Lab 07 - Children's word learning

Gabrielle Young

October 19, 2023

Load packages and data

```r
library(tidyverse)
```

```r
wordbank <- read_csv("data/wordbank-item-freq.csv")
```

```
## Rows: 609 Columns: 22
## -- Column specification -----------------------------------------------------
## Delimiter: ","
## chr   (3): item_definition, category, word
## dbl  (18): item_id, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, ...
## date  (1): downloaded
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
glimpse(wordbank)
```

```
## Rows: 609
## Columns: 22
## $ downloaded      <date> 2022-09-22, 2022-09-22, 2022-09-22, 2022-09-22, 2022-~
## $ item_id         <dbl> 4, 5, 6, 7, 9, 10, 13, 14, 15, 16, 17, 18, 19, 20, 21,~
## $ item_definition <chr> "grrr", "meow", "moo", "ouch", "uh oh", "vroom", "alli~
## $ category        <chr> "sounds", "sounds", "sounds", "sounds", "sounds", "sou~
## $ `16`            <dbl> 0.48, 0.41, 0.50, 0.34, 0.71, 0.38, 0.02, 0.03, 0.04, ~
## $ `17`            <dbl> 0.49, 0.48, 0.52, 0.41, 0.68, 0.40, 0.04, 0.07, 0.06, ~
## $ `18`            <dbl> 0.55, 0.56, 0.62, 0.51, 0.78, 0.51, 0.06, 0.11, 0.10, ~
## $ `19`            <dbl> 0.64, 0.68, 0.75, 0.61, 0.87, 0.57, 0.13, 0.20, 0.18, ~
## $ `20`            <dbl> 0.65, 0.71, 0.76, 0.64, 0.87, 0.57, 0.14, 0.21, 0.21, ~
## $ `21`            <dbl> 0.67, 0.78, 0.78, 0.67, 0.89, 0.61, 0.19, 0.28, 0.23, ~
## $ `22`            <dbl> 0.65, 0.77, 0.79, 0.70, 0.88, 0.60, 0.22, 0.33, 0.26, ~
## $ `23`            <dbl> 0.74, 0.81, 0.85, 0.77, 0.88, 0.69, 0.30, 0.44, 0.37, ~
## $ `24`            <dbl> 0.77, 0.86, 0.88, 0.78, 0.90, 0.70, 0.40, 0.50, 0.43, ~
## $ `25`            <dbl> 0.78, 0.89, 0.88, 0.86, 0.93, 0.75, 0.47, 0.56, 0.49, ~
## $ `26`            <dbl> 0.76, 0.88, 0.88, 0.82, 0.90, 0.71, 0.50, 0.58, 0.51, ~
## $ `27`            <dbl> 0.80, 0.92, 0.92, 0.83, 0.93, 0.70, 0.52, 0.61, 0.54, ~
## $ `28`            <dbl> 0.90, 0.96, 0.95, 0.93, 0.97, 0.86, 0.74, 0.78, 0.74, ~
## $ `29`            <dbl> 0.79, 0.91, 0.92, 0.86, 0.91, 0.72, 0.66, 0.71, 0.63, ~
## $ `30`            <dbl> 0.82, 0.93, 0.93, 0.88, 0.89, 0.74, 0.70, 0.73, 0.66, ~
## $ word            <chr> "grrr", "meow", "moo", "ouch", "uhoh", "vroom", "allig~
## $ freq            <dbl> 276325, 810281, 1631972, 1174728, 27389, 209222, 14176~
## $ freqrank        <dbl> 0.8327283, 0.9108456, 0.9405968, 0.9281497, 0.3541304,~
```

