Milestone 2

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.

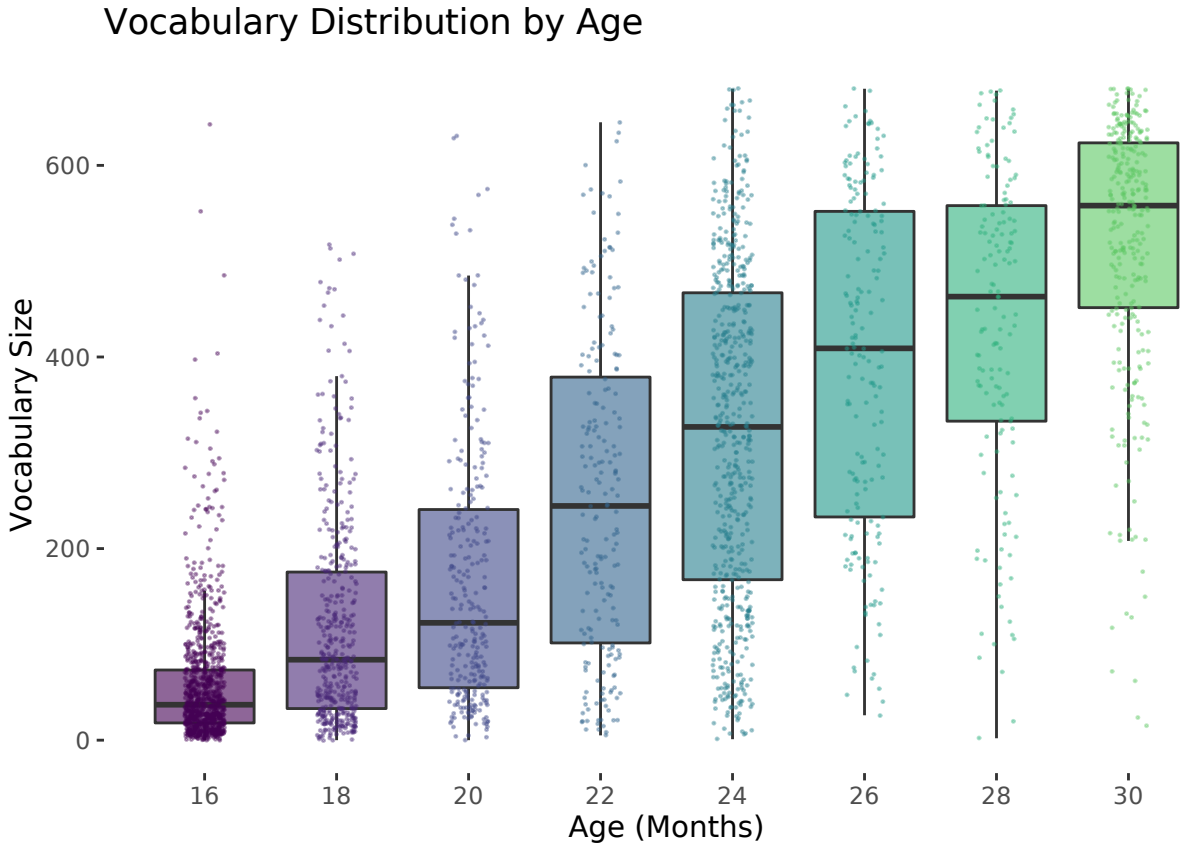
EDA

```
#load data
word_bank <- read.csv("vocabulary_norms_data.csv")

#filter out even ages
word_bank2 <- word_bank %>%
  filter(age %% 2 == 0)

font_add_google("Source Sans Pro")
ggplot(word_bank2, aes(x = age, y = vocab, fill = as.factor(age))) +
geom_boxplot(alpha = .6, outlier.shape = NA) +
geom_jitter(size = 0.2, alpha = 0.35, width = 0.3, aes(color = as.factor(age))) +
scale_fill_viridis_d(end = .75, option = "D", guide=FALSE) +
scale_color_viridis_d(end = .75, option = "D", guide=FALSE) +
```

```
labs(title = "Vocabulary Distribution by Age",
x = "Age (Months)",
y = "Vocabulary Size") +
scale_x_continuous(breaks = seq(from = 16, to = 30, by = 2)) +
theme(panel.background = element_blank(),
      text = element_text(family = "Source Sans Pro"))
```




```
#filter only first and second birth orders
word_bank3 <- word_bank2 %>%
  filter(birth_order == "First" | birth_order == "Second")

ggplot(word_bank3, aes(x = age, y = vocab, group = birth_order, col = birth_order)) +
  geom_jitter(size = 0.7, alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_fill_viridis_d(end = .75, option = "C") +
  scale_color_viridis_d(end = .75, option = "C", name = "Birth Order") +
  labs(title = "Vocabulary Distribution by Age",
       subtitle = "Comparing First and Second Birth Orders",
       x = "Age (Months)",
       y = "Vocabulary Size")  +
  theme(panel.background = element_blank(),
        text = element_text(family = "Source Sans Pro"),
```

```
        legend.position = "top",
        legend.justification = "left")
```

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