An Investigation of Early Childhood Vocabulary Development

Dohyun Lee, Satyaa Suresh

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 or linguistic 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.

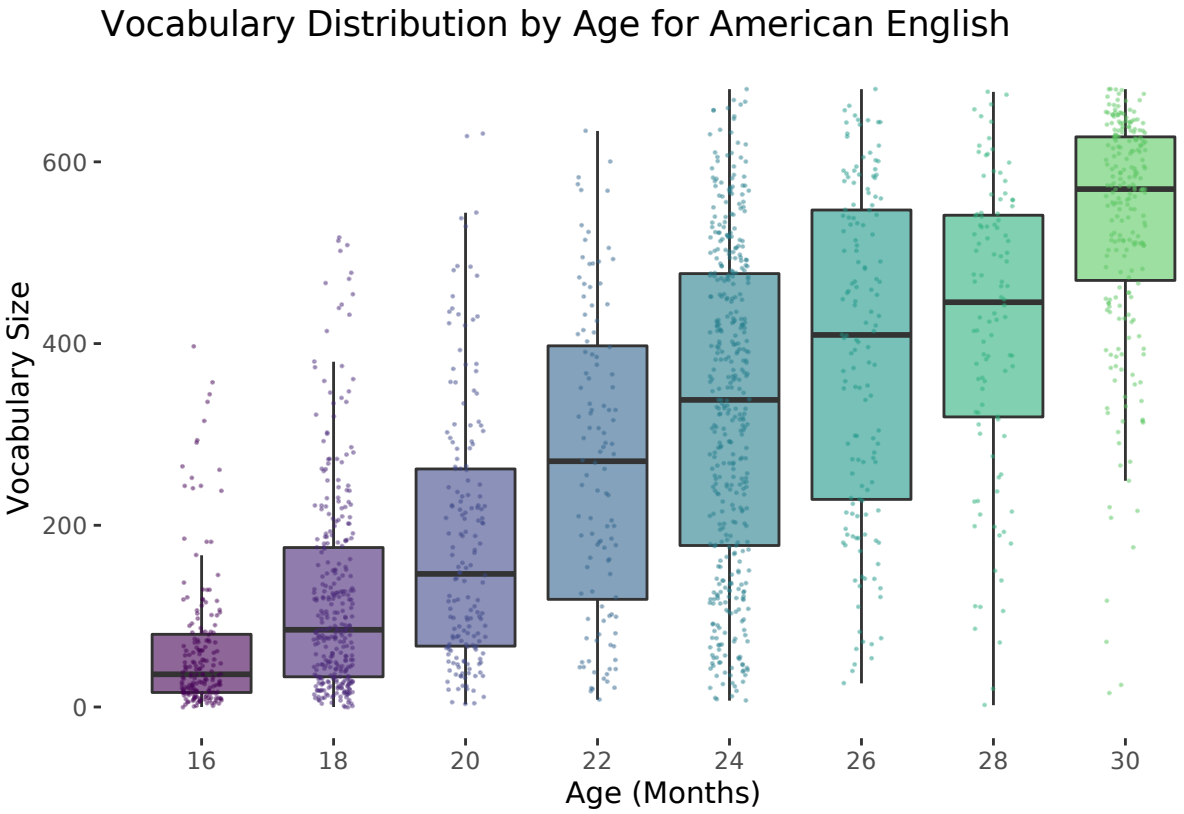

Figure 1

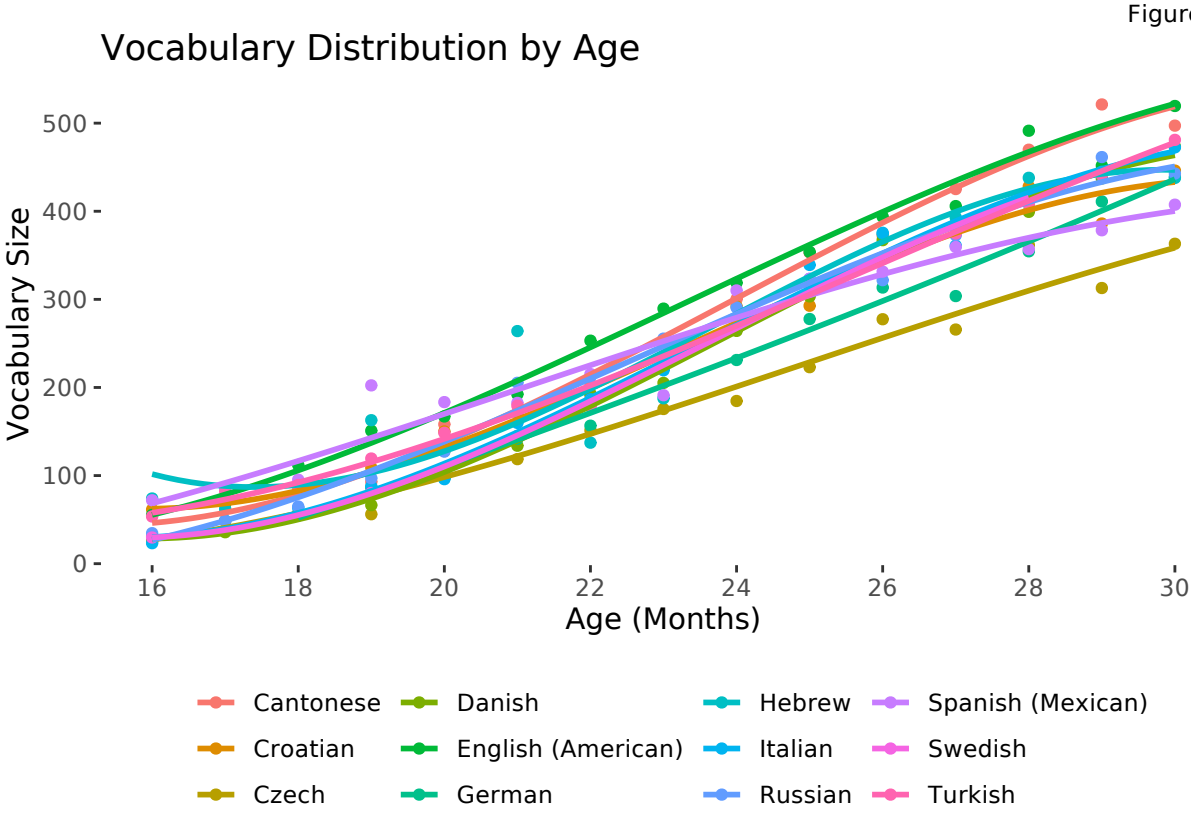

Figure 2

The first question of interest involves clarifying the general rate of vocabulary development during early childhood. Essentially, prior to scrutinizing the various factors that could influence development, we figured it appropriate to first design a clear portrayal of the linguistic ripening in itself. We will first use American English as an instrument to capture this foundational trend. Figure 1 involves a series of boxplots that showcase the overall trend in early childhood American English vocabulary development by age. The figure captures both the median vocabulary production size for each age, as well as the general distribution of the sizes. On the other hand, Figure 2 excludes the focus on distribution to provide a more clear portrait of the changes in the rate of development. In addition, Figure 2 also includes multiple languages to verify whether different languages share a common feature related to the developmental trend. Overall, the trends captured in the plot aligns relatively closely with the intuitive sense of early vocabulary development; they reveal that most children tend to speak at most only a few words before their first birthday, but that production accelerates going into the second year. A deeper look into the progression reveals that this growth at some point begins to decelerate. When looking at development between twenty-four and twenty-eight months, it is apparent that the vocabulary size increases at a slightly decreasing rate than in the few months prior. This somewhat logistic relationship serves as a microcosmic function of the vocabulary development in humans in general. Indeed, as a whole, the development in vocabulary is the most rapid during early childhood and then begins to level off in growth in adolescence and adulthood. Therefore, considering the consequential nature of early childhood in constructing the linguistic palette of an individual, we consider it a compulsory and provoking to analyze.