Exercises

Exercise 1

```
wordbank_long <- wordbank %>%
pivot_longer(cols= c(`16`:`30`), names_to = "AgeinMonths",
values_to = "Proportion")

wordbank_18mo <- wordbank_long %>% filter(AgeinMonths == "18") %>% select(word, Proportion) %>%
  arrange(desc(Proportion)) %>% head(3)
wordbank_30mo_sounds <- wordbank_long %>% filter(category == "sounds", AgeinMonths == "30") %>%
  select(word, Proportion) %>% arrange(Proportion) %>% head(1)
glimpse(wordbank_18mo)
```

```
## Rows: 3
## Columns: 2
## $ word       <chr> "ball", "bye", "uhoh"
## $ Proportion <dbl> 0.84, 0.80, 0.78
```

```
glimpse(wordbank_30mo_sounds)
```

```
## Rows: 1
## Columns: 2
## $ word       <chr> "vroom"
## $ Proportion <dbl> 0.74
```

The three most commonly produced words by 18 months are ball, bye, and uhoh. Vroom is the least likely sounds to be produced by 30 months.

Exercise 2

```
wordbank_long <- wordbank %>%
pivot_longer(cols= c(`16`:`30`), names_to = "AgeinMonths",
values_to = "Proportion")

wordbank_long_sounds <- wordbank %>% filter(category == "sounds") %>%
pivot_longer(cols= c(`16`:`30`), names_to = "AgeinMonths",
values_to = "Proportion")

wordbank_long_sounds %>% ggplot(aes(x = AgeinMonths, y = Proportion,
fill = word, color = word)) + geom_point()
```
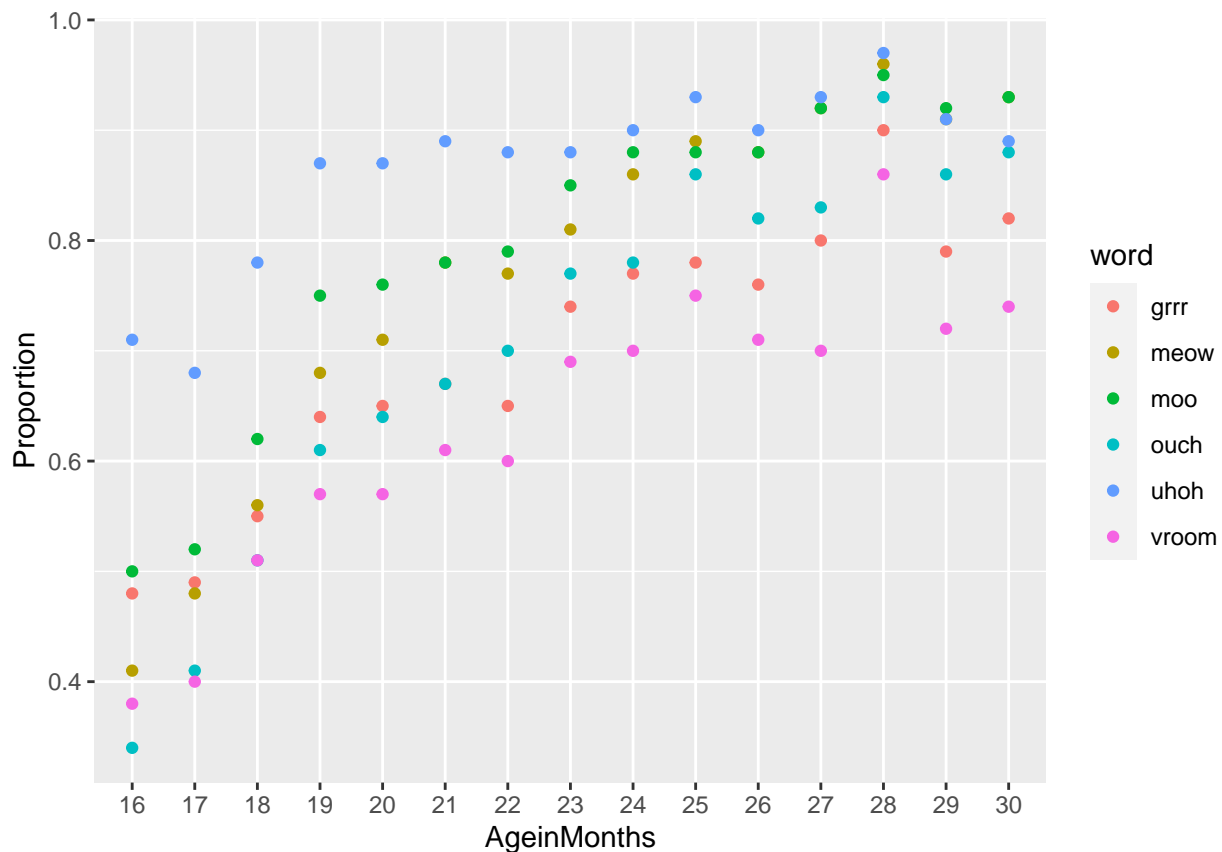
Children at 28 months tend to have a higher average of children that are able to pronounce these words. From 16 to 25 months,"uhoh" is learned the fastest. This word is probably vocalized more to younger children when they are in the learning process, as opposed to "meow" or "vroom."

