Click Sounds in the Spanish of Texas

Rebecca Twite & Stephen Robert

Introduction

In this research, Dr. Derrin Pinto and Dr. Donny Vigil aim to find evidence that the use of clicks, known as markers or fillers in conversation, serve different purposes, are used in different situations, and are used in various frequency, depending on a person's age, gender, place of birth, Spanish speaking ability, use of Spanish, and education. The data used in this research has been extracted from a corpus of interviews conducted of bilingual Spanish and English speakers in Texas. The data collected and analyzed includes 36 speakers, randomly selected with consideration for age, gender, place of birth, Spanish speaking ability, use of Spanish, and education, aiming to prevent excess skew in the data.

Research & Data

The data analyzed includes men and women who were born in both the United States and Mexico, between the ages of 18 and 74, with various educational backgrounds between people with less than 6^{th} grade through more than 12^{th} grade separated into three groups. Domain is used to measure how each participant used Spanish in up to eight domains, including educational settings, the home, and workplaces. The data also includes a speaking ability variable measured on a scale from 1 to 5, with self-reported ability varying from not very good to excellent. Also, the data includes a placement variable used to determine where participants used clicks during their conversation and a function variable used to determine how each participant used clicks. Function is more complicated than placement, as any single click may be used in up to two out of five ways, with options for opening, continuing, search, stance, and reformulation. Words per click was also provided in the data, as a ratio of the number of words in the interview and the number of clicks recorded.

Previous research on clicks has indicated that the use of clicks seems to be idiosyncratic, as not all Spanish speakers click and those who do, use clicks with varying frequency. This research has provided three hypotheses to test. All three share the assumption that gender, age, place of birth, domain, speaking ability, and education will have no statistically significant different effects on each of the three response variables; the ratio of words per click, the function of clicks, and the placement of clicks.

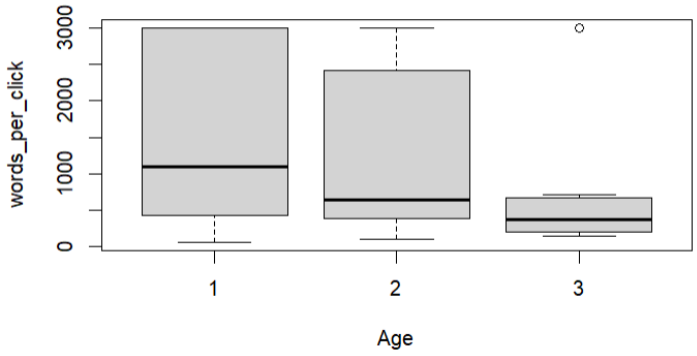
The statistical model chosen, non-parametric t-tests, allowed the data to be analyzed without making major assumptions about the distribution of the data, as much of it is skewed and may be more difficult to analyze fairly and accurately by making assumptions about the data.

It was possible to maintain and analyze the entire dataset almost as provided, however our early analysis indicated that the participants who didn't click provided us with a challenge which we avoided by flipping the ratio to be clicks per word. This allowed us to keep the dataset intact and include the non-clickers' demographic data in the analysis, as it may still include information which impacts the analysis and may contribute to incorrect or incomplete results if removed. For the first hypothesis, which deals with the ratio of words per click, we chose to use the ratio of clicks per word due to the issue of non-clickers skewing the data in the original variable.