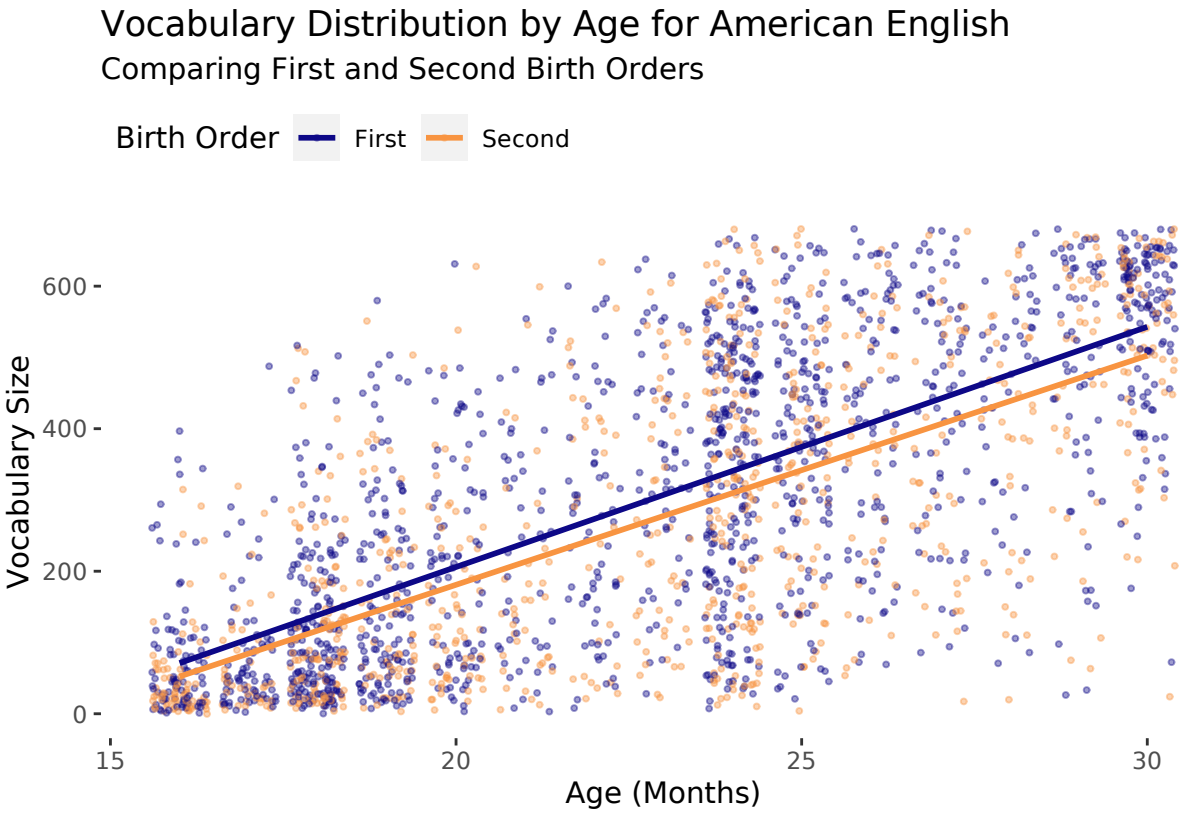

Figure 3

Now that we have an idea of the general trend associated with early childhood vocabulary development, we will next begin to explore various questions pertaining to factors that influence the structure of the aforementioned models. The first factor we decided to investigate included the birth order. Essentially, we desired to gauge whether birth order had a significant impact on vocabulary size. To accomplish this, we fit two linear models on a scatterplot, as shown in Figure 3. The linear models were filtered to include the first and second birth orders. Although the dataset included children who were later born, we figured it more appropriate to only include the first and second born children. Firstly, there was not enough data on later born children to produce a convincing and statistically significant model. Secondly, we assumed that the inclusion would be quite redundant; the effect of being second-born was almost identical to the effect of being later-born and thus the inclusion would likely only serve to hamper visual clarity. As for the visual features, we colored the first borns as navy blue and the second borns as orange. We assumed this color contrast

would allow for the comparison to stand out.

The linear model revealed that first born children have a consistently higher vocabulary size for each age. Indeed, the regression line for the first born children stays completely above the regression line of second born children, which alludes to the statistically significant difference between the two that will be confirmed later on. As for discerning the reasoning behind this difference, we assume that it could correlate with parental behavior. As previously mentioned, children typically accrue linguistic skills through external stimulation. Naturally, parents are the most likely figures to supply this stimulation. During their early years, or at least their first year, first-born children do not have to share parental attention or resources with their younger siblings. Thus, based on the luxury of having their resources maximized, it can be assumed that first-born children have an advantage over later-born siblings during early childhood. In essence, this advantage could contribute to the linguistic gap observed in the dataset. Of course, we do not desire to extrapolate beyond the dataset and thus cannot make the claim that this trend persists beyond early childhood.