Exercise 3

```
wordbank_long_category <- wordbank  %>% filter(category== c("sounds",
"games_routines", "body_parts", "animals","clothing","people","pronouns",
"connecting_words","quantifiers","helping_verbs"))%>%
pivot_longer(cols= c(`16`:`30`), names_to = "AgeinMonths",
values_to = "Proportion")
```

```
## Warning in category == c("sounds", "games_routines", "body_parts", "animals", :
## longer object length is not a multiple of shorter object length
```

```
wordbank_long_category %>% group_by(category,AgeinMonths) %>%
  mutate(Meanproportion=mean(Proportion)) %>%
  ggplot( aes(x=AgeinMonths, y=Meanproportion, group=category,
              color=category))+
  geom_point()+
  geom_line()+
  facet_wrap(~category)
```
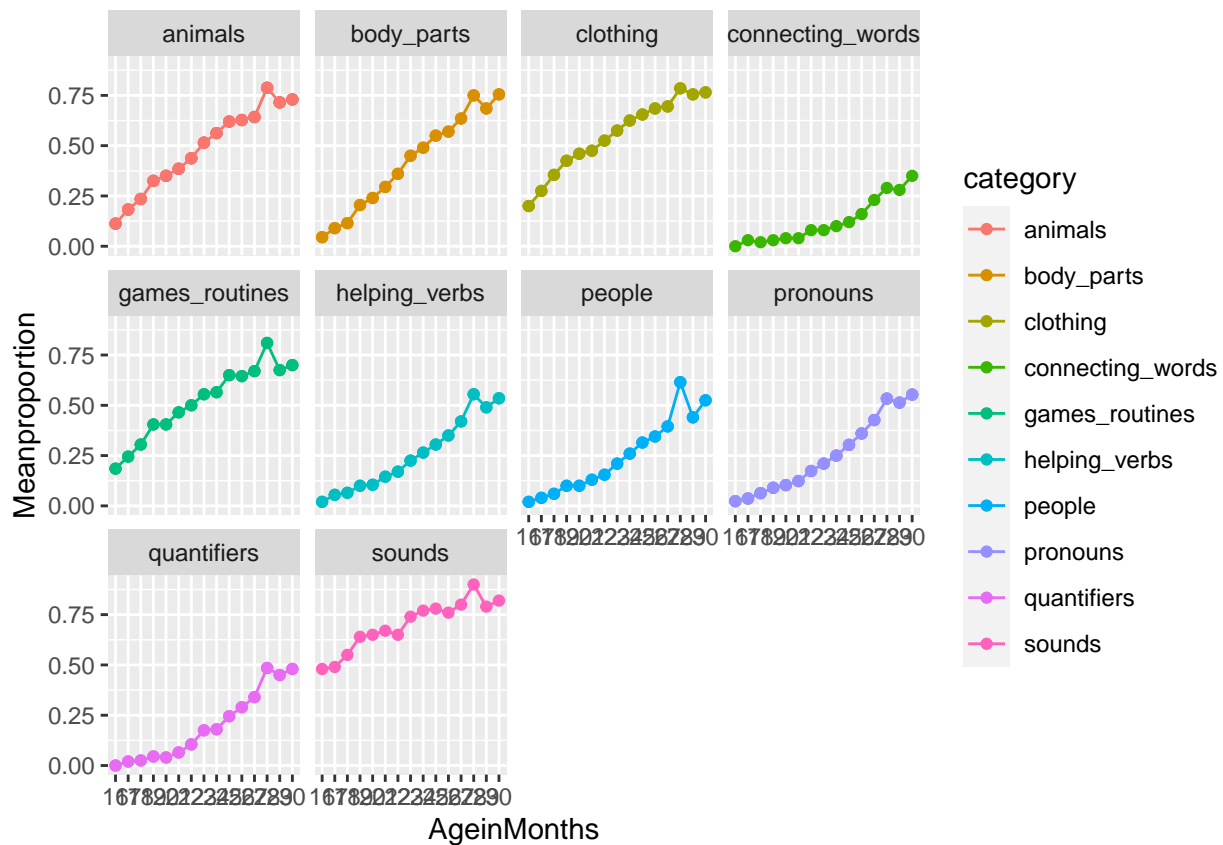
Exercise 4

```r
wordbank_10_most_frequent <- wordbank[order(-wordbank$freq), ]%>% select(word,
                                                                 freq) %>% head(10)

wordbank_10_least_frequent <- wordbank[order(wordbank$freq), ]%>% select(word,
                                                                 freq) %>% head(10)

print(wordbank_10_least_frequent)
```

```
## # A tibble: 10 x 2
##    word          freq
##    <chr>        <dbl>
##  1 uhoh          27389
##  2 pattycake     28656
##  3 playdough     44892
##  4 peanutbutter  46161
##  5 teddybear     51233
##  6 firetruck     71915
##  7 snowsuit      76873
##  8 hafta         82485
##  9 peekaboo     117785
## 10 yucky        154277
```

```r
print(wordbank_10_most_frequent)
```

```
## # A tibble: 10 x 2
##    word          freq
```

```
##    <chr>          <dbl>
##  1 the    23135851162
##  2 and    12997637966
##  3 a       9081174698
##  4 is      4705743816
##  5 that    3400031103
##  6 this    3228469771
##  7 i       3086225277
##  8 you     2996181025
##  9 it      2813163874
## 10 not     2633487141
```

