WORD ACQUISITION OF INFANTS AND TODDLERS

Hypotheses

This study has 2 hypotheses to prove.

First of them is that children memorize words of closed class after acquisition of particular amount of words from opened class. In this paper the author tries to prove that toddlers assimilate nouns earlier than verbs and adjectives, when closed class items are scarcer in their speech.

The author also expects that frequency of each category in child directed speech and adults' speech have linear negative association with age of acquisition of each word class. A general idea of connection between these factors is that frequencies of each reproduced group of words could be the highest when children are the youngest (in other words, infants have more probability to learn more frequent words as their first words). During their growing up children commence to remember rarer words. It is the second hypothesis of the study hereto.

Background

The previous research on this topic is provided by Pernille Hansen (Hansen, 2016). The purpose of her article *"What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development"* is to study two ideas. First, how individual and joint effects and also word class, frequency, imageability and phonological neighbourhood density influence lexical development of children. Second, how age of acquisition and vocabulary size are connected. Of her particular interest is the competition and interaction between the factors.

"Norwegian Words: A lexical database for clinicians and researchers" (Lind, M. et al., 2015) describes the Norwegian Words database. Frequencies of words in adults' speech used in this paper are based on the results of that work.

In *"The Phonology of Children's Early Words: Trends, Individual Variation, and Parents' Accommodation in Child-Directed Speech"* (Garmann et al., 2019) the authors focus on the phonological characteristics of children's first words and on how parents accommodate to childrens' patterns in their speech.

In contrast to previous works, this paper is mostly based on numeric variables with general parts of speech.

Description of the data

Data collection method

The main data consists of 732 words included in the Norwegian adaptation of MacArthur-Bates Communicative Development Inventories (abbreviated as “CDI”) form, a parental questionnaire demonstrated to give a valid and reliable measure of early lexical development. The Norwegian CDI norms consist of averaged data from 6500 monolingual Norwegian toddlers. Important thing about CDI: their norms are typically cross-sectional and can thus not be used to determine when each child acquired each word, only which words she or he currently produces (or understands). This limitation was circumvented by calculating the age in months where at least 50% of the children are reported to produce each word. It is commonly referred to the data as a word’s age of acquisition.

This data was collected by Pernille Hansen, PhD in linguistics at MultiLing, University of Oslo.

Description of the data

The dataset includes two files. The main table consists of information about 732 Norwegian words with translation into English, age of acquisition and vocabulary size reproduced by more that 50% of children when they learn words, lexical and broad lexical categories of each word and frequencies of words in the Norwegian Web Corpus, which is used to evaluate frequencies of words in adults’ speech, and child-directed speech (abbreviated as “CDS”). The “Norwegian CDS frequency” table replicates some columns from main data. The second table includes measures of how frequently each word is used in Norwegian both on the internet (as observed in the Norwegian Web as Corpus dataset) and when an adult is talking to a child. It is a reason why the author of the piece of research explore only the first file.

Main data columns

- **ID_CDI_I** (ID in the Norwegian CDI I (Kristoffersen et al., 2012));
- **ID_CDI_II** (ID in the Norwegian CDI II (Kristoffersen et al., 2012));
- **Word_NW** (Entry as given in the database Norwegian Words (Lind et al, 2015));
- **Word_CDI** (Entry as given in the CDI (Kristoffersen et al., 2012));

- **Translation** (translation to English of each word);
- **AoA** (Age of acquisition: the age (in months) where 50% of the norming sample of children (Simonsen et al., 2014) are reported to produce the word);
- **VSoA** (Vocabulary size of acquisition: the vocabulary size (in spans of 20 words) where 50% of the norming sample of children (Simonsen et al., 2014) are reported to produce the word);
- **Lex_cat** (Lexical category as given by Kristoffersen et al. (2012));
- **Broad_lex** (Broad lexical category based on Bates et al. (1994));
- **Freq** (Frequency in the corpus Norwegian Web (Guevara, 2010), taken to approximate frequency in adults' speech);
- **CDS_freq** (Frequency in child-directed speech based Simonsen (1990) and Garmann-Norwegian (Garmann, 2016; Garmann et al., in press)).

Main factors

In the current dataset nouns are named as *common nouns*, *places to go*, *people* (in **Lex_cat**) and *nominals* (in **Broad_lex**). Some nouns are included in *games&routines* (both in **Lex_cat** and **Broad_lex**), which also consists of interjections ('ja' - 'yes', 'nei' - 'no','hallo' - 'hello'), verbs ('vent' - 'wait') and collocations ('gå på do' - 'go to the bathroom', 'god natt' - 'good night').

As a broad category, *nominals* comprise some interjections as *sound effects* (in **Lex_cat**).

Most of all verbs are placed into *action words* category (in **Lex_cat**) and *predicates* (in **Broad_lex**). Furthermore, it is important to notice that adjectives are included in *predicates* (in **Broad_lex**) group, but they are named as *descriptive words* (in **Lex_cat**).

Closed class words are supplied in *closed-class items* (in **Lex_cat**), or *closed-class* (in **Broad_lex**).

There appears to be *temporal* category which is not identified in the main data ('#N/A' category). It is composed of temporal nouns, prepositions and adverbs such as 'formiddag' - 'morning', 'nå' - 'now' and 'senere' - 'later'.

According to this information, the author decided to analyze *common nouns*, *places to go* and *people* as nouns, *action words* as verbs, *descriptive words* as adjectives and *closed-class items* as closed class words. The reason of this choice is that other categories are mixed. *games&routines* and *temporal* categories are able to confuse models and results. The author decided to remain *sound effects* category for further research on this topic.

Additional relevant factors for the paper are frequencies of words in adults' and child directed speech and age of each word's reproducing.

```
## # A tibble: 6 x 11
##   ID_CDI_I ID_CDI_II Word_NW Word_CDI Translation AoA   VSoA 
Lex_cat Broad_lex
##   <chr>    <chr>     <chr>   <chr>    <chr>       <chr> <chr>
<chr>   <chr>    
## 1 i_4_1    i_1_1     'au'    'au'     'ouch'      16    40   
sound … nominals 
## 2 i_4_2    i_1_2     'bææ'   'bææ'    'baa baa'   15    40   
sound … nominals 
## 3 i_4_3    i_1_3     'brrr … 'brrr (… 'vroom'     13    20   
sound … nominals 
## 4 i_4_4    i_1_4     'gakk … 'gakk g… 'quack qua… 17    40   
sound … nominals 
## 5 i_4_5    i_1_5     'grr'   'grr'    'grr'       22    220  
sound … nominals 
## 6 i_4_6    i_1_6     'kykel… 'kykeli… 'cock-a-do… 22    220  
sound … nominals 
## # … with 2 more variables: Freq <chr>, CDS_freq <chr>
```

Preprocessing

The data have missing values. First, we should rename gaps in lexical categories. '#N/A' strings there are located in rows with temporal nouns and prepositions (such as, for instance, 'morgen' - 'morning', 'før'- ('before').

```
main_data$Lex_cat <- gsub('#N/A', 'temporal', main_data$Lex_cat)
main_data$Lex_cat <- as.factor(main_data$Lex_cat)
main_data$Broad_lex <- gsub('#N/A', 'temporal', main_data$Broad_l
ex)
main_data$Broad_lex <- as.factor(main_data$Broad_lex)
main_data$Word_CDI <- as.factor(main_data$Word_CDI)
```

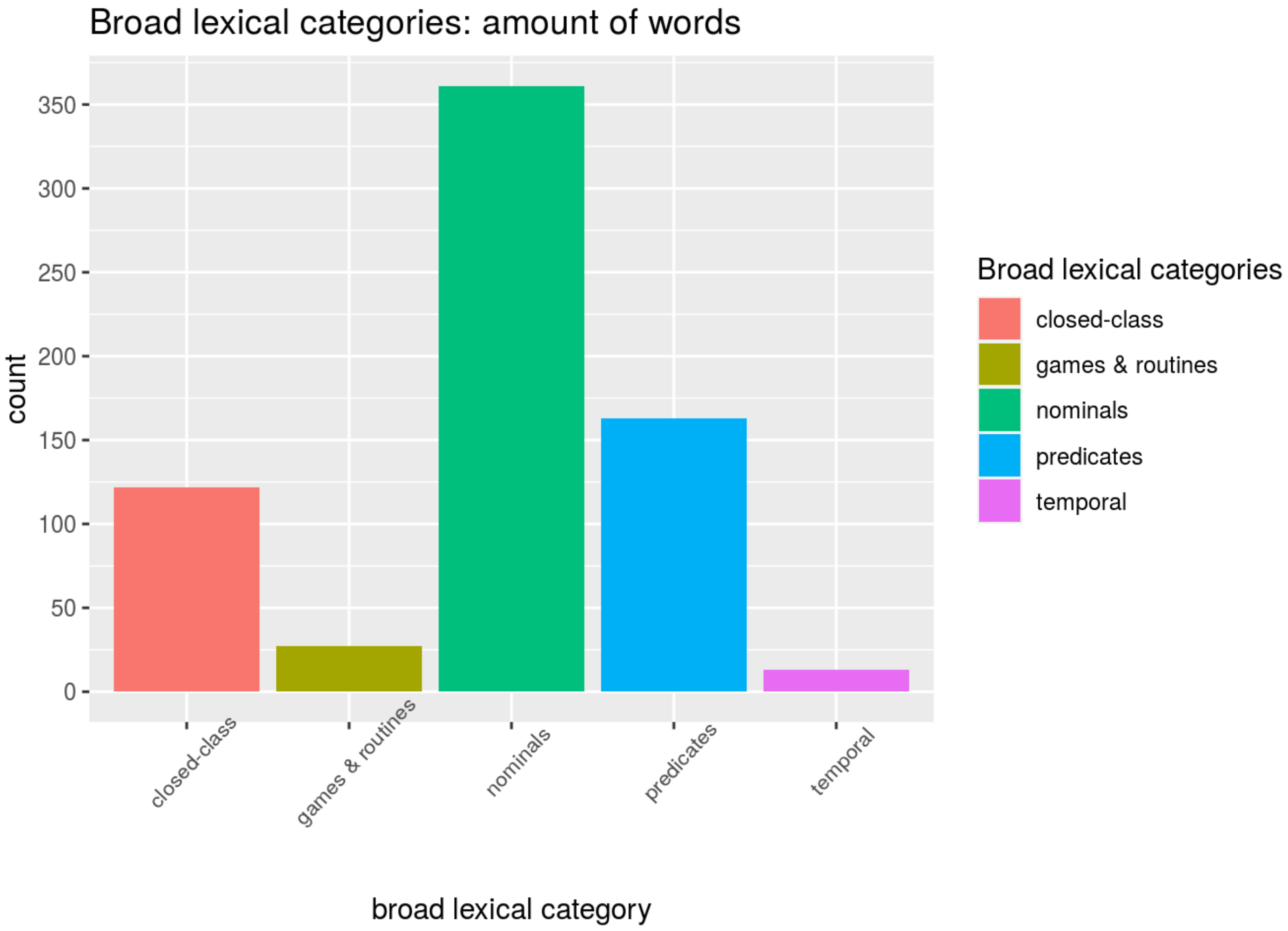
Missing values are observed in columns with numeric content. It is necessary to remove them.

```
main_data$AoA <- as.numeric(main_data$AoA)
main_data$VSoA <- as.numeric(main_data$VSoA)
main_data$CDS_freq <- as.numeric(main_data$CDS_freq)
main_data$Freq <- as.numeric(main_data$Freq)
main_data<-na.omit(main_data)
```

