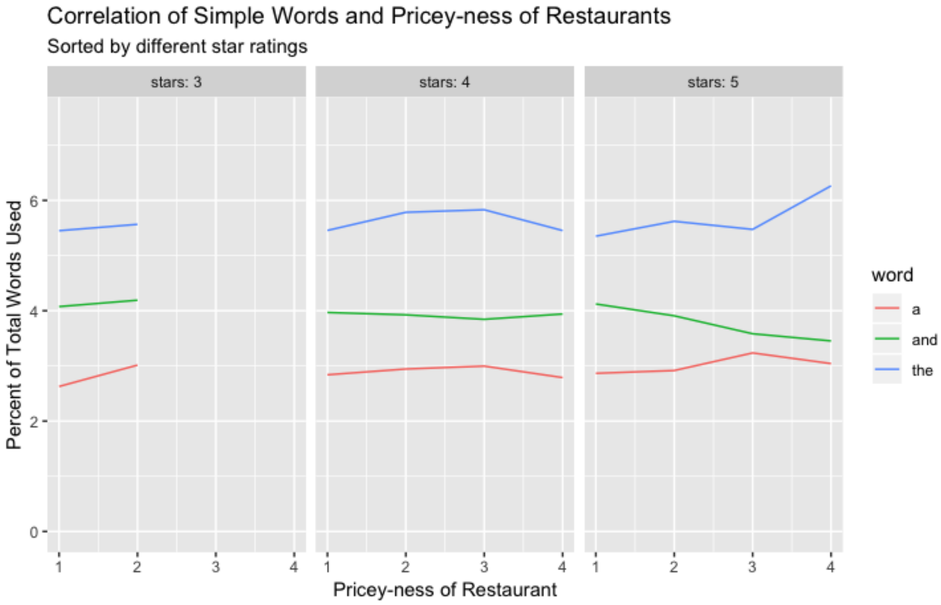

Figure 2: Line Graph: sorted by star rating

On this graph one can observe the tendency for the reviewer to use each of the three most common words ”a”, ”the”, and ”and” in their reviews, with the dependent variable being the pricey-ness of the restaurant. There are 3 different subgraphs, one for each of the different star ratings of the restaurant. From this graph, it is easiest to see the areas in which the corpus lacked data. In the left segment of the figure, the graph of restaurants rated as 3-star on average, there are only two data points for each word, which means that only the lower priced restuarants were ever averaging below a 4-star rating. Other than this, in all of the graphs in this figure, there do not seem to be significant correlations between the pricey-ness of the restaurant and the percent use of these three words, as there is no strong upward or downward trend that is consistent.

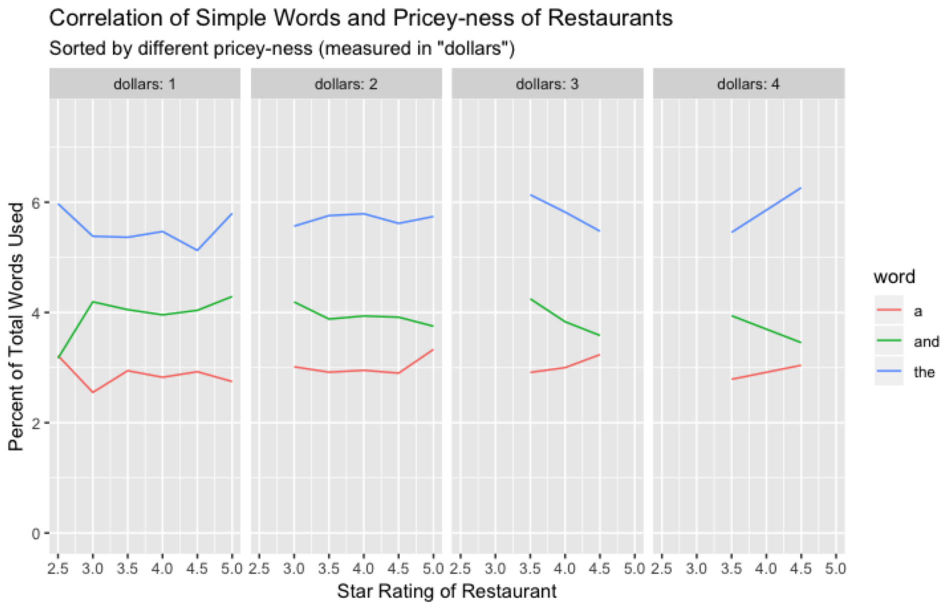

Figure 3: Line Graph: sorted by price

On this graph one can observe the tendency for the reviewer to use each of the three most common words ”a”, ”the”, and ”and” in their reviews, with the dependent variable being the star-rating of the restaurant. There are 4 different subgraphs, one for each of the different price levels of restaurant. This graph is slightly different from the 2 before it because the data it uses is slightly more specific. The ratings rounded to every 0.5, rather than to every whole number, which increases the amount of specificity of the x axis. The most interesting of the graphs in this figure is the leftmost one, where the price rating of the restaurants in this category are the lowest. This is the most informative part of the graph, because it has the widest valid range of average star ratings. From this, it is clear that the percentages of usages of the words ”a,” ”and,” and ”the” do not correlate with eachother, and do not trend in one way or another as the star rating of the restaurant goes up. This indicates that the average star rating of the restaurant has no bearing on the use of these common words in the cheapest category of restaurants. In the right graphs in this figure, the graphs with less overall data (3 dollar signs and 4 dollar signs) tend to have a strong apparent correlation, that as the average rating goes up, the use of the words tend to go up or down (depending on the word and the price category of the restaurant). From these two observations, it is possible to conclude that review text for restaurants that are very inexpensive tend to have no pattern of use of frequent functional words, whereas restaurants that are more expensive have trends of type of text in their reviews based on the average star rating of that restaurant.

Discussion and Conclusion:

With the above graphs, not many conclusions can be drawn. The most interesting results come from the last graphs, figure 3, and revolve around the correlation between the pricey-ness of restaurant and the amount of correlation of percent of functional words within the star rating for each category of restaurant, namely that the lower priced restaurants have less correlation and the higher priced restaurants have more correlation in terms of percent of functional words used in reviews. Since correlation does not indicate causation, no conclusions about causation can be made with certainty from this data.

Though this data does not lead to many conclusions on its own, it does lead to many other questions for future exploration. For example, is there a linguistic explanation of the higher frequency of the word "the" in the priceyest and highest rated restaurants? Does the use of the word "the" change in this kind of review, which is likely all positive? Since the reviewer knows the average ratings of the restaurant they are reviewing already, does this effect the kind of review they leave for the restaurant? How does the number of words used in a review overall change based on the kind of restaurant being reviewed? There are also many other non-linguistic questions one might explore from this data: Why is it that less expensive restaurants have a wider range of reviews, and more expensive restaurants are mostly rated with the highest average star ratings? Or, an interesting computational question, is it possible to identify false or fictitious reviews solely by looking at the connectives or functional words in the review text? If able to continue exploring with this data set or a similar one, I would begin my exploration with one of these questions.