- Most frequent: the, and, a, is, that, this, i, you, it, not
- Least frequent: uhoh, pattycake, playdough, peanutbutter, teddybear, firetruck, snowsuit, hafta, peekaboo, yucky

Exercise 5

```r
wordbank_long_2_years <- wordbank_long %>% filter(AgeinMonths == "24")

wordbank_long_2_years %>% mutate(log_freq = log(freq)) %>%
  ggplot(aes(x = log_freq, y = Proportion)) + geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```
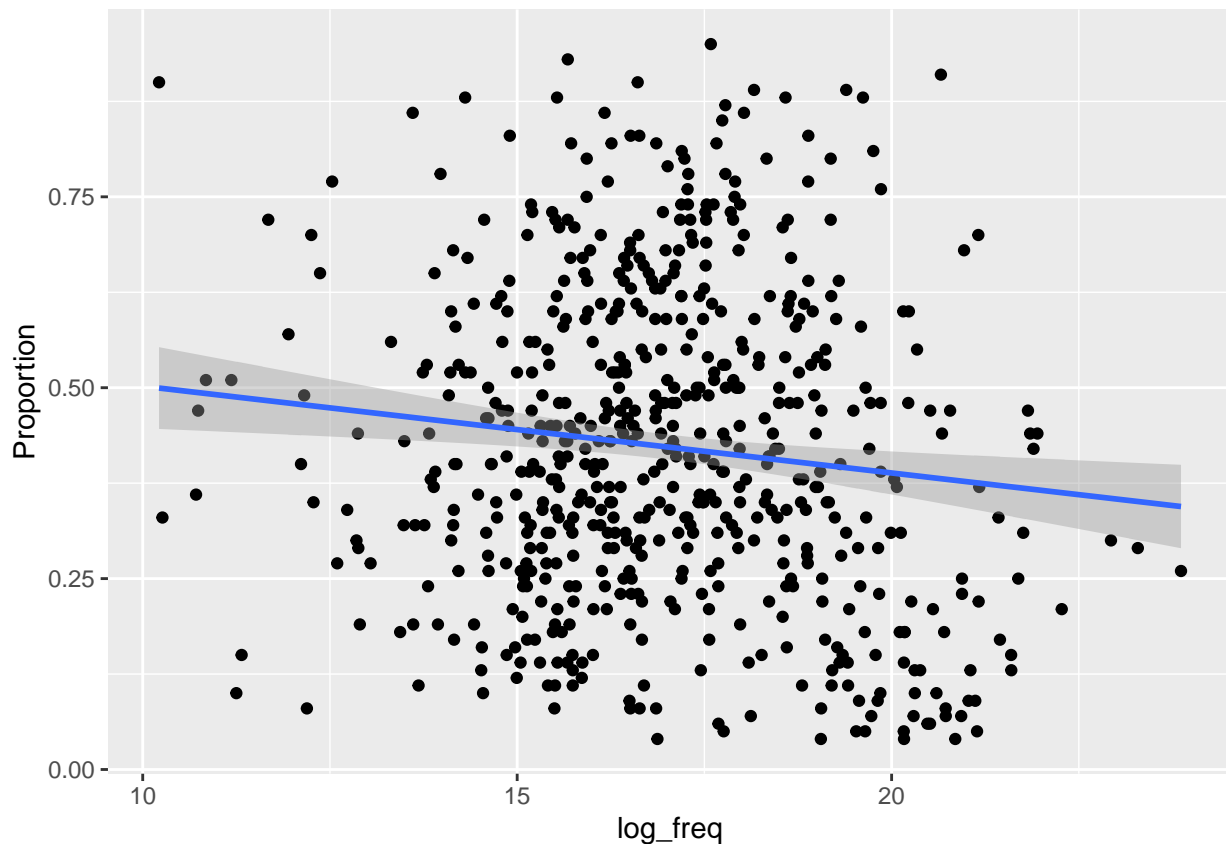
Exercise 6

This relationship is negative. There seems to be a negative relationship between the frequnecy a word is used and the proportion of children who use it. This does seem surprising because I would think that if a word is heard and used more in general, children would pick up on learning it quicker, and also use it more. This relationship may be explained by the combination of many types of words into one analysis.

Exercise 7

