Milestone 3

Stat041 Final Project

Dohyun Lee, Satyaa Suresh

Introduction

Language prevails as a unique trademark of the human species and the bedrock of human development. Although most species possess the ability to communicate, humans remain the sole organism to wield the gift of intricate and thorough language, a tool that allows us to sculpt the full narrative characterizations and social coordination that has provided the means to dominate the planet. In essence, individual words remain the components that comprise the apparatus of language. The vocabulary of an individual formulates the building blocks of their ability to comprehend and express ideas. Without a precise system of words, the profoundness of language would degrade merely into the brute rudimentary communication found in other species. Indeed, it is quite crucial to recognize the role of vocabulary as a catalyst in the advancement of the human species.

Humans quickly learn words and assemble their vocabulary from a very young age. In fact, the most rapid development in vocabulary takes place during childhood, where humans tend to efficiently interpret and absorb the words they hear from adults through conversation and play. The growth of vocabulary during this childhood period hinges deeply on consistent nurturing and interactivity with peers and adults. Typically, the first few months of young childhood involves establishing the ability to vocalize pleasure and displeasure. This would include simple gestures such as laughing or crying. The next few months usually involve communicating through physical action as well as attempting to repeat words heard from the parents. This time period, six to eleven months, also includes the first words. However, the most rapid and parabolic rate of vocabulary development typically starts around sixteen months. As such, the basis of our study and the trends that we depict will begin at this age. Considering the overarching significance of human language previously outlined, as well as the consequential nature of early childhood in cultivating it, this project aims to explore the vocabulary of children through analyzing the trends in their development of language. The project will also look into different dialects across the world and in turn attempt to identify differences in development between each one. These distinctions between dialects may provide insight into cultural differences that influence the maturation process of language in children. In essence, we regard these patterns as enormously meaningful as it concerns the foundation of human relationships.

To reach the goal of identifying trends in the development of language during childhood, the project will aim to incorporate data and visualizations that showcase the growth of vocabulary by age. We will attempt to do so by including several marquee statistical

techniques and visualizations such as multiple regression and scatter plots respectively. Particularly, we will investigate the vocabulary trends as it relates to different languages. These trends would also include the respective percentiles and in turn, provide readers with an understanding of the distribution of vocabulary growth amongst children. A test of proportions will also be included to identify differences in means between languages. In addition to simply revealing the overall trends, we will also aim to explore several factors that could contribute to the variance in vocabulary development. The variables we have in mind include key factors such as parental education, the language being analyzed, birth order, and gender. Considering the drastic effect each of these factors is known to have on development in general, it is safe to assume that they would also influence the progression of vocabulary. Essentially, we aim to explore the effects of these variables and rationalize their existence.

On top of delving into the trends and rate of vocabulary development as a whole, we also aim to scrutinize the role linguistic differences play on children retaining certain words in different languages. For instance, we suspect that there exists words or gestures that may be relatively ubiquitous amongst young children in Western countries, but quite rare in the vocabulary palette of children in Eastern countries. We aim to recognize these words or gestures and attempt to provide a rationale for them. We will also attempt to provide significant evidence for these developmental differences identified between languages and cultures through conducting meaningful statistical tests. Specifically, we will likely utilize a two-sample test of proportions to provide proof for a statistically significant difference in the portion of children knowing a particular word in one language versus another. Essentially, identifying these linguistic differences could yield valuable insights into the expressions and gestures of children from different cultures and backgrounds are exposed to.

For the sake of transparency, the data involved in this study comes from Wordbank, a Stanford based research directory that contains a database of the vocabulary development of children. The database contains data from tends of thousands of children across dozens of languages. Wordbank compiles responses from norming studies but also includes data that individual researchers have contributed from their own projects.

Paper Outline

Methods

- Question 1: We decided to use a boxplot to explain the relationship between vocabulary and age in months because it provided an intuitive visualization of how vocabulary size increases with age. It also includes the quantiles, which we thought was a good detail to have.

- Question 2: To see how birth order affects vocabulary size, we fit two linear models on a scatterplot that we filtered to only include the first and second birth orders, since every other birth order below the two lacked a large enough sample size to provide statistically significant results. We colored the lines orange and navy blue so that the contrast makes the comparison stand out.