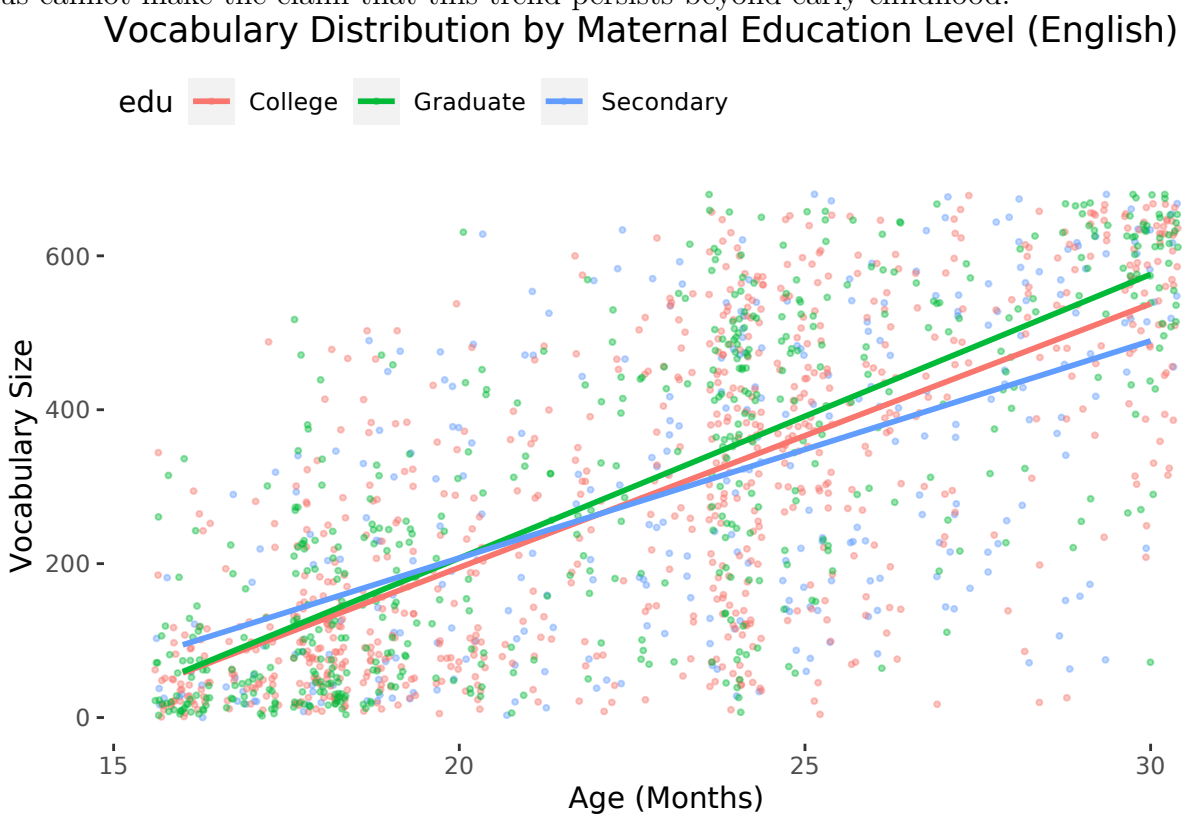

Figure 4

term	estimate	std.error	statistic	p.value
(Intercept)	-476.08	20.50	-23.23	0.00
age	34.15	0.86	39.80	0.00
eduGraduate	12.60	10.30	1.22	0.22
eduSecondary	8.71	12.98	0.67	0.50
birth_orderSecond	-32.22	10.93	-2.95	0.00
eduGraduate:birth_orderSecond	15.27	17.05	0.90	0.37
eduSecondary:birth_orderSecond	-30.65	19.71	-1.56	0.12

+———————+———————-+ | Variable | Meaning | +———————+———————-+ | eduGraduate | Graduate School | +———————+———————-+ | eduSecondary | Secondary School | +———————+———————-+ | Birth_orderSecond | Second Born | +———————+———————-+

Our next research question of interest once again revolved analyzing a particular variable that could influence development. In this case, we chose to focus on maternal education. Our methodology for analyzing this variable was quite straightforward. Similar to the birth order analysis from before, we simply created a scatterplot and conducted a regression on two levels of maternal education: secondary and graduate. However, it is worth to note that the original dataset contained many more levels. Specifically, the dataset included the following choices for education: “Primary”, “Some Secondary”, “Secondary”, “Some College”, “College”, “Some Graduate”, and “Graduate”. Out of these levels, we chose to filter the data by Secondary, College, and Graduate as the difference between these

three groups were the most pronounced. Finally, for clarity on the statistical significance of these variables, we also decided to conduct a multiple regression analysis. The variables involved in the regression model included vocabulary size, age, education, birth order, and an interaction term involving both birth order and maternal education. Considering that there is likely no correlation between the three included variables, we can say that the model is safe from multicollinearity. However, we would like to note that there could exist an issue involving omitted variable bias. For instance, one could suspect that there could be a particular variable, say income, that could correlate with both the explanatory variable of maternal education and the response variable of vocabulary size.