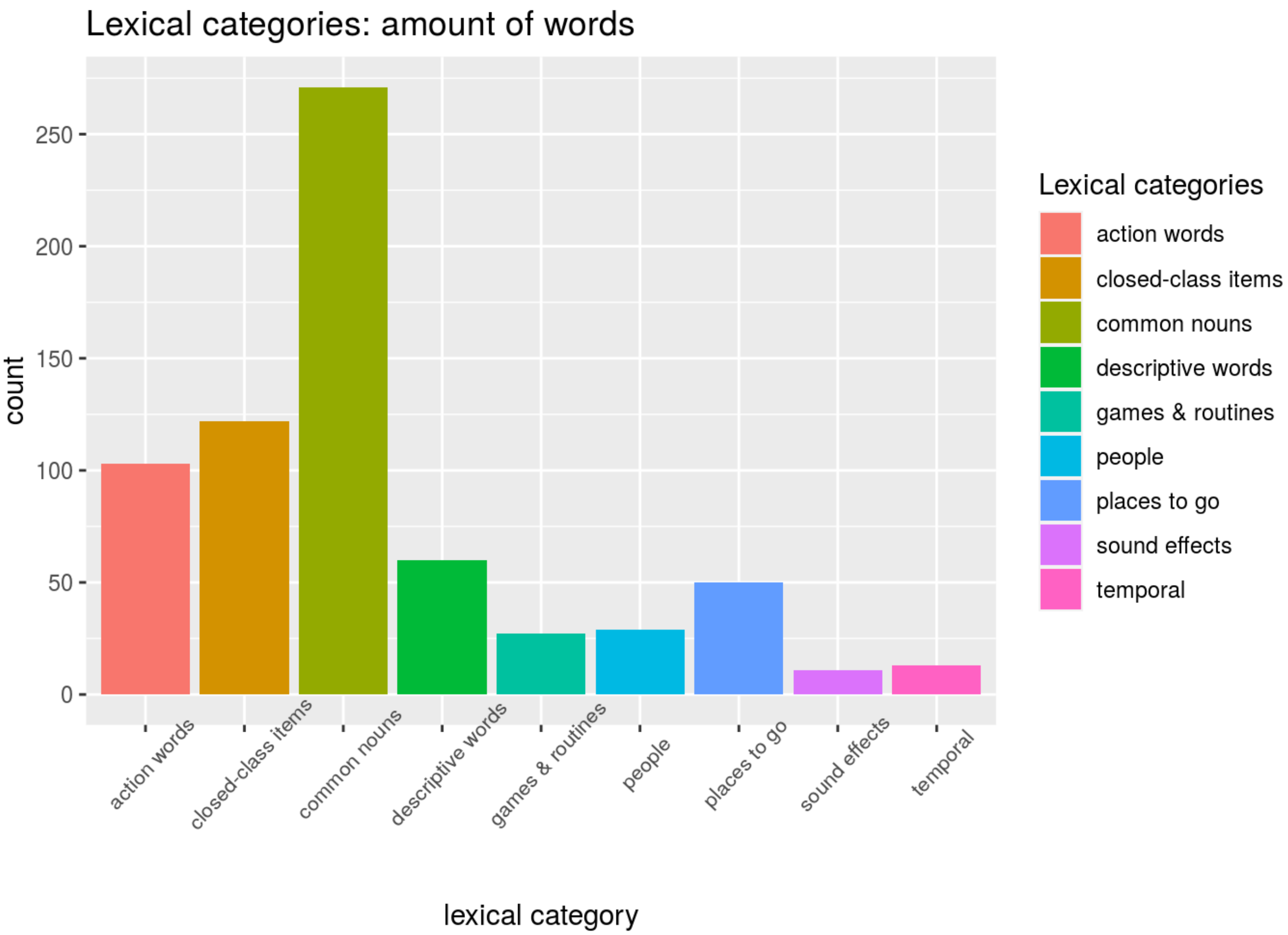
Introduction to the main data

On the bar charts below quantities of lexical and broad lexical categories are represented.

```
main_data %>%
ggplot(aes(Broad_lex, fill=Broad_lex))+geom_bar()+
        labs(title = "Broad lexical categories: amount of words",
             x = "broad lexical category")+scale_fill_discrete("B
road lexical categories") +
  theme(axis.text.x = element_text(size=8, angle=47, margin = mar
gin(t=10)))+
  scale_y_continuous(breaks=seq(0,400,50))
```

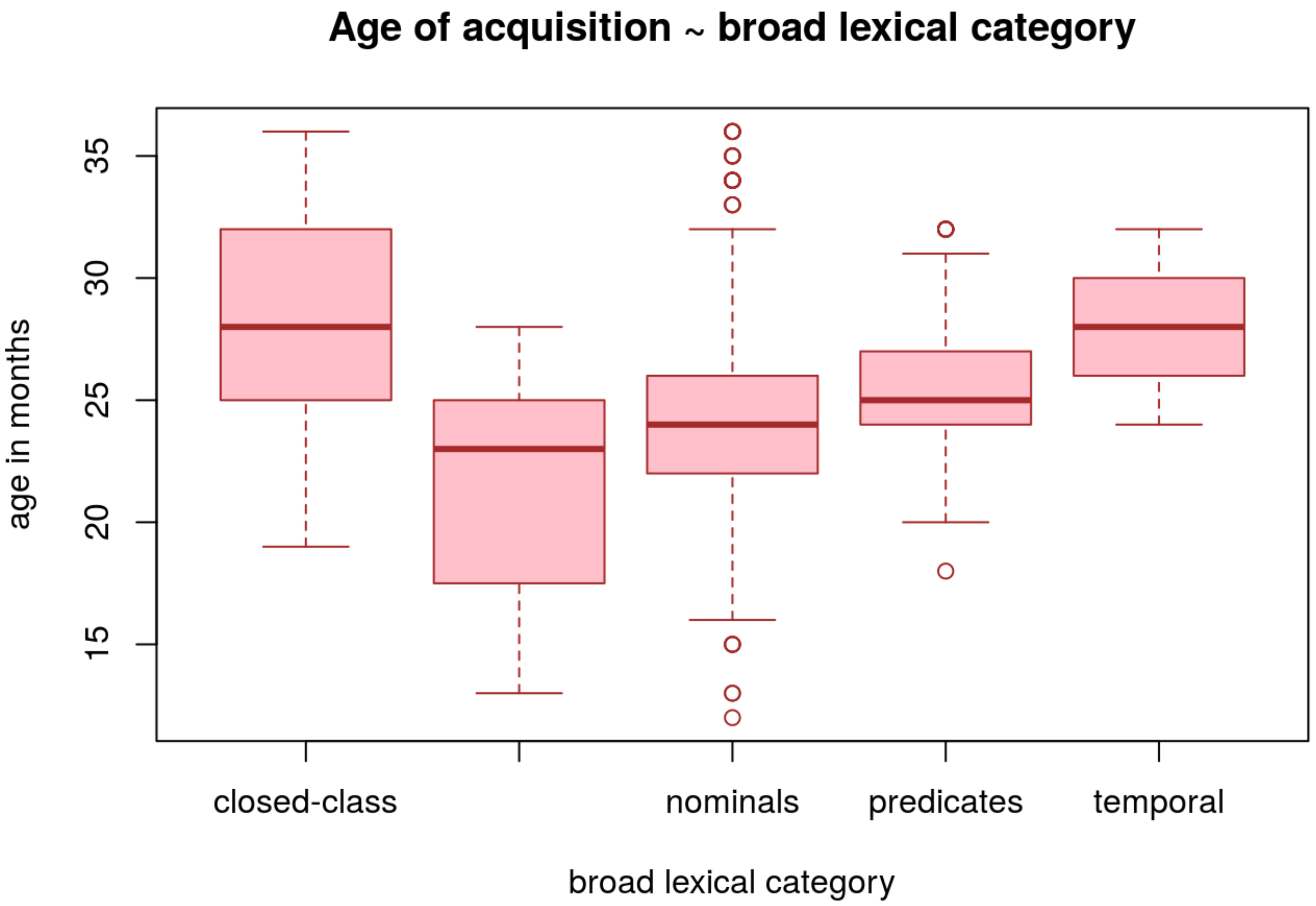


```
par(mar=c(7,7,1,1)+3.5,mgp=c(6,1,0))
main_data %>%
ggplot(aes(Lex_cat, fill=Lex_cat))+geom_bar()+
        labs(title = "Lexical categories: amount of words",
             x = "lexical category")+scale_fill_discrete("Lexical
categories")+
  theme(plot.margin = margin(1, 0, 1, 0))+
  theme(axis.text.x = element_text(size=8, angle=47, margin = mar
gin(t=10)))+
  scale_y_continuous(breaks=seq(0,400,50))
```



We are able to notice some outliers in the dataset.