- Question 3: Like with our second question, we fit a multiple regression model, filter-

ing for mother's education. Out of "Primary, Some Secondary", "Secondary", "Some College", "College", "Some Graduate", and "Graduate", we chose to filter the data by "Secondary", "College", and "Graduate" because the differences among these groups were more pronounced. Along with this, we also filtered the data by first and second birth orders. (Talk about the equation we produced for the multiple regression model)

- Question 4: Among each age group, we wanted to compare the proportion of children who retained a certain word across every language in the dataset. To do this, we created a faceted scatterplot for every language and fit a curve using LOESS regression to better highlight the growth in proportion of children who were able to produce that word. "American Sign Language" was too long to fit on the plot, so we shortened it to "ALS". To make it easier on the eyes of the audience, each language's data points and curve were colored differently. (Talk about the t-test conducted). To see the differences in the proportion of children of different ethnic backgrounds we took the word "chocolate" and conducted a two-sample t-test, comparing the mean proportions of children from Croatia and South Korea. In the dataset, the age ranges from 8 to 30 months, but we cut it to 8 to 20 months because by looking at the data, at around 30 months the proportion is close to 1 for every language for most words, which would skew our data and produce statistically insignificant conclusions.

Results

- The first graph included in the project reveals the relationship between vocabulary and age for English speaking children. As one would expect, vocabulary size increases with age. The box plots also give the viewer an idea of the distribution of the vocabulary size by age.

- The next graph involves a scatterplot that also relates vocabulary size to age. However, this plot filters out for birth order. The plot reveals that first borns typically have a greater vocabulary size for every age. Considering that children typically develop their vocabulary from parental exposure, this may suggest that parents pay more attention to their first born children while they're infants than others.

- Similar to the previous graph, Graph 3 relates vocabulary and age while accounting for maternal education. As expected, the graph reveals that maternal education has a positive impact on a child's vocabulary size. A more education mother might possess a larger vocabulary palette themselves, which the child would then absorb through nurture. Alternatively, or additionally, the education level of a mother could correlate with their intelligence, which an infant might inherit. This inheritance of intelligence might allow a child to accrue a greater vocabulary at a faster rate.

- Alongside the two graphics, we also included a multiple regression analysis. The regression model included birth order, maternal education, and an interaction term involving both maternal education and birth order. Disregarding age, the regression output suggested that birth order has the greatest effect on vocabulary size. It indicated that being born second had a negative effect of -32 on vocabulary, holding all other variables

constant. Furthermore, the interaction term suggested that being born second, as well as having only a secondary maternal education leads to a 62 (-32+-30) point decrease in vocabulary size, on average.

- The final graph involved a facet grid that showcased that mean proportion of children who knew of a certain word for various languages. In our case, the chosen word was chocolate. As we saw a substantial difference in the proportions for chocolate between Croation and Korean, we decided to run a two-sample test of proportions to verify whether there was any significant difference between the two groups. The t-test resulted in a p-value of 0.02, which verified our hypothesis that there was a significant difference in the proportions of the two groups. The reasoning behind this is quite unclear. We speculate that since Croatia is a massive exporter of chocolate, there exists a substantial cultural attachment to it. This attachment could cause children to garner more exposure and thereby attach it to their vocabulary sooner than children from other countries or languages.

Discussion

- Some of the languages have limited data, which made it difficult to conduct tests involving them. For instance, for chocolate, we would have liked to done a t-test comparing the proportion in English with other languages, since everyone in the class is familiar with English. Surprisingly, however, there was a lack of data points in English for chocolate. Perhaps we overestimated how common it is for Americans children to understand or produce the word.

- Considering that we have a limited number of variables, we also can't be absolute certain of the magnitude of the coefficients in the multiple regression analysis. For instance, the regression model may suffer from omitted variable bias, which could induce bias in the coefficients and skew the results.

- Future analysis could involve continuously tracking these vocabulary sizes as the children get older. This could provide insight into the influence of the education system of country on the growth of vocabulary. It could also reveal with discrepancies that existed between infants were maintained through adolescence and adulthood. For instance, we could observe whether or not first-born children maintained a greater vocabulary size than their later born counterparts.