The resulting plot (Figure 4) and linear regression regarding maternal education displayed some peculiar characteristics. For one, the regression line for secondary maternal education starts off the highest and then intersects with the other two maternal education lines. Then, the graduate regression line overtakes both the secondary and college line, implying a greater vocabulary size for later ages. We believe that many would be quick to label maternal education as a significant variable. Intuitively, the relationship between maternal education level and vocabulary size seems quite defined; 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. However, the regression model revealed that the education coefficients were not significant. For instance, both p-values for the coefficients for the Graduate (0.22) and Secondary (0.50) education variables were greater than the chosen alpha level of 0.05. On the other hand, the p-value for the coefficient for Birth Order (0.00) came out as significant.

The results produced by the regression analysis were quite surprising. Fuel for discussion could involve conjecturing the reasoning behind the seemingly dampened effect of maternal education on early childhood vocabulary development. One possible explanation

could be that young children experience similar lexical exposure regardless of the educational status of the mother. For instance, book reading is a crucial element in constructing lexical diversity and may be administered relatively similarly by parents across varying educational backgrounds. Another possible explanation could be the limits of the model itself. Perhaps the divergence in vocabulary size between children with mothers of different educational backgrounds is more pronounced in later childhood or in adolescence, which would involve information that the data set does not contain.