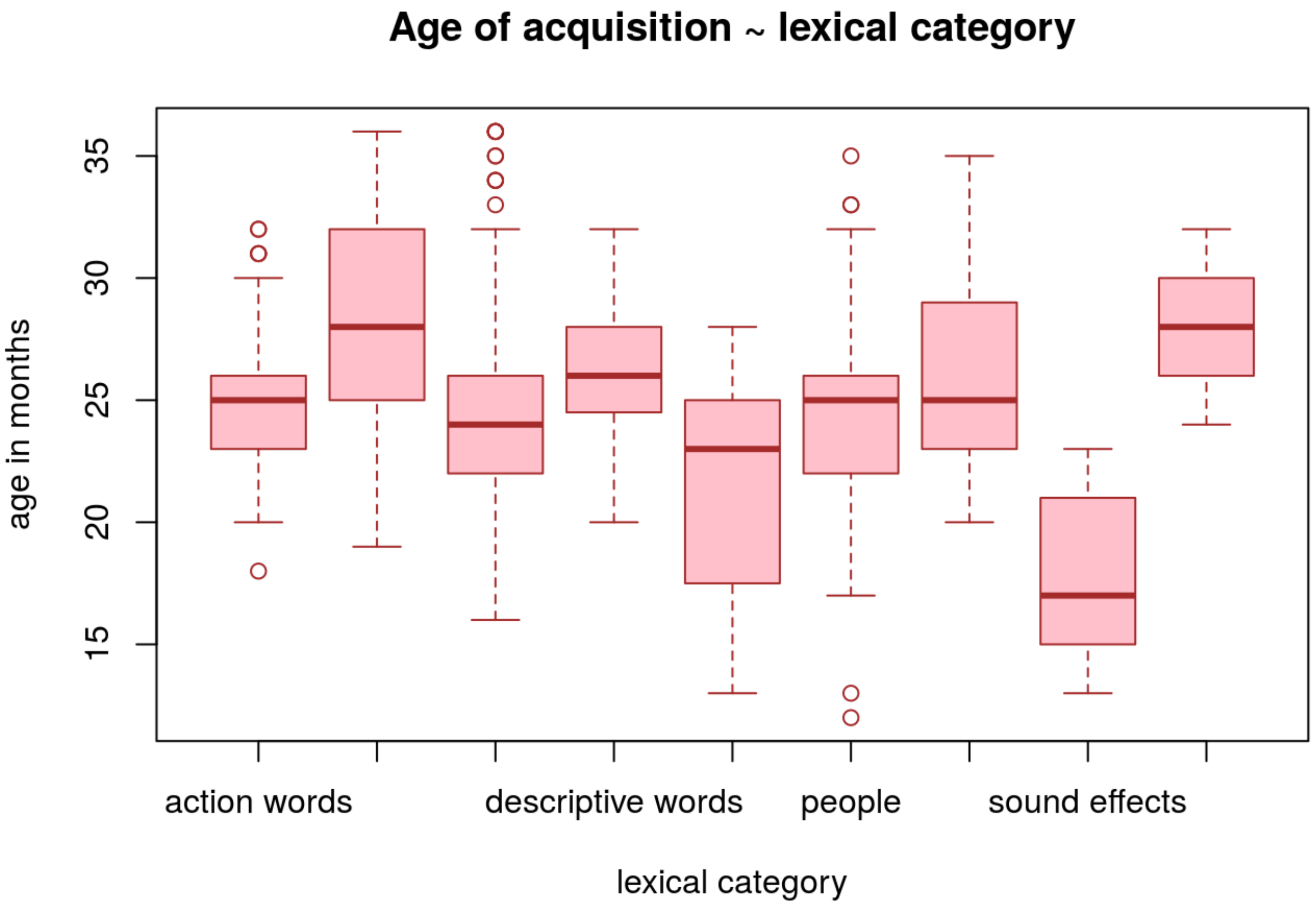
```
length(boxplot(main_data$AoA ~ main_data$Broad_lex,
        main = "Age of acquisition ~ broad lexical category",
        xlab = "broad lexical category",
        ylab = "age in months",
        col = "pink",
        border = "brown")$out)
```



```
## [1] 32
```

```
# [1] 32
```

```
length(boxplot(main_data$AoA ~ main_data$Lex_cat,
        main = "Age of acquisition ~ lexical category",
        xlab = "lexical category",
        ylab = "age in months",
        col = "pink",
        border = "brown")$out)
```

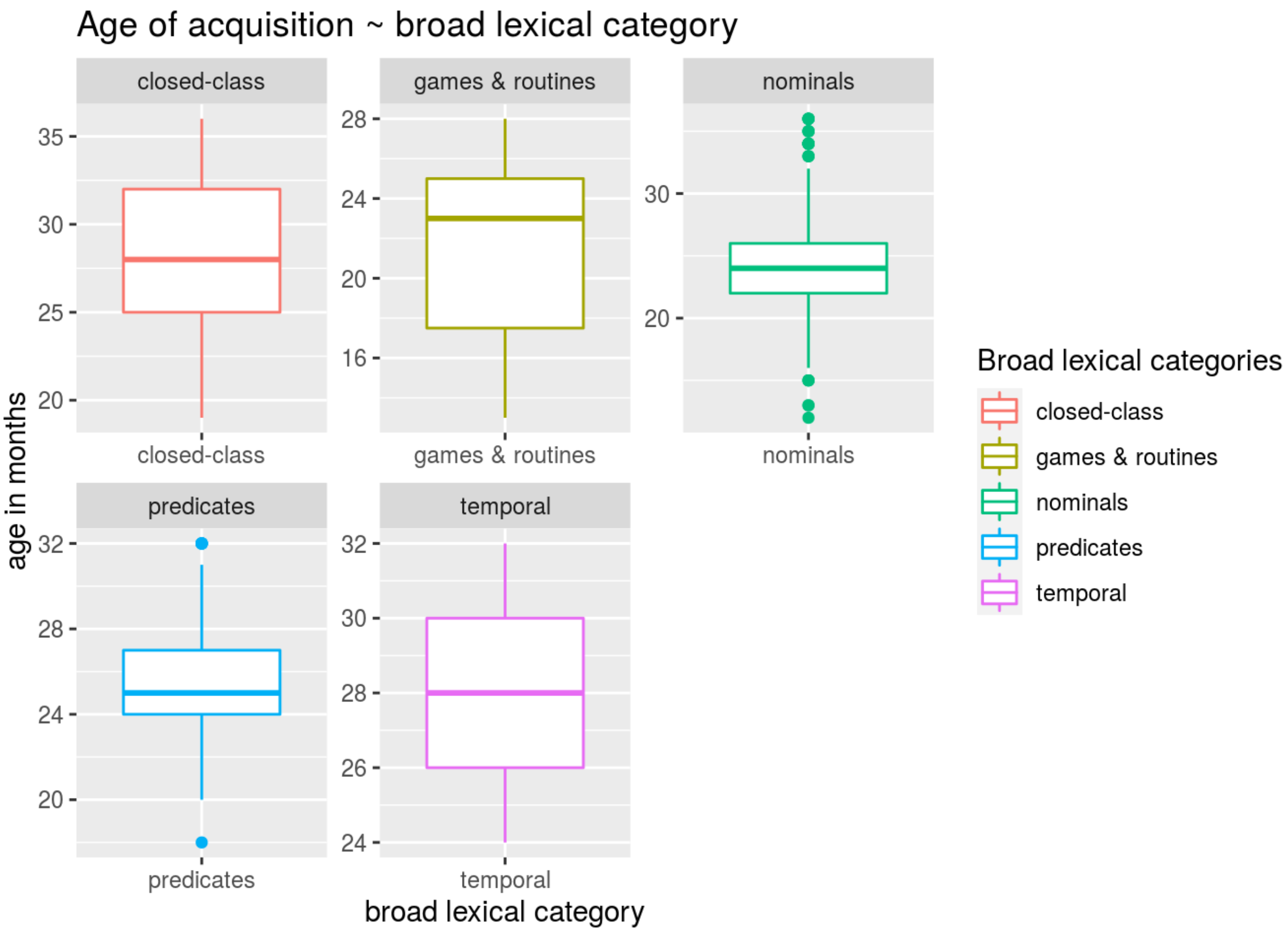


```
## [1] 22
```

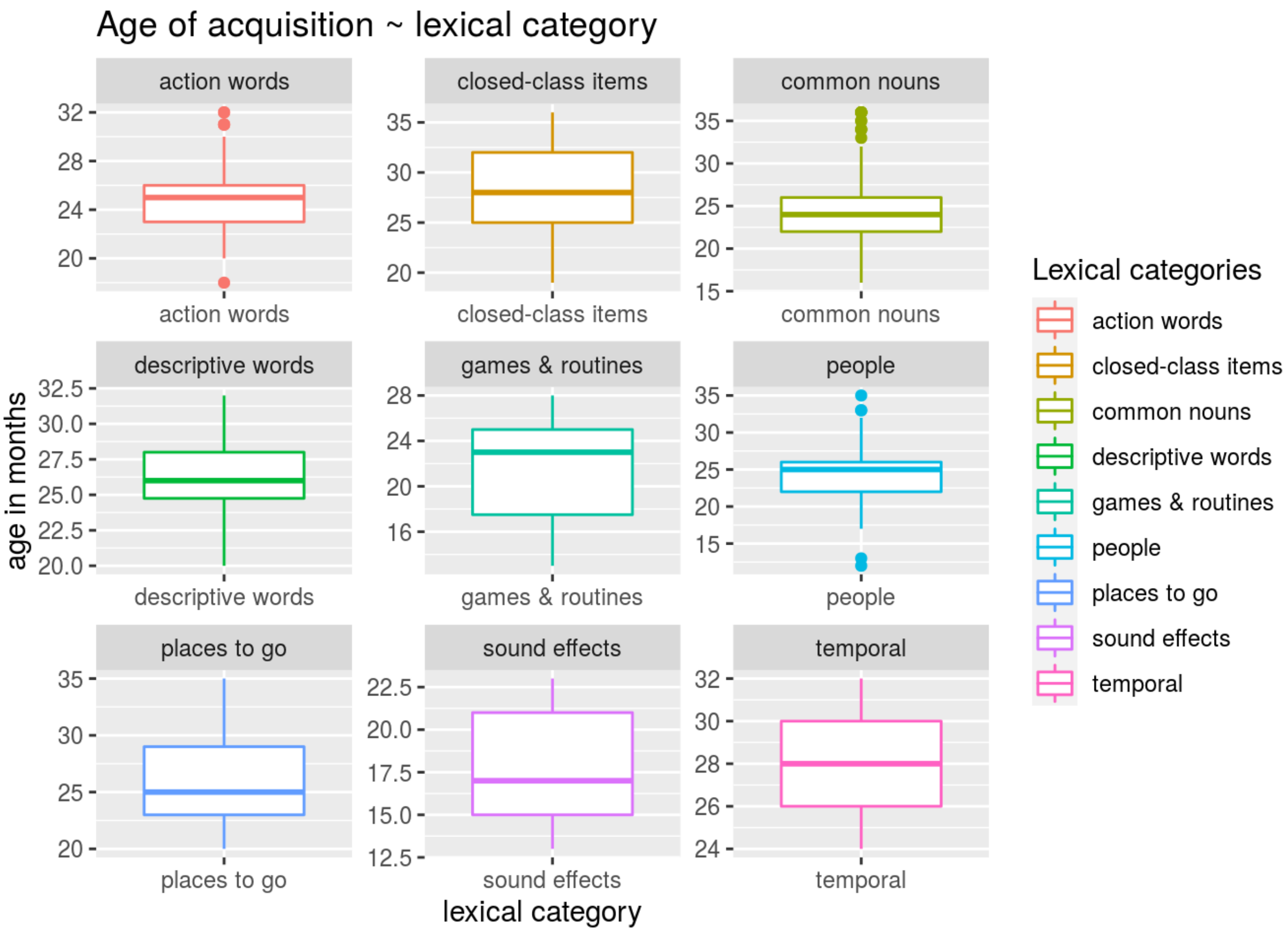
```
# [1] 22
```

The boxplots below demonstrate outliers (categories children reproduce earlier or later than majority of the studied words) separated by categories.

```
p <- ggplot(main_data, aes(Broad_lex, AoA, color=Broad_lex))
p + geom_boxplot() + facet_wrap(~Broad_lex, scale="free") +
        labs(title = "Age of acquisition ~ broad lexical category
",
       x = "broad lexical category",
       y = "age in months") + scale_colour_discrete("Broad lexica
l categories")
```