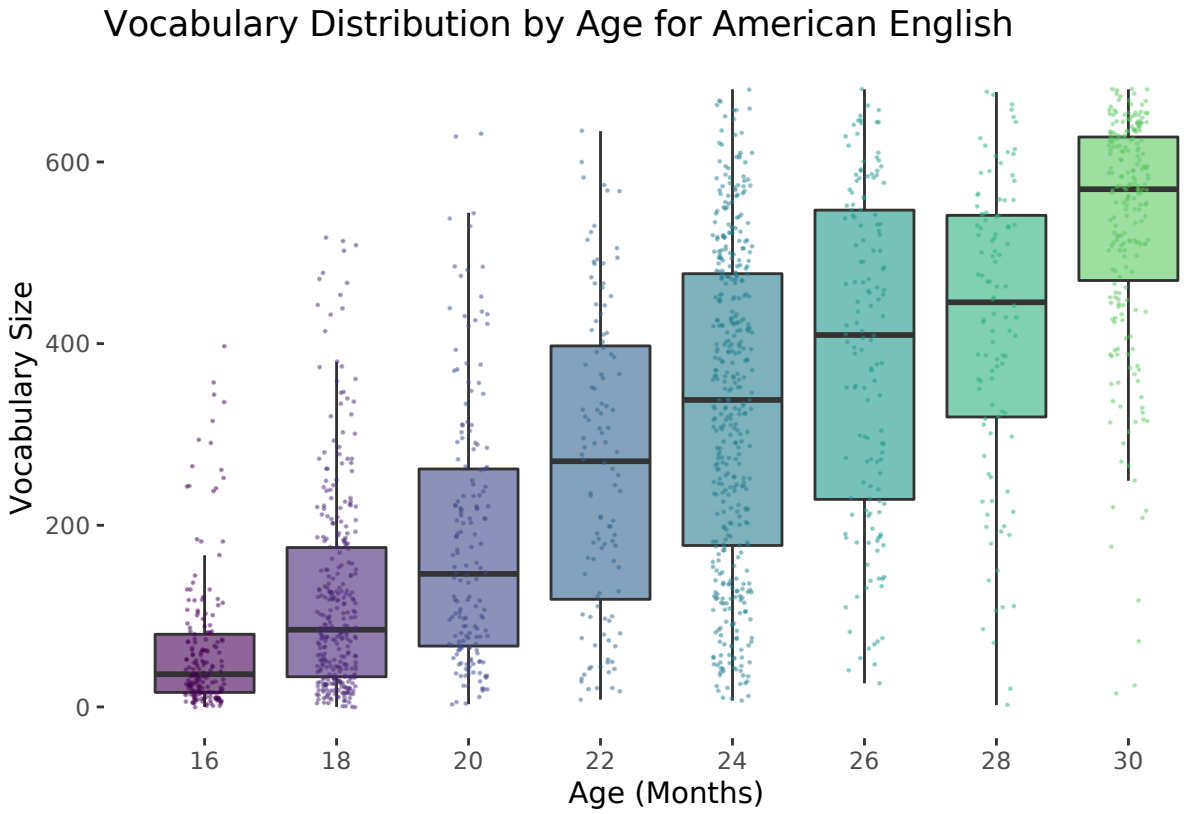

Figure 1

```
## `geom_smooth()` using formula 'y ~ x'
```

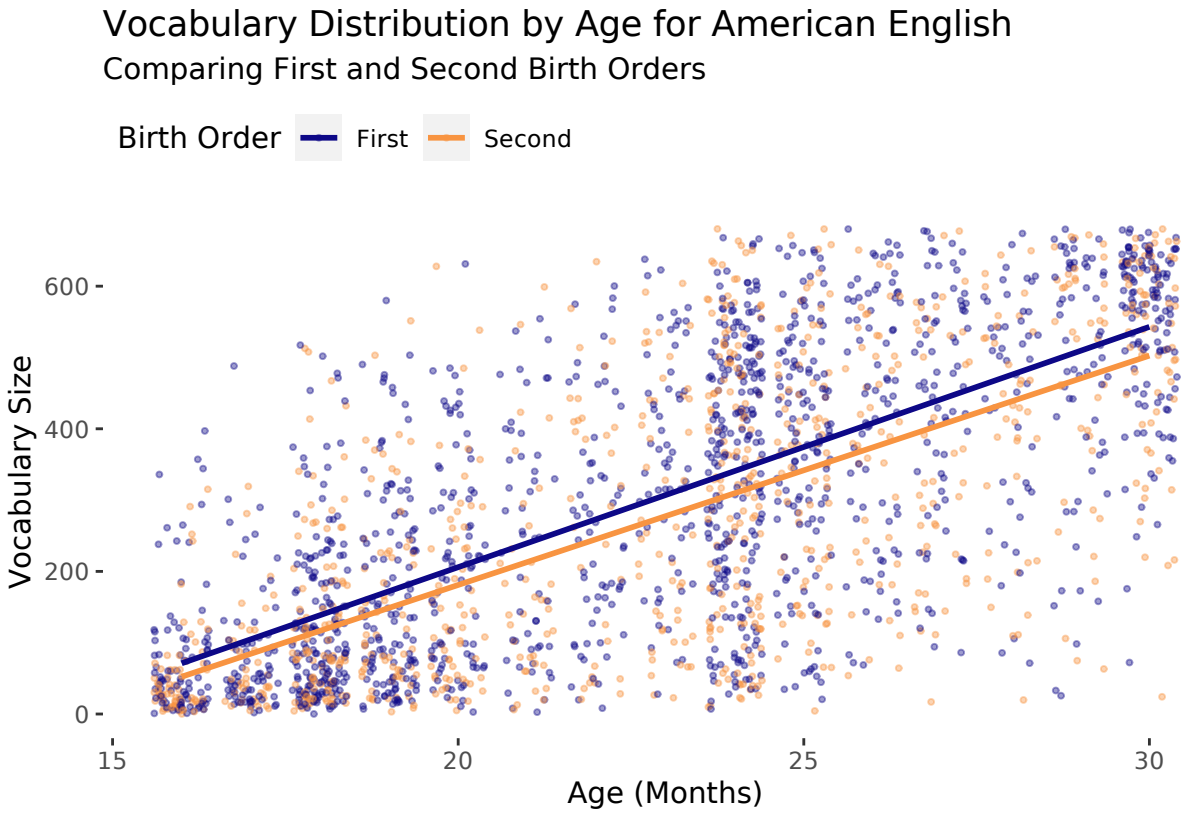

Figure 2

```
## `geom_smooth()` using formula 'y ~ x'
```

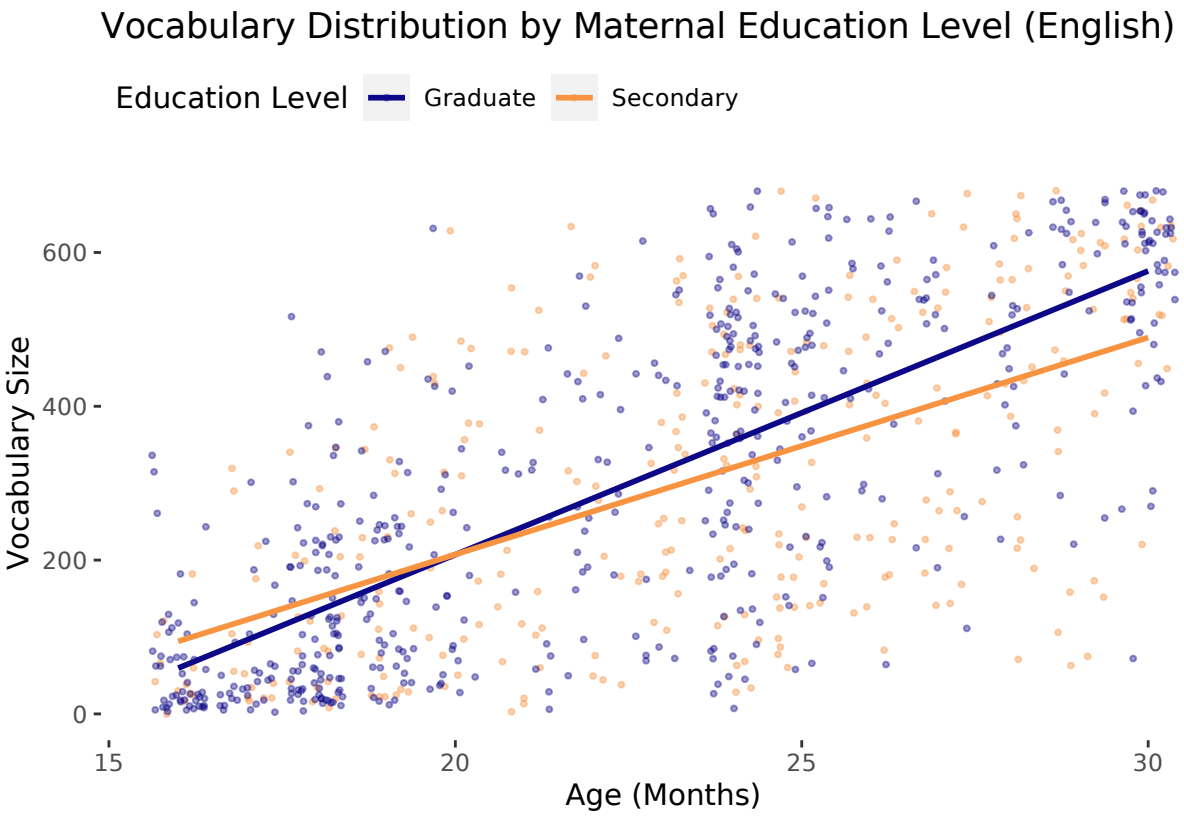

Figure 3

term	estimate	std.error	statistic	p.value
(Intercept)	-460.30	27.82	-16.54	0.00
age	34.01	1.19	28.67	0.00
eduSecondary	-3.76	13.94	-0.27	0.79
birth_orderSecond	-17.02	13.28	-1.28	0.20
eduSecondary:birth_orderSecond	-45.82	21.30	-2.15	0.03