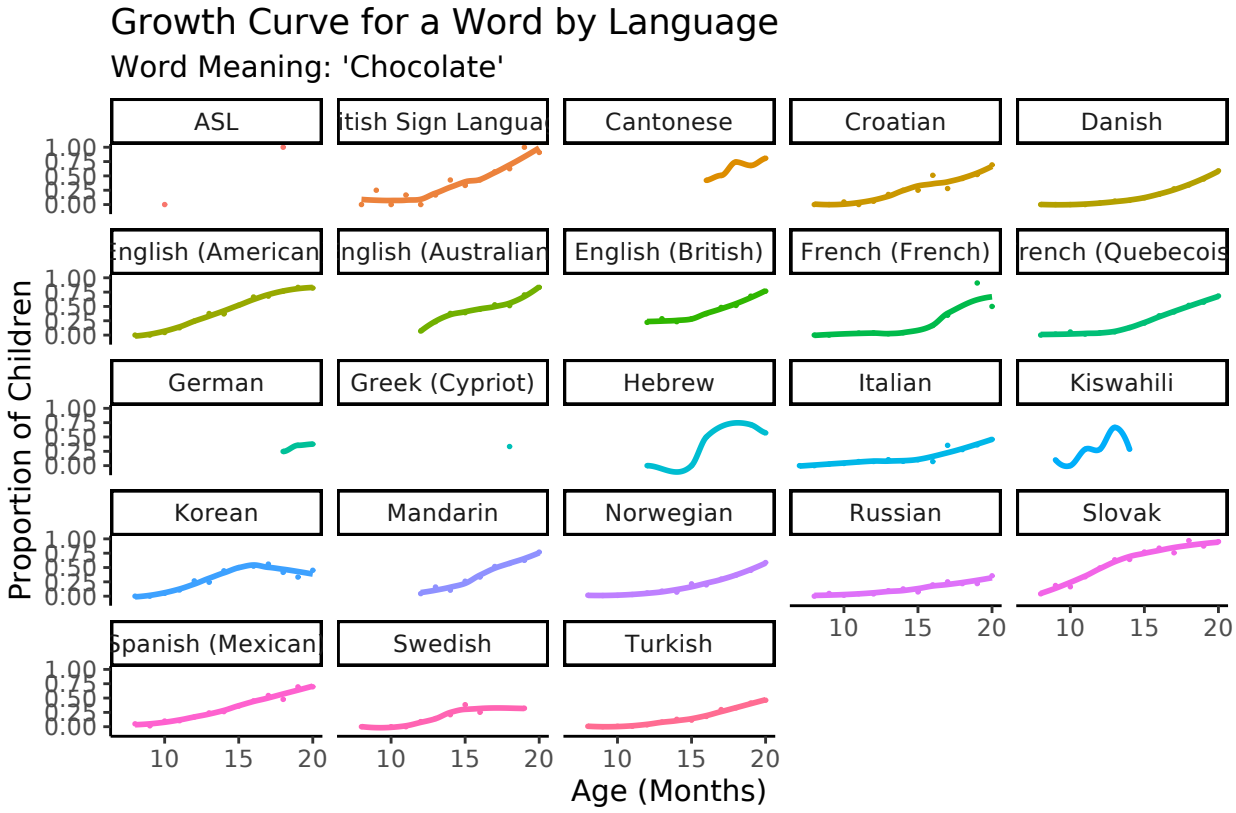

Figure 5

```
##
##  Welch Two Sample t-test
##
## data:  ling_data_croat$prop and ling_data_kor$prop
```

```
## t = -0.58443, df = 23.577, p-value = 0.5645
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2277647  0.1273136
## sample estimates:
## mean of x mean of y
## 0.2492592 0.2994847
```