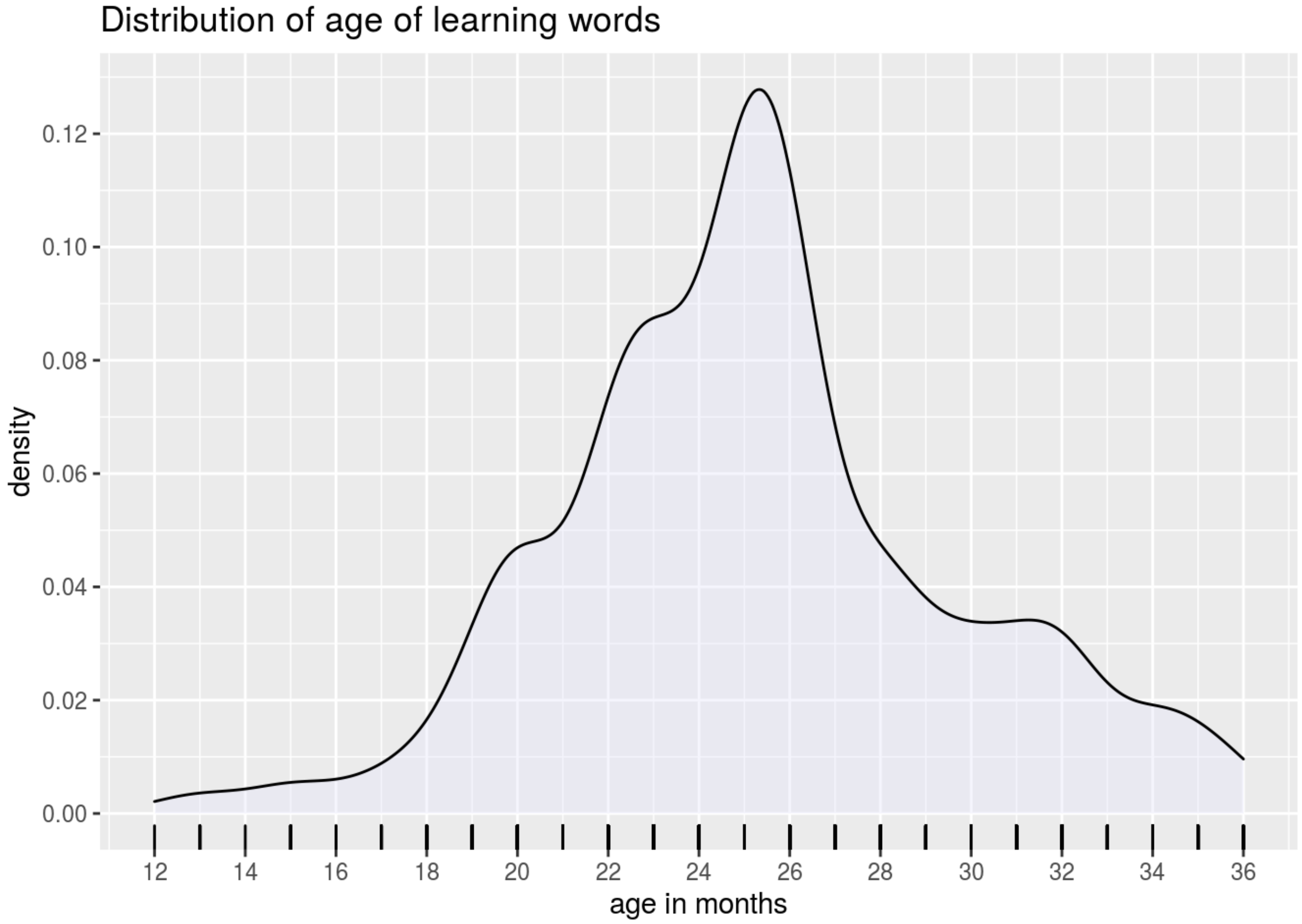


```
p <- ggplot(main_data, aes(Lex_cat, AoA, color=Lex_cat))
p + geom_boxplot() + facet_wrap(~Lex_cat, scale="free") +
        labs(title = "Age of acquisition ~ lexical category",
       x = "lexical category",
       y = "age in months") + scale_colour_discrete("Lexical cate
gories")
```



The distribution of children’ ages.

```
ggplot(main_data, aes(x = AoA)) +
  geom_density(alpha=0.4, fill="lavender")+
  geom_rug()+
  labs(title = "Distribution of age of learning words",
       x = "age in months")+
  scale_x_continuous(breaks=seq(12,36,2)) +
  scale_y_continuous(breaks=seq(0,0.2,0.02))
```



Statistical hypotheses to prove

The main hypotheses of the piece of research are:

- children commence to reproduce words of closed class after acquisition of nouns, verbs and adjectives;

- frequency of words in child directed speech and adults' speech within some classes have strong correlation with age of acquisition.

The statistical hypotheses based on the linguistic ones are:

First

- **H0**: ages of acquisition of different classes are similar.
- **H1**: ages of acquisition of different classes have variations.

Second

- **H0**: there appears to be no correlation between frequencies of categories in parents' speech and age of words' acquisition, the true correlation Pearson's coefficient is 0.
- **H1**: there appears to be correlation between frequencies of categories in parents' speech and age of words' acquisition, the true correlation Pearson's

coefficient is not 0.

Third

- **H0**: there appears to be no linear (negative or positive) association between frequencies of categories in parents' speech and age of words' acquisition.
- **H1**: there appears to be linear (negative or positive) association between frequencies of categories in parents' speech and age of words' acquisition.

Both second and third hypotheses are connected with the second task of the study.

Statistical tools and methods

The author of the study hereto assert the first hypotheses with analysis of variance (*ANOVA*) method. *ANOVA* permits to show connections between more than 2 groups which is benefit for our main factors - parts of speech.

The second hypothesis is examined with *Pearson's correlation coefficient*. As we observe above, the data have ties and small amount of outliers, so it appears to be the most suitable evaluation of connection between word frequencies and age.

It is decided to use *linear regression* for understanding and analyzing if the connection exists and how strong it is.

Connection between part of speech and age (ANOVA)

Preprocessing for the first part

```
main_data %>% filter(Lex_cat=="action words")-> verbs
main_data %>% filter(Lex_cat=="descriptive words")-> adjectives
main_data %>% filter(Broad_lex=="closed-class")-> closedclass

main_data %>% filter(Lex_cat=="common nouns")-> common_nouns
main_data %>% filter(Lex_cat=="places to go")-> places_to_go
main_data %>% filter(Lex_cat=="people")-> people

bind_rows(common_nouns, places_to_go, people) -> nouns
bind_rows(nouns, verbs, adjectives, closedclass) -> target_cats

head(target_cats)
```

```
## # A tibble: 6 x 11
##   ID_CDI_I ID_CDI_II Word_NW Word_CDI Translation   AoA  VSoA
Lex_cat Broad_lex
##   <chr>    <chr>     <chr>   <fct>    <chr>       <dbl> <dbl>
<fct>   <fct>
## 1 i_5_1    i_2_2     'ei an… 'and'    'duck'         24   240
common… nominals
## 2 i_5_3    i_2_3     'en ap… 'apekat… 'monkey'       22   200
common… nominals
## 3 i_5_4    i_2_4     'en bj… 'bjørn'  'bear'         23   240
common… nominals
## 4 i_5_5    i_2_5     'et dy… 'dyr'    'animal'       25   340
common… nominals
## 5 i_5_6    i_2_6     'et ek… 'ekorn'  'squirrel'     25   380
common… nominals
## 6 i_5_7    i_2_7     'en el… 'elefan… 'elephant'     22   200
common… nominals
## # … with 2 more variables: Freq <dbl>, CDS_freq <dbl>
```

First attempt with suggested categories

```
categories_acquisition <- aov(target_cats$AoA ~ target_cats$Lex_c
at)
categories_acquisition
```

```
## Call:
##    aov(formula = target_cats$AoA ~ target_cats$Lex_cat)
##
## Terms:
##                 target_cats$Lex_cat Residuals
## Sum of Squares             1413.200  9469.518
## Deg. of Freedom                   5       629
##
## Residual standard error: 3.880062
## Estimated effects may be unbalanced
```

```
summary(categories_acquisition)
```

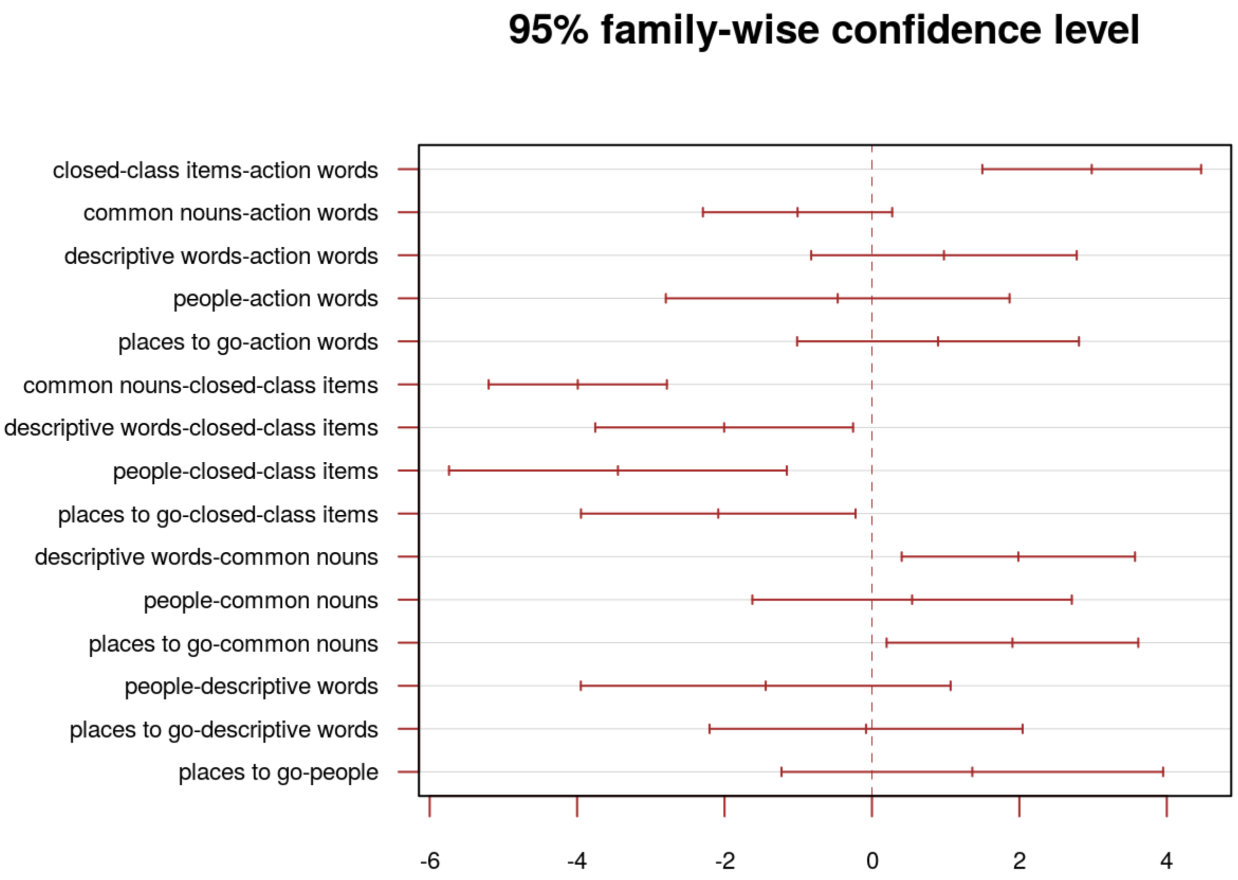
```
##                       Df Sum Sq Mean Sq F value Pr(>F)    
## target_cats$Lex_cat    5   1413  282.64   18.77 <2e-16 ***
## Residuals            629   9470   15.05                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value (<2e-16) for F statistics is less than our significance level, so we reject the null hypothesis. Mean ages of acquisition of some different classes are not the same. We can check confidence values and visualise them.