```
wordbank_long_2_years %>% mutate(log_freq = log(freq)) %>%
  ggplot(aes(x = log_freq, y = Proportion)) + geom_point() +
  geom_smooth(method="lm") + facet_wrap( ~category)
```

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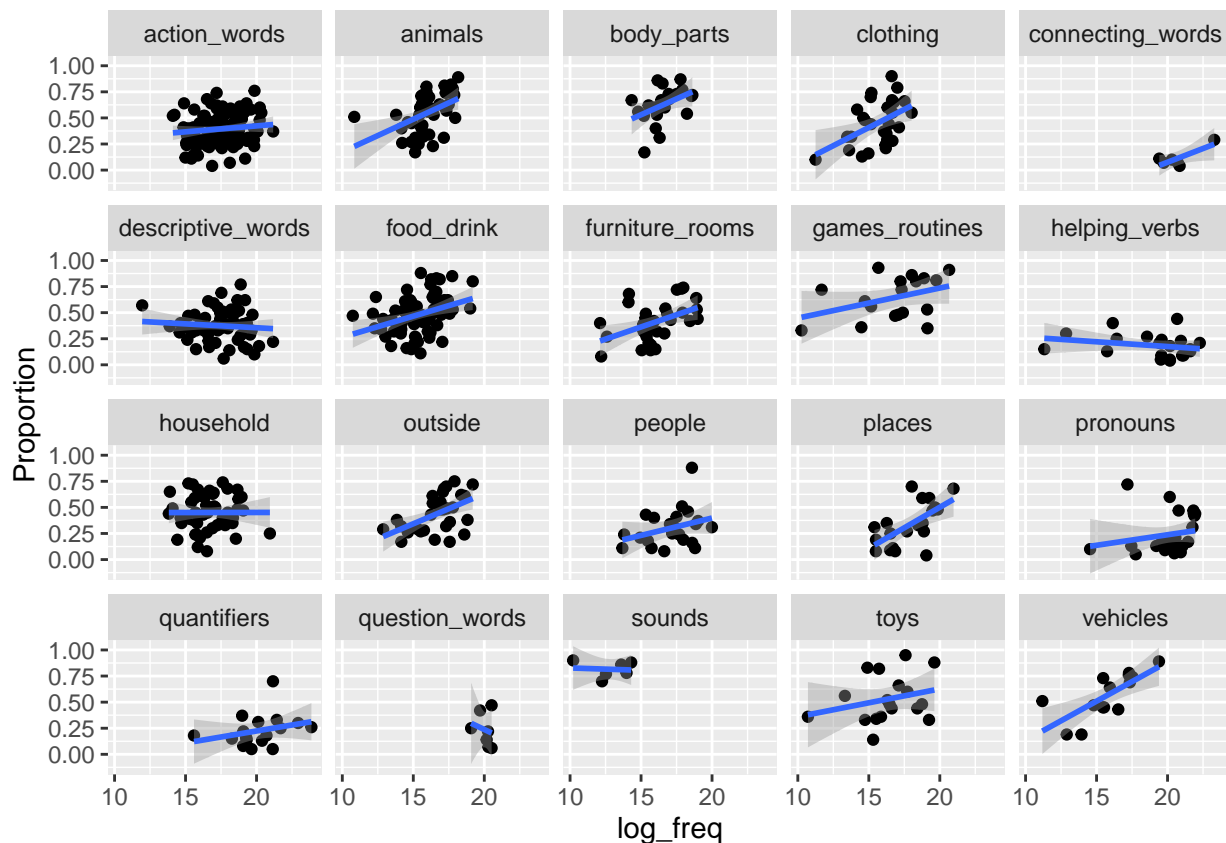

This graph illustrates the differences in log frequencies of words used and the proportion of words used by category. Here, differences between types of words are much more visible. For example, there is a positive relaitonship in log frequency and proportion for vehicles, game_routines, animals, and body_parts. There is a neutral relationship for household, sounds, and pronouns. The first plot illustrates a combination of all points, regardless of category, while the second plot shows more specific relationshi between the log frequency of a word and the proportion of 2 year olds using it based on category.

Exercise 8

People, sounds, furniture_rooms seem to be categories that should be heard more often in real life than written in books.

Exercise 9

```r
wordbank_high_freq <- wordbank_long_2_years %>% filter(category == c("connecting_words",
"quantifiers")) %>% mutate(log_freq = log(freq))
```

```
## Warning in category == c("connecting_words", "quantifiers"): longer object
## length is not a multiple of shorter object length
```

```r
wordbank_high_freq %>% select(log_freq, word,category) %>%
arrange(desc(log_freq)) %>% head(10)
```

```
## # A tibble: 10 x 3
##     log_freq word    category
##        <dbl> <chr>   <chr>
## 1      23.9 the      quantifiers
## 2      21.7 not      quantifiers
## 3      21.4 all      quantifiers
## 4      21.1 an       quantifiers
## 5      20.8 if       connecting_words
## 6      20.4 any      quantifiers
## 7      19.7 then     connecting_words
## 8      19.4 because  connecting_words
## 9      19.3 much     quantifiers
## 10     19.3 same     quantifiers
```

The most frequent words are the, not, all, an, if, any, then, because, much, and same. I would assume that children hear these words very often. They also probably see these words written fairly often.