Models & Results

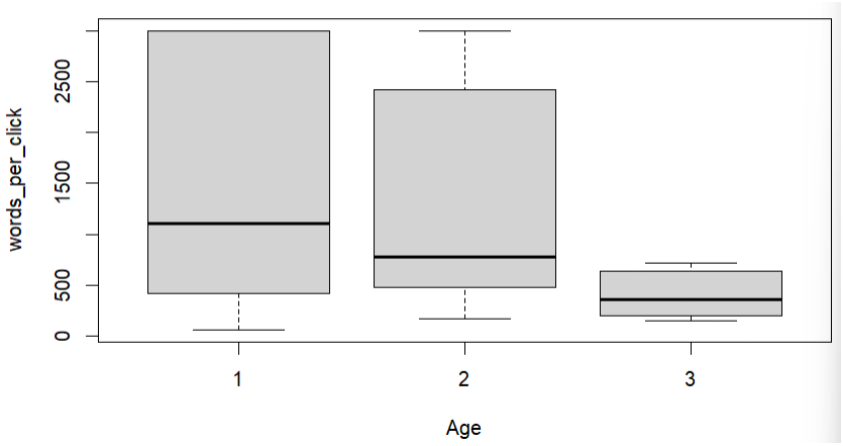
Hypothesis #1: There will not be a statistically significant difference in the ratio of [clicks per word] based on the following variables: gender, age, place of birth, domain, speaking ability, and education.

```
            Df   Sum Sq Mean Sq F value Pr(>F)
Age          1  5682915 5682915   4.818 0.0351
Residuals   34 40103081 1179502
```



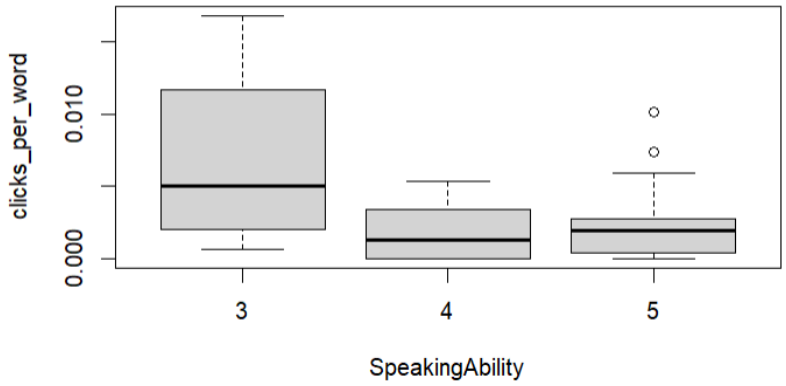
First, our brief analysis of words per click indicates that age is a factor in determining how often people will click. The p-value for the age variable and words per click was 0.0351, indicating that there is a statistically significant relationship between the two variables in this dataset. The above side-by-side boxplot indicates that the mean number of words per click used by each age group decreases by a wide margin as we look at older populations. Additionally, the spread between the minimum and maximum number of words per click decreases to be small when looking at the age group 3, which spans 51-74 years old, with one outlier at almost 3000 words per click.

```
            Df   Sum Sq Mean Sq F value Pr(>F)
Age          1  7182576 7182576   6.736 0.0143
Residuals   31 33053962 1066257
```



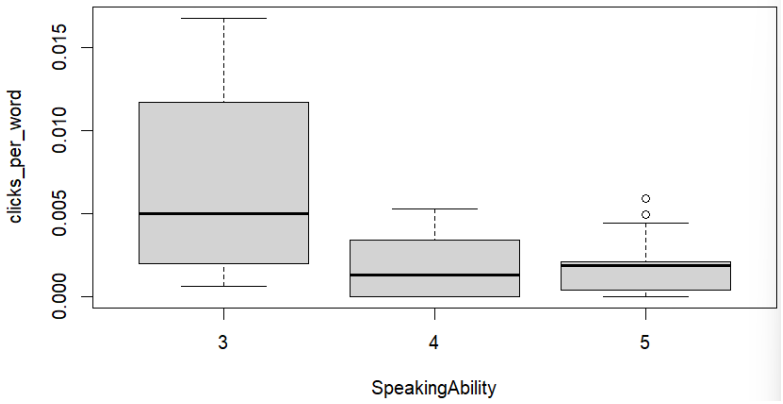
Removing the outlier in age group 3 further decreased the p-value to 0.0143 between age and words per click, supporting the previous statement that age is a statistically significant predictor of the number of words said by the participant for each click used.

```
                Df    Sum Sq   Mean Sq F value Pr(>F)
SpeakingAbility  1 0.0000386 3.863e-05   3.582 0.0669
Residuals       34 0.0003667 1.078e-05
```



Our analysis of clicks per word indicated that there may be a statistically significant relationship between how often people click and their speaking ability, with a p-value of 0.0669. The data only included people who ranked themselves as speaking Spanish moderately well through very well, as indicated by the presence of speaking ability ranks 3, 4, and 5. As expected, the number of clicks per word decreases as we look across the three rankings, with more people who are less fluent using clicks more often than their more fluent peers.