```
tukey <- TukeyHSD(x=categories_acquisition, conf.level=0.95)
tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = target_cats$AoA ~ target_cats$Lex_cat)
##
## $`target_cats$Lex_cat`
##                                            diff        lwr
upr     p adj
## closed-class items-action words       2.9816171  1.4974820  4.
4657521 0.0000002
## common nouns-action words            -1.0092788 -2.2931257  0.
2745681 0.2178133
## descriptive words-action words        0.9766990 -0.8245770  2.
7779750 0.6319050
## people-action words                  -0.4646803 -2.7962588  1.
8668983 0.9929367
## places to go-action words             0.8966990 -1.0150144  2.
8084125 0.7620841
## common nouns-closed-class items      -3.9908959 -5.2001356 -2.
7816562 0.0000000
## descriptive words-closed-class items -2.0049180 -3.7538014 -0.
2560347 0.0140457
## people-closed-class items            -3.4462973 -5.7376411 -1.
1549535 0.0002857
## places to go-closed-class items      -2.0849180 -3.9473482 -0.
2224878 0.0180114
## descriptive words-common nouns        1.9859779  0.4035117  3.
5684440 0.0048255
## people-common nouns                   0.5445985 -1.6223953  2.
7115924 0.9797042
## places to go-common nouns             1.9059779  0.1988598  3.
6130960 0.0185049
## people-descriptive words             -1.4413793 -3.9498041  1.
0670454 0.5705359
## places to go-descriptive words       -0.0800000 -2.2038137  2.
0438137 0.9999980
## places to go-people                   1.3613793 -1.2274907  3.
9502493 0.6623698
```

```
par(mar=c(1,7,1,1)+3.5,mgp=c(6,1,0))
plot(tukey, las=1, col="brown", cex.lab=2.5, cex.axis=0.7)
```



The diagram and results of TukeyHSD analysis prove that some categories have not significant differences in age of learning: *common nouns-action words*, *descriptive words-action words*, *places to go-action words*, *people-common nouns*, *people-descriptive words*, *places to go-descriptive words* and *places to go-people*. *people-common nouns* and *places to go-people* could be explained by the fact that all words from that groups are nouns. But *common nouns-action words*, *descriptive words-action words*, *places to go-action words*, *people-descriptive words* and *places to go-descriptive words* are interesting cases because words from that pairs belong to different parts of speech. We are able to assume that in some cases part of speech or lexical category of some words do not influence age of reproducing that words.

Preprocessing for the second part

Estimates of **Main factors** part, some categories of words should be renamed for further analysis.

```
main_data <- read_csv('https://raw.githubusercontent.com/wksmirnowa/LDA_project/master/main_data.csv')
```

```
## Parsed with column specification:
## cols(
##   ID_CDI_I = col_character(),
##   ID_CDI_II = col_character(),
##   Word_NW = col_character(),
##   Word_CDI = col_character(),
##   Translation = col_character(),
##   AoA = col_character(),
##   VSoA = col_character(),
##   Lex_cat = col_character(),
##   Broad_lex = col_character(),
##   Freq = col_character(),
##   CDS_freq = col_character()
## )
```

```
main_data$Lex_cat <- gsub('#N/A', 'temporal', main_data$Lex_cat)
main_data$Broad_lex <- gsub('#N/A', 'temporal', main_data$Broad_lex)
main_data$Word_CDI <- as.factor(main_data$Word_CDI)
main_data$Lex_cat <- gsub('action words', 'verbs', main_data$Lex_cat)
main_data$Lex_cat <- gsub('descriptive words', 'adjectives', main_data$Lex_cat)
main_data$Lex_cat <- gsub('closed-class items', 'closedclass', main_data$Lex_cat)
main_data$Lex_cat <- gsub('common nouns', 'nouns', main_data$Lex_cat)
main_data$Lex_cat <- gsub('places to go', 'nouns', main_data$Lex_cat)
main_data$Lex_cat <- gsub('people', 'nouns', main_data$Lex_cat)

main_data$Broad_lex <- as.factor(main_data$Broad_lex)
main_data$Lex_cat <- as.factor(main_data$Lex_cat)
```

```
main_data$AoA <- as.numeric(main_data$AoA)
main_data$VSoA <- as.numeric(main_data$VSoA)
main_data$CDS_freq <- as.numeric(main_data$CDS_freq)
main_data$Freq <- as.numeric(main_data$Freq)
main_data<-na.omit(main_data)
```

```
main_data %>% filter(Lex_cat=="verbs")-> verbs
main_data %>% filter(Lex_cat=="adjectives")-> adjectives
main_data %>% filter(Lex_cat=="closedclass")-> closedclass
main_data %>% filter(Lex_cat=="nouns")-> nouns

bind_rows(nouns, verbs, adjectives, closedclass) -> target_cats_n
ew
head(target_cats_new)
```

```
## # A tibble: 6 x 11
##   ID_CDI_I ID_CDI_II Word_NW Word_CDI Translation   AoA  VSoA
Lex_cat Broad_lex
##   <chr>    <chr>     <chr>   <fct>    <chr>       <dbl> <dbl>
<fct>   <fct>
## 1 i_5_1    i_2_2     'ei an… 'and'    'duck'         24   240
nouns   nominals
## 2 i_5_3    i_2_3     'en ap… 'apekat… 'monkey'       22   200
nouns   nominals
## 3 i_5_4    i_2_4     'en bj… 'bjørn'  'bear'         23   240
nouns   nominals
## 4 i_5_5    i_2_5     'et dy… 'dyr'    'animal'       25   340
nouns   nominals
## 5 i_5_6    i_2_6     'et ek… 'ekorn'  'squirrel'     25   380
nouns   nominals
## 6 i_5_7    i_2_7     'en el… 'elefan… 'elephant'     22   200
nouns   nominals
## # … with 2 more variables: Freq <dbl>, CDS_freq <dbl>
```

Second attempt with parts of speech

```
categories_acquisition <- aov(target_cats_new$AoA ~ target_cats_n
ew$Lex_cat)
categories_acquisition
```

```
## Call:
##    aov(formula = target_cats_new$AoA ~ target_cats_new$Lex_cat
)
##
## Terms:
##                 target_cats_new$Lex_cat Residuals
## Sum of Squares                 1258.223  9624.495
## Deg. of Freedom                       3       631
##
## Residual standard error: 3.905479
## Estimated effects may be unbalanced
```

```
summary(categories_acquisition)
```

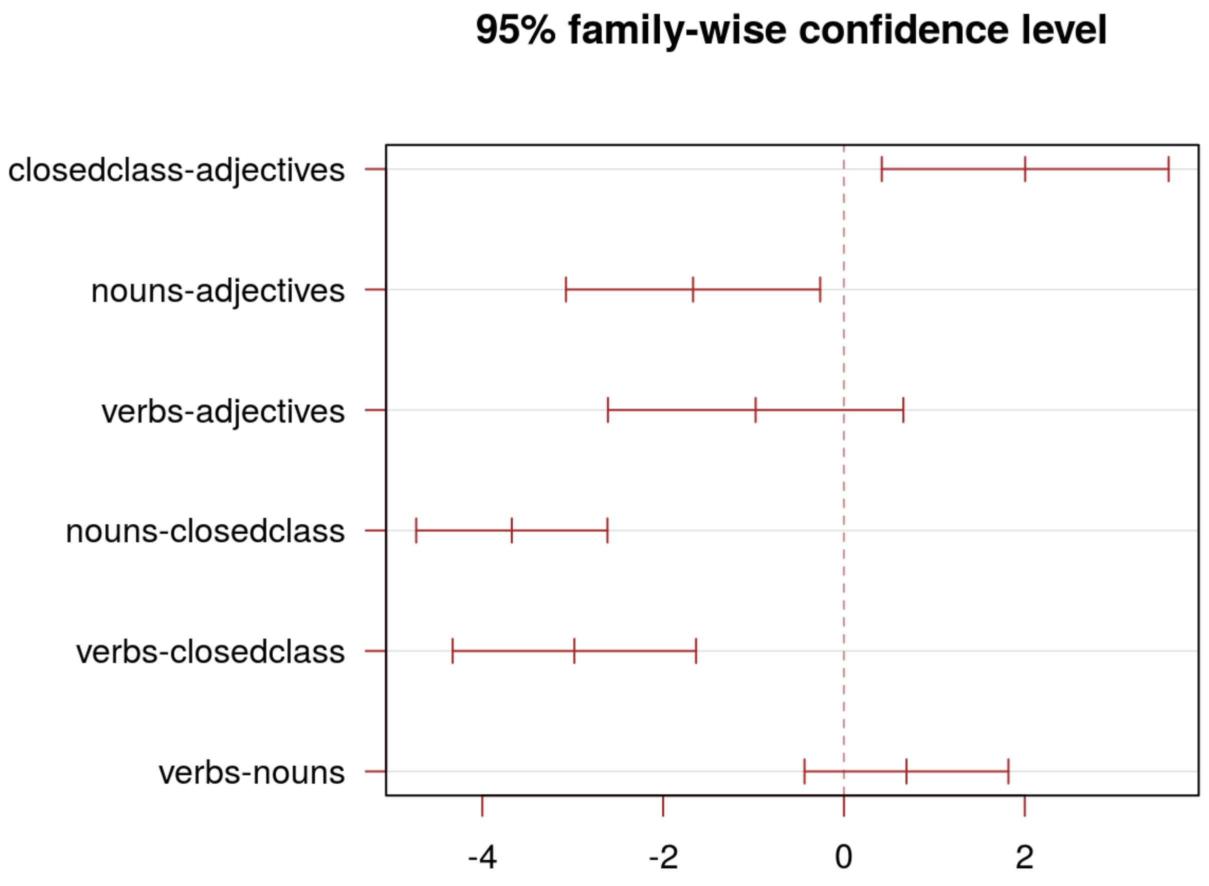
```
##                          Df Sum Sq Mean Sq F value Pr(>F)    
## target_cats_new$Lex_cat   3   1258   419.4    27.5 <2e-16 ***
## Residuals               631   9624    15.3                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value (<2e-16) for F statistics is same. It is still less than our significance level, so we reject the null hypothesis again. Mean ages of acquisition of some classes are not the same.

```
tukey <- TukeyHSD(x=categories_acquisition, conf.level=0.95)
tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = target_cats_new$AoA ~ target_cats_new$Lex_c
at)
##
## $`target_cats_new$Lex_cat`
##                               diff        lwr        upr     p
adj
## closedclass-adjectives  2.0049180  0.4186310  3.5912051 0.0065
286
## nouns-adjectives       -1.6685714 -3.0742431 -0.2628997 0.0124
182
## verbs-adjectives       -0.9766990 -2.6105077  0.6571096 0.4142
345
## nouns-closedclass      -3.6734895 -4.7311802 -2.6157988 0.0000
000
## verbs-closedclass      -2.9816171 -4.3277698 -1.6354643 0.0000
001
## verbs-nouns             0.6918724 -0.4358393  1.8195841 0.3905
872
```