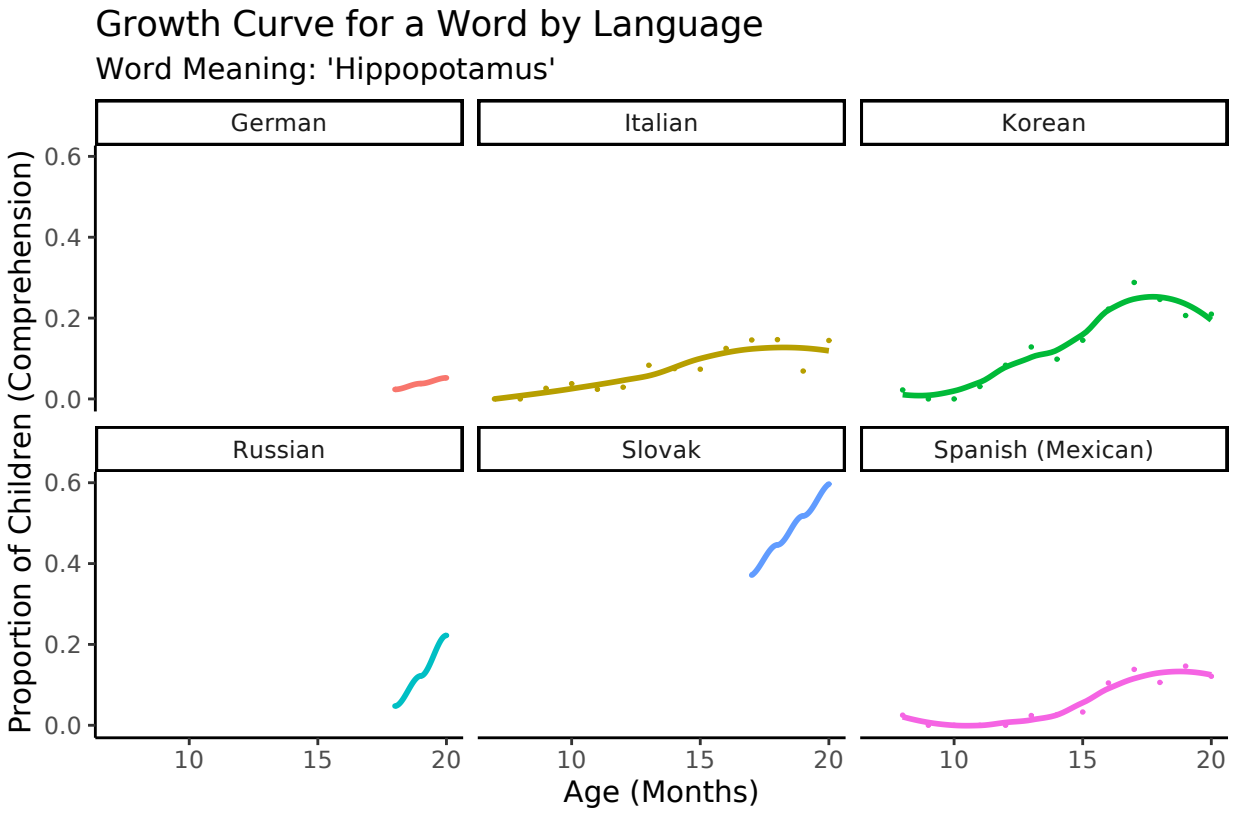

Figure 6

```
##
##  Welch Two Sample t-test
##
## data:  hippo_spn$prop and hippo_kor$prop
## t = -2.3132, df = 19.3, p-value = 0.03189
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.140075549 -0.007071681
## sample estimates:
##  mean of x  mean of y
## 0.05573073 0.12930434

##
##  Welch Two Sample t-test
##
## data:  hippo_ita$prop and hippo_kor$prop
## t = -1.9159, df = 18.116, p-value = 0.0713
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.124343282  0.005699347
## sample estimates:
##  mean of x  mean of y
## 0.06998237 0.12930434
```

The final research goal involved investigating the rate at which common words are learned in various languages. Essentially, we desired to get a better understanding of the extent that vocabulary development, especially for particular words, differed between languages. We also aimed to provide either a cultural or linguistic rationale for these differences. To easily showcase the differences in the rate of the acquisition of particular words between languages, we decided to a construct a Shiny application involving an interactive facet grid. The application allows the user to select a desired word from the dropdown menu. Then, the facet grid changes accordingly to showcase all the available languages that have a significant amount of data for the selected word. The plots indicate the mean proportion of children of that language who can produce the selected word. We decided to set an upper limit for age at eighteen months, which differs from our previous plots that have an upper

limit of thirty months. Considering the included words are fairly basic, the mean proportion of children able to produce the word, regardless of language, almost all converge to around one hundred percent. This would not be very useful in our goal to investigate differences. Overall, we hoped that this application would allow the user to develop a sense of familiarity with how vocabulary development differs between languages, and also function as a catalyst for delving into linguistic features that could contribute to certain disparities. In fact, we found two words that displayed clear disparities between certain languages. We sought to verify the statistical significance of these differences through conducting a two-sample t-test of proportions and investigate for possible explanations.

The first notable feature that piqued our interest was the production of the word chocolate. Figure 5 showcases the facet grid for the verbal production of chocolate. Note the clearly lower mean proportion in Korean than in any other language. This seemingly outlandish difference compelled us to conduct a test of statistical significance. In the respective t-test, we compared the mean proportion of children in Croatian and Korean who could produce the word chocolate. The usage of Croatian as a comparison was somewhat of an arbitrary decision; we simply assumed it as a benign comparison with many data points. The t-test outputted a p-value (0.02) less than an alpha level of 0.05, suggesting a statistically significant difference between the mean proportion in Korean and Croation.

Another test that we desired to conduct involved the difference in the mean proportion of children who were able to produce a word meaning hippopotamus in Spanish and Korean. Once again, the desire to conduct the test was simply a product of plot revealing a notable difference. This is seen in Figure 6. We conducted a two-sample t-test of proportions and found the difference statistically significant, producing a p-value of 0.03.

After verifying that there indeed existed a statistically significant difference between the these two particular words for the chosen languages we then attempted to provide a rationale. For the difference of the production of chocolate between Croatian and Korean, we assume it to be correlated with cultural differences. Chocolate is extremely popular in

Western culture and less so in Asian culture. We wish to be more specific but then run into the risk of unintentionally producing unfounded causality statements. For the comparison between Korean and Spanish, we assume it to be correlated with linguistic differences. The word meaning hippopotamus in Spanish is five syllables long, whereas it is a mere two syllables in Korean. Hippopotami live in neither region, which could reduce the influence of cultural familiarity or significance.

Citations

- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. Journal of Child Language. doi: 10.1017/S0305000916000209.

- Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. Developmental Science, 16, 234–248. http://doi.org/10.1111/desc.12019

- Pae, S., & Kwak, K. (2011). Korean MacArthur-Bates Communicative Development Inventories (K M-B CDI). Seoul: Mindpress.

- Kovacevic, M., Babic, Z., & Brozovic, B. (1996). A Croatian language parent report study: Lexical and grammatical development. Paper presented at the VIIth International Congress for the Study of Child Language, July 1996, Istanbul, Turkey.

- Jackson-Maldonado, D., Thal, D., Marchman, V., Newton, T., Fenson, L, & Conboy, B. (2003). MacArthur Inventarios del Desarrollo de Habilidades Comunicativas. User´s Guide and Technical Manual. Brookes, Baltimore.