Variable	Meaning
eduGraduate	Graduate School
eduSecondary	Secondary School
Birth_orderSecond	Second Born

```
## `geom_smooth()` using formula 'y ~ x'
```

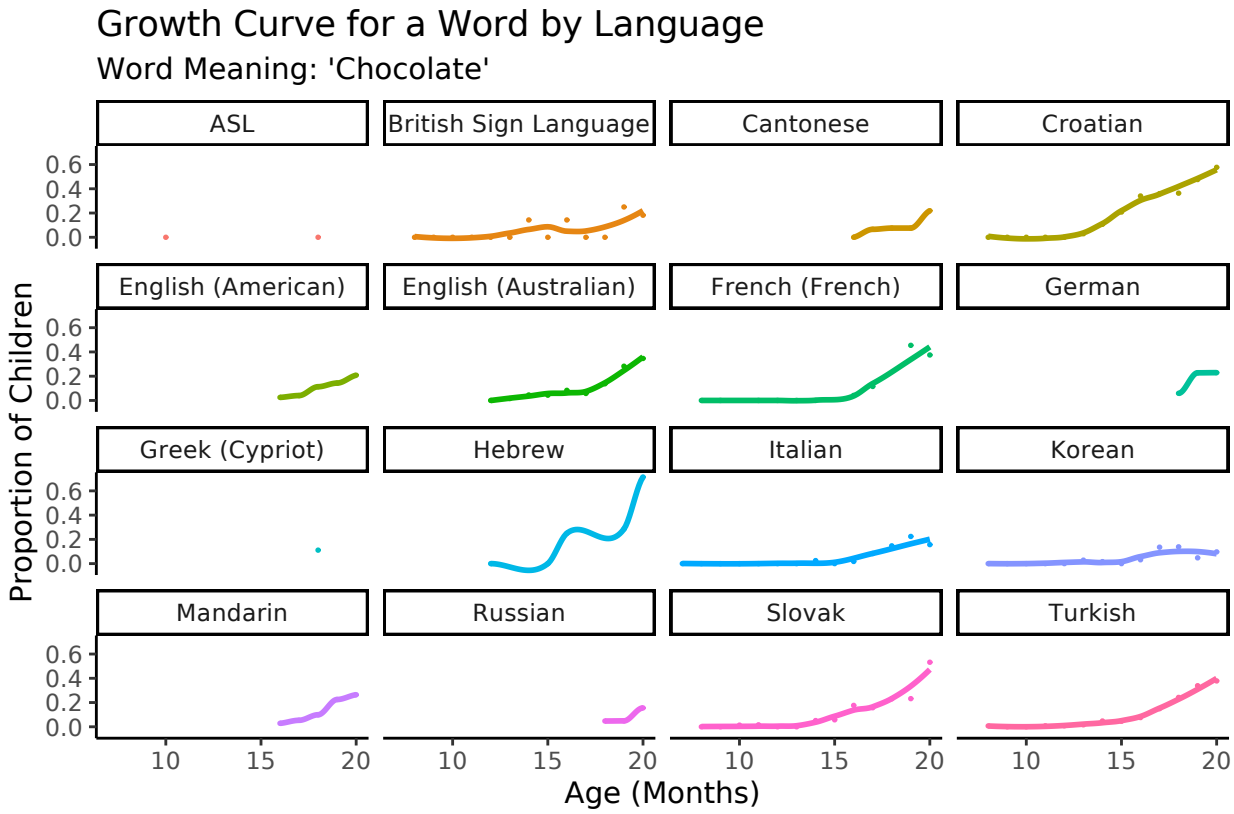

Figure 4

```
##
##  Welch Two Sample t-test
##
## data:  ling_data_croat$prop and ling_data_kor$prop
## t = 2.5271, df = 13.472, p-value = 0.02472
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02241431 0.28018638
## sample estimates:
##  mean of x  mean of y
## 0.18938949 0.03808915
```