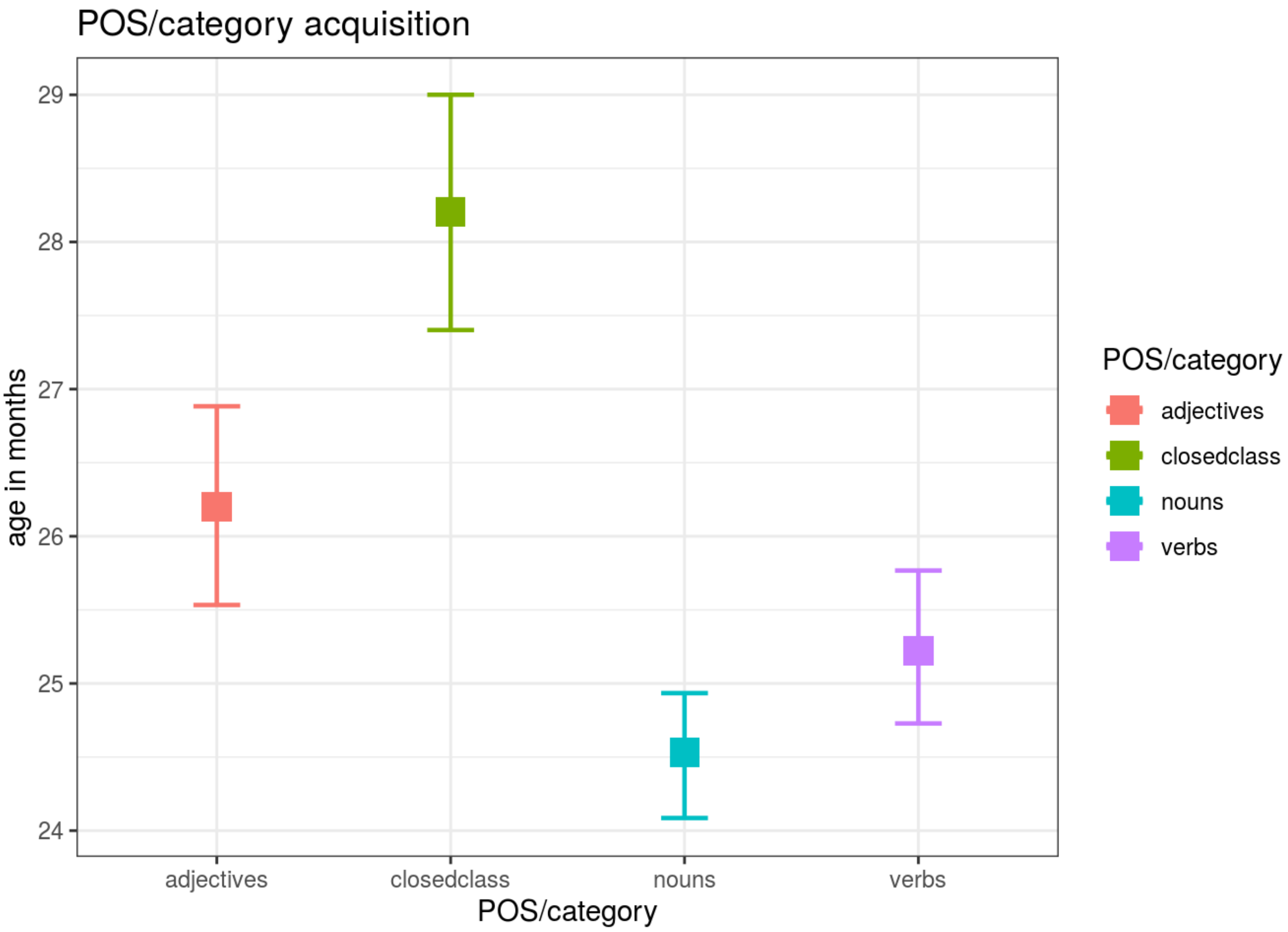
```
par(mar=c(1,7,1,1)+3.5,mgp=c(6,1,0))
plot(tukey , las=1 , col="brown")
```



On this graph *verbs-nouns* and *verbs-adjectives* intervals cross the 0 line, and adjusted p-values of that pairs confirm that difference between that parts of speech is not statistical significant. It could be explained by the fact that children might acquire verbs together with nouns and adjectives in collocations and sentences. But other parts of speech demonstrate that some categories are reproduced by children of different ages. Here is the visualisation of connection between parts of speech and their acquisition.

```
pd = position_dodge(0.4)
ggplot(target_cats_new, aes(Lex_cat, AoA, color = Lex_cat)) +
stat_summary(fun.data = mean_cl_boot, geom = 'errorbar', width =
0.2, lwd = 0.8, position = pd)+
    stat_summary(fun.data = mean_cl_boot, geom = 'line', size = 1
.5, position = pd) +
    stat_summary(fun.data = mean_cl_boot, geom = 'point', size =
5, position = pd, pch=15) +
    theme_bw() +
   ggtitle("POS/category acquisition")+
    xlab('POS/category') +
    ylab('age in months') + scale_colour_discrete("POS/category")
```

```
## geom_path: Each group consists of only one observation. Do you
need to adjust
## the group aesthetic?
```



Correlation

Preprocessing

```
main_data <- read_csv('https://raw.githubusercontent.com/wksmirnowa/LDA_project/master/main_data.csv')
main_data$Lex_cat <- gsub('#N/A', 'temporal', main_data$Lex_cat)
main_data$Broad_lex <- gsub('#N/A', 'temporal', main_data$Broad_lex)
main_data$Word_CDI <- as.factor(main_data$Word_CDI)
main_data$Lex_cat <- gsub('action words', 'verbs', main_data$Lex_cat)
main_data$Lex_cat <- gsub('descriptive words', 'adjectives', main_data$Lex_cat)
main_data$Lex_cat <- gsub('closed-class items', 'closedclass', main_data$Lex_cat)

main_data$Lex_cat <- gsub('common nouns', 'nouns', main_data$Lex_cat)
main_data$Lex_cat <- gsub('places to go', 'nouns', main_data$Lex_cat)
main_data$Lex_cat <- gsub('people', 'nouns', main_data$Lex_cat)

main_data$Broad_lex <- as.factor(main_data$Broad_lex)
main_data$Lex_cat <- as.factor(main_data$Lex_cat)
```

```
main_data$AoA <- as.numeric(main_data$AoA)
main_data$VSoA <- as.numeric(main_data$VSoA)
main_data$CDS_freq <- as.numeric(main_data$CDS_freq)
main_data$Freq <- as.numeric(main_data$Freq)
main_data<-na.omit(main_data)
```

In the tests below the author investigates if there appears to be correlation between frequencies of words in adults’ speech and age, frequencies of words in child directed speech and age, both of that frequencies and age.

```
cor.test(main_data$Freq, main_data$AoA)
```

```
##
##  Pearson's product-moment correlation
##
## data:  main_data$Freq and main_data$AoA
## t = 2.1293, df = 684, p-value = 0.03359
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.006330421 0.155061372
## sample estimates:
##        cor
## 0.08114762
```

```
cor.test(main_data$CDS_freq, main_data$AoA)
```

```
##
##  Pearson's product-moment correlation
##
## data:  main_data$CDS_freq and main_data$AoA
## t = -0.6355, df = 684, p-value = 0.5253
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09896747  0.05065598
## sample estimates:
##         cor
## -0.02429178
```

```
cor.test(main_data$Freq+main_data$CDS_freq, main_data$AoA)
```

```
##
##  Pearson's product-moment correlation
##
## data:  main_data$Freq + main_data$CDS_freq and main_data$AoA
## t = 2.1288, df = 684, p-value = 0.03363
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.006311961 0.155043354
## sample estimates:
##        cor
## 0.08112928
```

The results are:

1. Pearson’s coefficient 0.08114762 and p-value 0.03359 for *Freq* and *AoA*
2. Pearson’s coefficient -0.02429178 and p-value 0.5253 for *CDS_freq* and *AoA*
3. Pearson’s coefficient 0.08112928 and p-value 0.03363 for *Freq* + *CDS_freq* and *AoA*

The first result proves that there appears to be correlation, but the coefficient is close to zero value. Though p-value is less than 0.05, it is rather close to 0.05, so the correlation is not strong. Note that it is observed similar situation in the third result. The null hypothesis in these results should be rejected. There appears to be not strong correlation between this factors. The author emphasises that the second general hypothesis does not approved with this results. It was suggested that toddlers have higher probability to learn more frequent words earlier but contrasting results are demonstrated.

The second p-value is more than 0.05. It means that we should accept the null hypothesis and deny the second main hypothesis in case of child directed speech frequency.