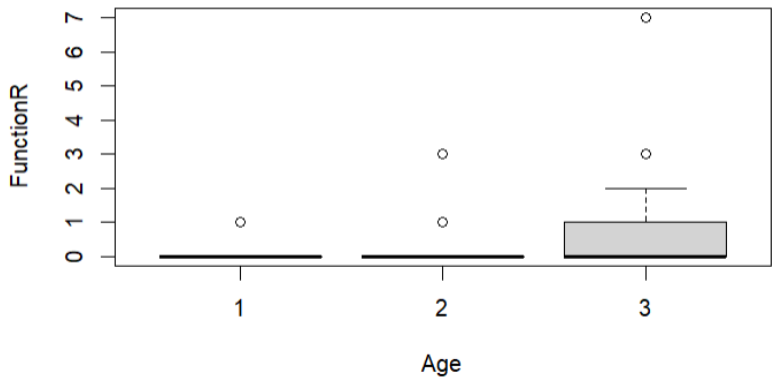
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SpeakingAbility	1	5.538e-05	5.538e-05	6.486	0.0161
Residuals	31	2.647e-04	8.540e-06		



As seen above, removing the outliers did improve the p-value for this analysis, as the p-value for speaking ability in relation to clicks per word is now 0.0161, indicating that this relationship is statistically significant. This analysis did add two new outliers, but they are closer to the

Hypothesis #2: There will not be a statistically significant difference in the functions of clicks based on the following variables: gender, age, place of birth, domain, speaking ability, and education.

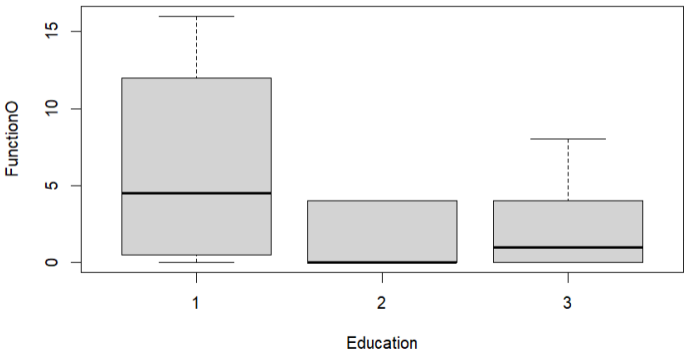
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	5.04	5.042	2.86	0.0999
Residuals	34	59.93	1.763		



The above side-by-side boxplot indicates that most people across all three age groups are likely not regularly using reformulation clicks, indicated by Function R, although people who are in the oldest age group, ranging from 51-74 years old, may be more likely to use clicks for reformulation. This analysis has a p-value of 0.0999, which does not meet our 0.05 threshold for a 95% confidence, further research with a larger dataset may find that there is a strong relationship between age, especially older participants, and their use of clicks for reformulation.

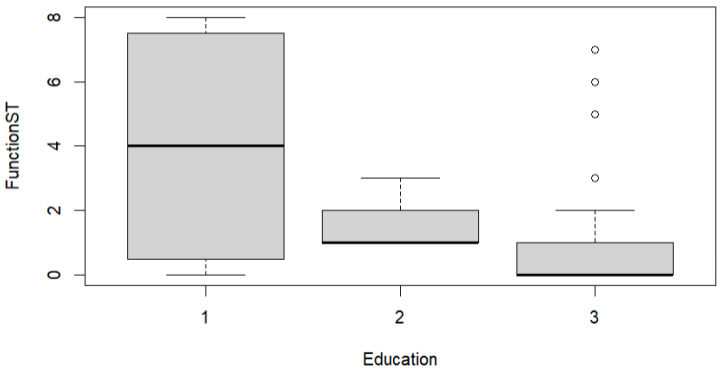
By removing the outliers, we found a few more interactions which are statistically significant for this dataset, as follows.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Education	1	45.3	45.33	4.055	0.0534
Residuals	29	324.2	11.18		



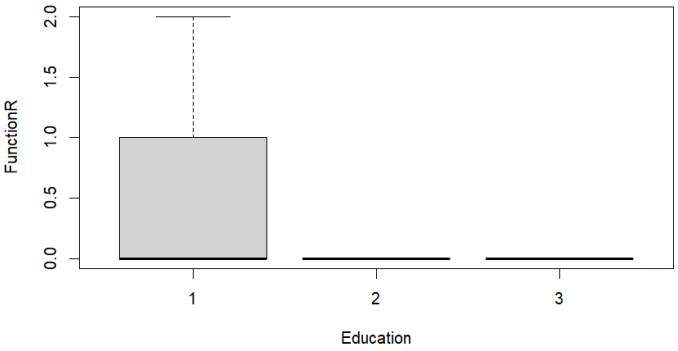
The above boxplot shows the relationship between clicks used as opening operations and just barely misses the threshold of 0.05 but may with some additional data.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Education	1	22.28	22.281	4.284	0.0475
Residuals	29	150.82	5.201		



After removing the previous outliers, education level and stance clicks now has a statistically significant relationship, with a p-value of 0.0475, even with an additional four outliers in the highest education group. Overall, we see a decrease in the mean number of clicks used for stance as education increases.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Education	1	0.643	0.6428	5.774	0.0229
Residuals	29	3.228	0.1113		



Removing the outliers we saw earlier, the p-value for the relationship between clicks used for reformulation and education level indicates significance, at 0.0229. All people who used reformulation clicks in this revised set of data were in the least educated group, possibly having completed 6th grade.

Hypothesis #3: There will not be a statistically significant difference in the placement of clicks based on the following variables: gender, age, place of birth, domain, speaking ability, and education.

Our analysis provided no evidence to reject hypothesis 3, although the lower p-values of Age and Speaking Ability with both Placement M and N may indicate a statistically significant relationship in a larger dataset, however our limited data prevents us from rejecting the hypothesis. The results of our analysis regarding Placement F are likely unreliable given that there are only four clicks categorized as Placement F, but it may be possible that there is a relationship between any of the predictor variables and final placement of clicks.

Conclusions

With the data that we were working with, we only found a statistically significant relationship between age and words per click. However, there are a few relationships which may be statistically significant, as we saw relatively low p-values for many more interactions, but we

were limited by the data we had available. Future research would benefit by using a larger dataset, which may indicate that there are more statistically significant interactions across the board. As we were only able to indicate that there could be a slight relationship between Placement M and Placement N with any of the predictor variables, it would be interesting to focus on the placement of clicks across a wider sample of Spanish speakers, as age may contribute significantly to both Placements M and N, and place of birth may contribute to Placement N.

The lack of diversity in the speaking ability category, as no one participating in our data study ranked themselves as speaking Spanish either 1 or 2, indicating less than moderate Spanish abilities, may be a factor in making it an indicator of how often a person clicks, although it would be interesting to have a wider variety of speaking abilities and to see whether or not there is a statistically significant relationship across the board of people who speak Spanish in Texas.

Although the dataset includes a wide variety of people with various educational backgrounds and domains, there was not a statistically significant relationship between education or domain and any of the response variables we considered. The group was split fairly evenly between men and women, but gender was not a valuable prediction measure, as we may have expected.

Overall, removing the outliers helped us to find more significant relationships between the predictor and response variables, although we were dealing with smaller datasets. It would be beneficial to have a larger, less skewed set of data.