Linear regression

Model with frequency of words in adults’ speech

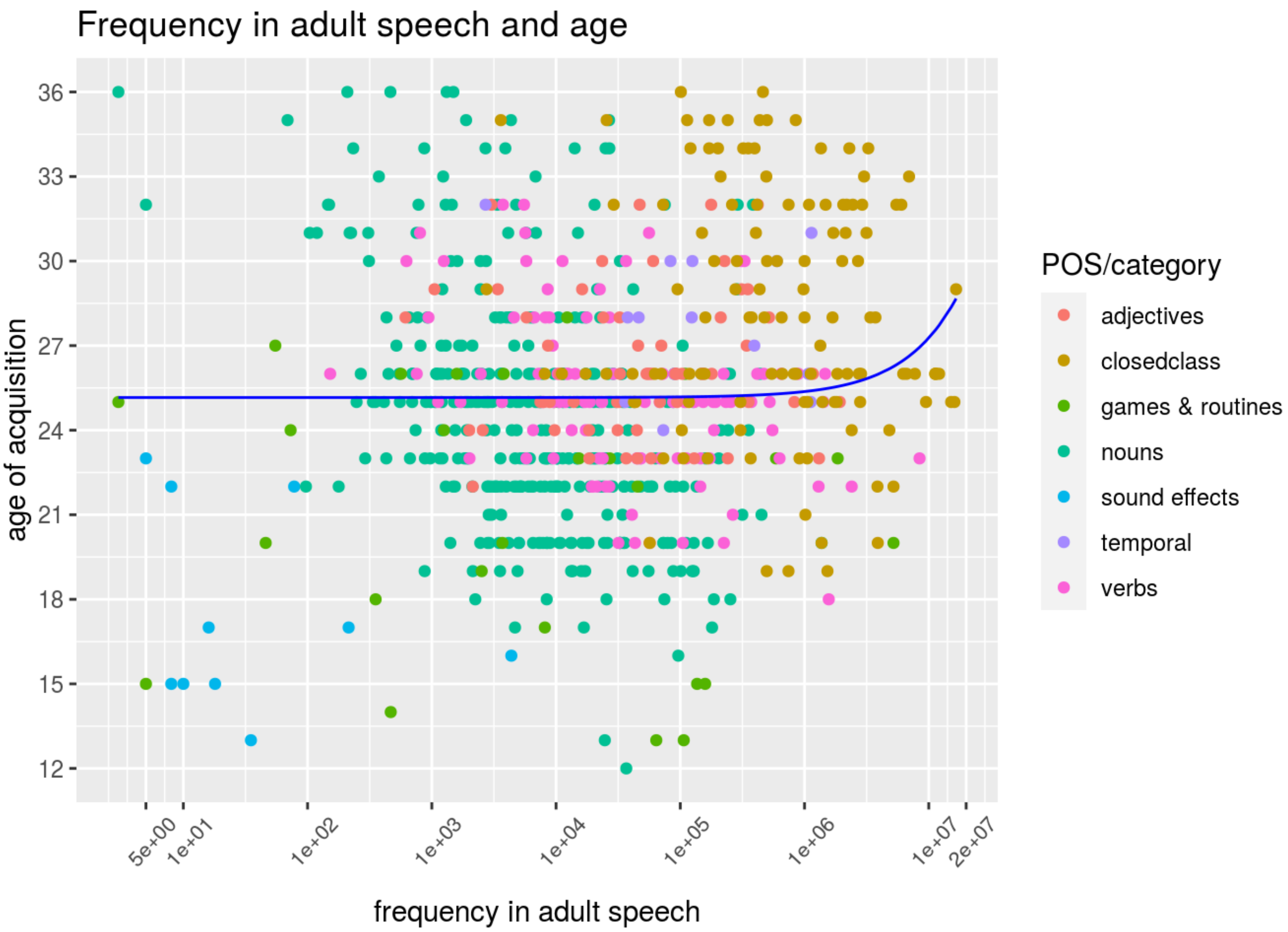
```
require("scales")
linreg1 <- lm(data = main_data, AoA ~ Freq)
summary(linreg1)
```

```
##
## Call:
## lm(formula = AoA ~ Freq, data = main_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1670  -3.0070  -0.1618   2.7561  10.8408
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.516e+01  1.713e-01 146.884   <2e-16 ***
## Freq        2.115e-07  9.935e-08   2.129   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.324 on 684 degrees of freedom
## Multiple R-squared:  0.006585,   Adjusted R-squared:  0.005133
## F-statistic: 4.534 on 1 and 684 DF,  p-value: 0.03359
```

```
main_data$linreg1 <- predict(linreg1)

base_breaks <- function(n = 10){
    function(x) {
        axisTicks(log10(range(x, na.rm = TRUE)), log = TRUE, n =
n)
    }
}

main_data %>%
  ggplot(aes(Freq, AoA, color=Lex_cat))+
  geom_point()+
  geom_line(aes(Freq, linreg1), color = "blue")+
  labs(x = "frequency in adult speech",
       y = "age of acquisition",
       title = "Frequency in adult speech and age")+ scale_colour
_discrete("POS/category") +
  theme(axis.text.x = element_text(size=8, angle=47, margin = mar
gin(t=10)))+
  scale_y_continuous(breaks=seq(0,60,3))+
  scale_x_continuous(trans = log_trans(), breaks=c(1,5, 10, 100,
1e3, 1e4, 1e5, 1e6, 1e7, 2e7))
```

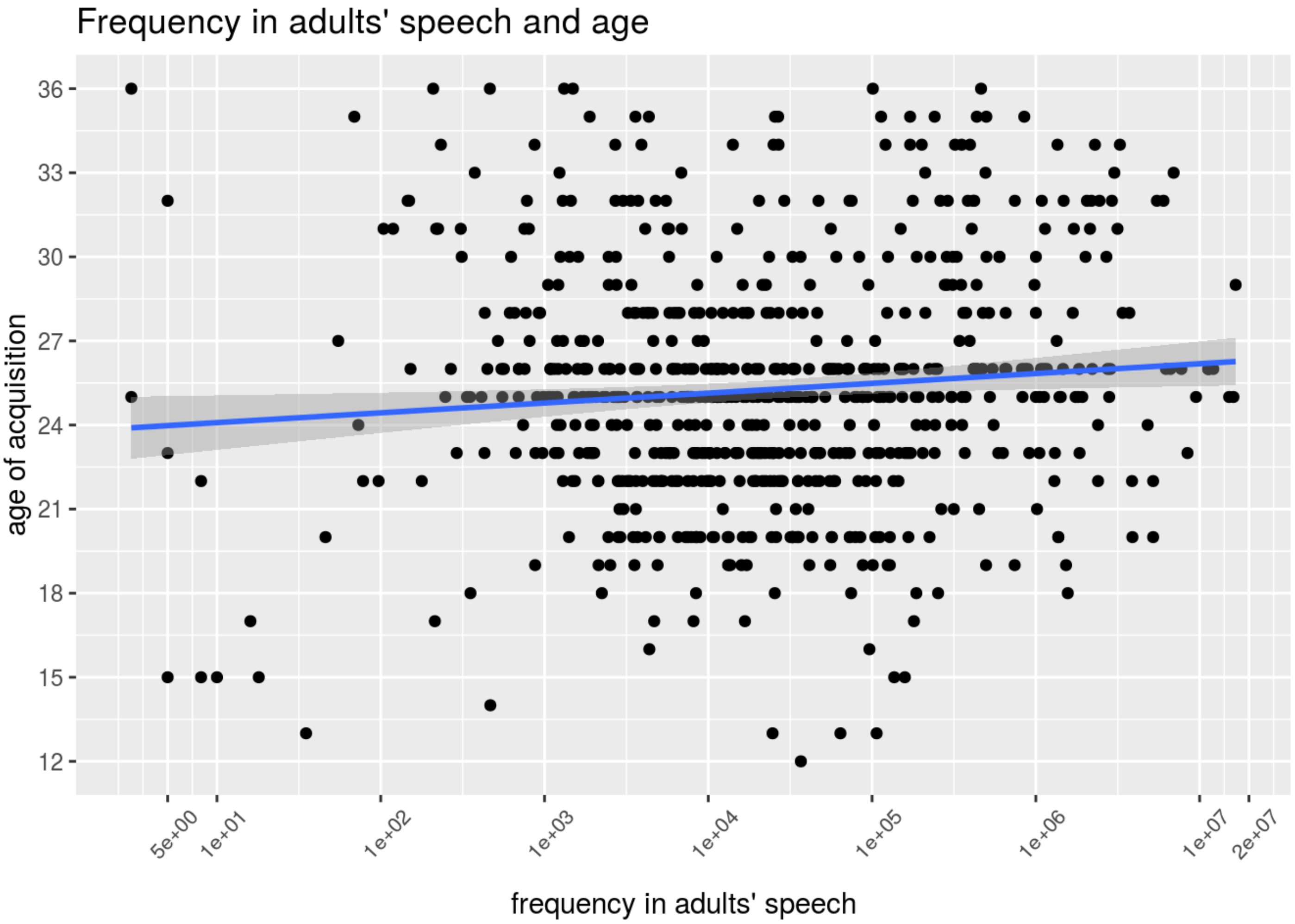


The provided graph represents that there appears to be large spread of frequencies' values. It is a problem for finding the most reliable regression's line and verifing the second general hypothesis. As we are able to notice, p-value for F-statistics is 0.03359 and p-value for influece of frequencies is 0.0336. It is statistically significant but close to critical value.

The graph with straight regression line and confidence interval is demonstrated below.

```
ggplot(data = main_data, aes(x = Freq, y = AoA)) +
  geom_point() +
  labs(x = "frequency in adults' speech",
       y = "age of acquisition",
       title = "Frequency in adults' speech and age") +
  geom_smooth(method=lm)+ theme(axis.text.x = element_text(size=8
, angle=47, margin = margin(t=10)))+
  scale_y_continuous(breaks=seq(0,60,3))+
  scale_x_continuous(trans = log_trans(), breaks=c(1,5, 10, 100,
1e3, 1e4, 1e5, 1e6, 1e7, 2e7))
```

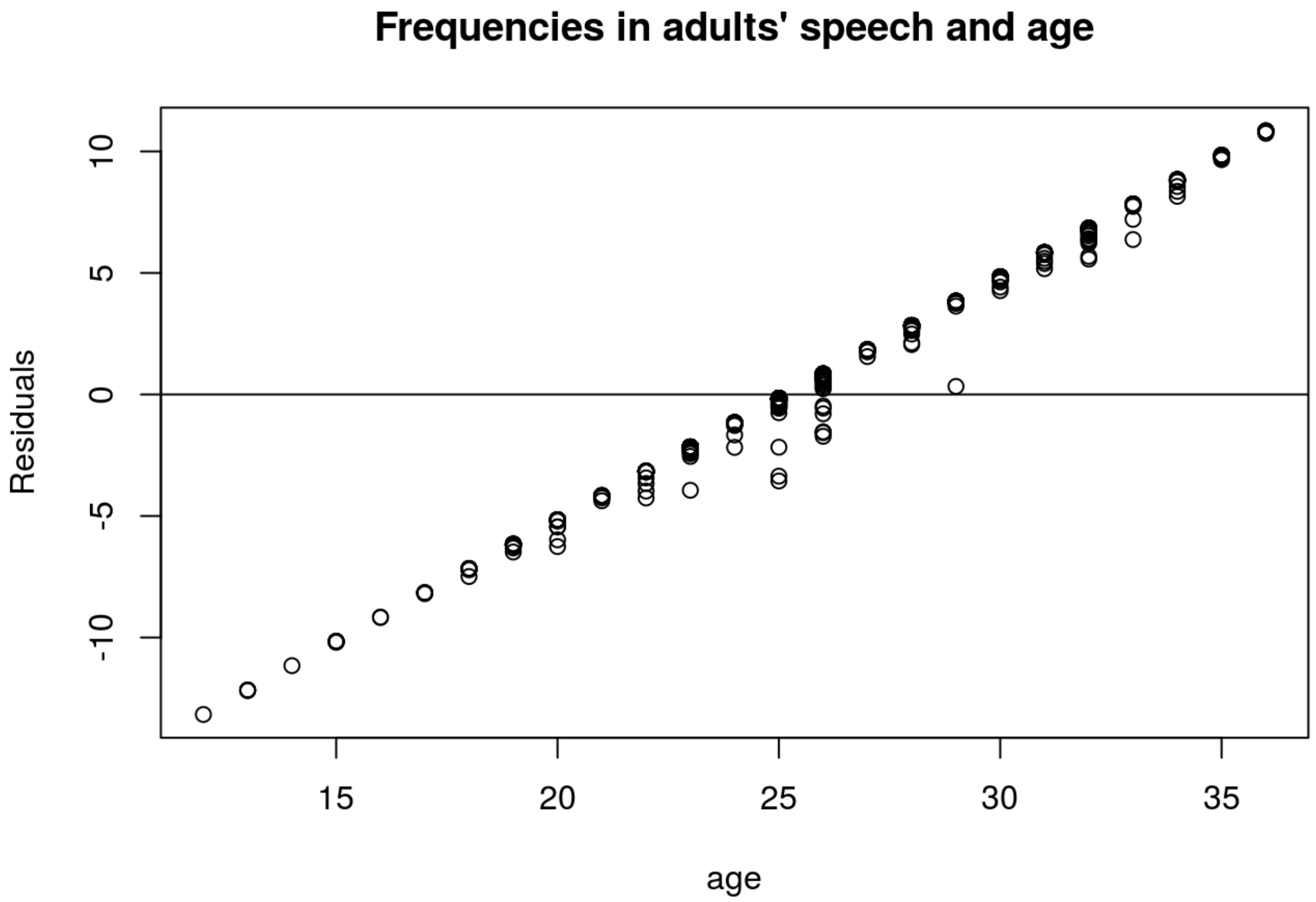
```
## `geom_smooth()` using formula 'y ~ x'
```



The picture below detects residuals. The most part of the main data is below or above zero line. It proves that linear regression line poorly represents our data.

```
linreg1.resid = resid(linreg1)

plot(main_data$AoA, linreg1.resid,
     ylab="Residuals", xlab="age",
     main="Frequencies in adults' speech and age")
abline(0, 0)
```



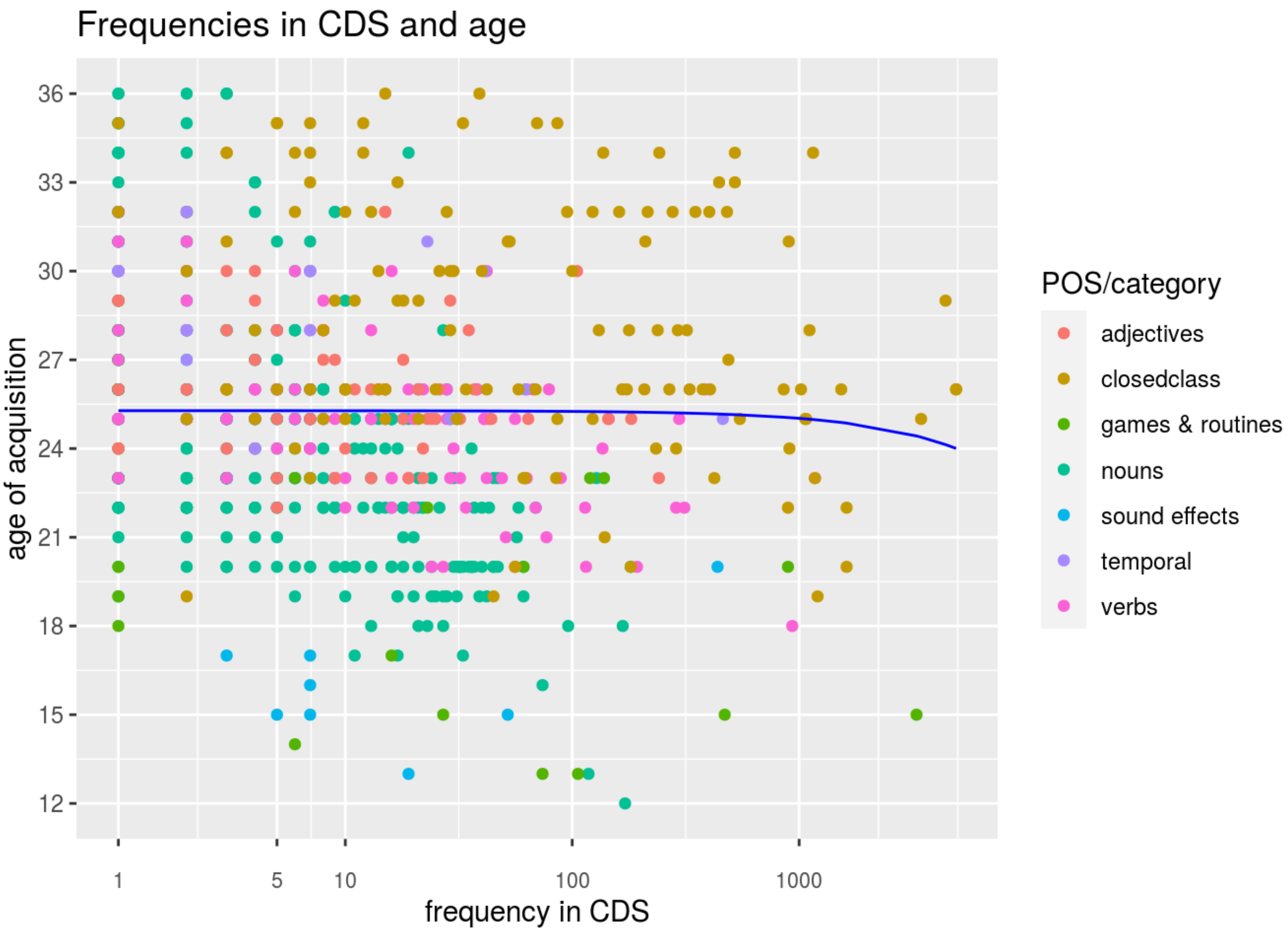
Model with frequency of words in CDS

```
linreg2 <- lm(data = main_data, AoA ~ CDS_freq)
summary(linreg2)
```

```
## 
## Call:
## lm(formula = AoA ~ CDS_freq, data = main_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -13.236  -2.280  -0.278   2.720  10.730 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 25.2804390  0.1697816 148.900   <2e-16 ***
## CDS_freq    -0.0002599  0.0004090  -0.635    0.525    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.337 on 684 degrees of freedom
## Multiple R-squared:  0.0005901,  Adjusted R-squared:  -0.00087
1 
## F-statistic: 0.4039 on 1 and 684 DF,  p-value: 0.5253
```

```
main_data$linreg2 <- predict(linreg2)

main_data %>%
  ggplot(aes(CDS_freq, AoA, color=Lex_cat))+
  geom_point()+
  geom_line(aes(CDS_freq, linreg2), color = "blue")+
  labs(x = "frequency in CDS",
       y = "age of acquisition",
       title = "Frequencies in CDS and age")+ scale_colour_discre
te("POS/category") +
  theme(axis.text.x = element_text(size=8, margin = margin(t=10))
)+
  scale_y_continuous(breaks=seq(0,60,3))+
  scale_x_continuous(trans = log_trans(), breaks=c(1,5, 10, 100,
1e3, 1e4, 1e5, 1e6, 1e7, 2e7))
```

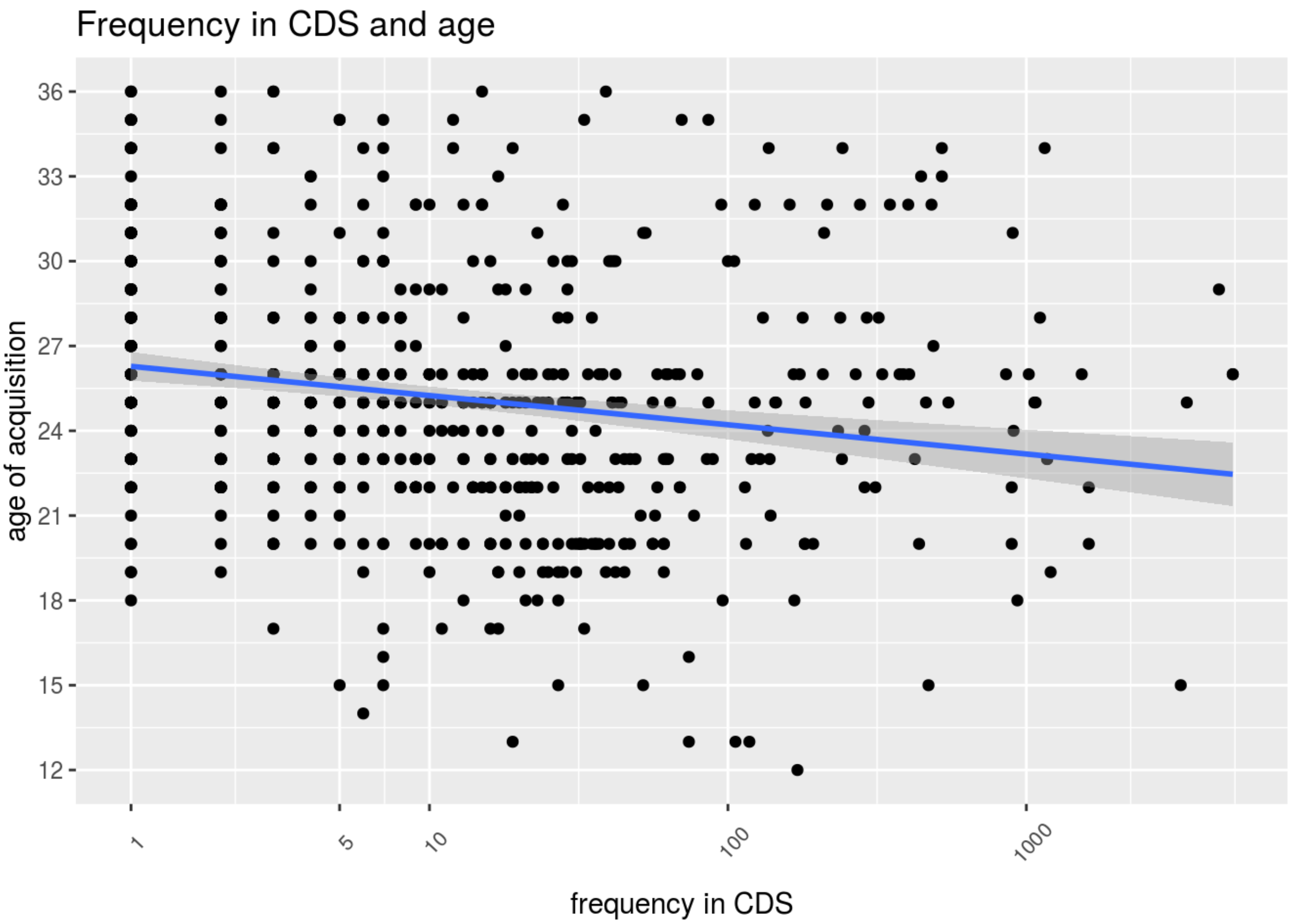


The picture displays that CDS frequencies also have large spread of values. High level of p-value claims that there does not appear to be a significant relation between frequency of word in CDS and age of learning of word.

The similar graph with straight regression line and confidence interval is presented below.

```
ggplot(data = main_data, aes(x = CDS_freq, y = AoA)) +
  geom_point() +
  labs(x = "frequency in CDS",
       y = "age of acquisition",
       title = "Frequency in CDS and age") +
  geom_smooth(method=lm)+ theme(axis.text.x = element_text(size=8
, angle=47, margin = margin(t=10)))+
  scale_y_continuous(breaks=seq(0,60,3))+
  scale_x_continuous(trans = log_trans(), breaks=c(1,5, 10, 100,
1e3, 1e4, 1e5, 1e6, 1e7, 2e7))
```

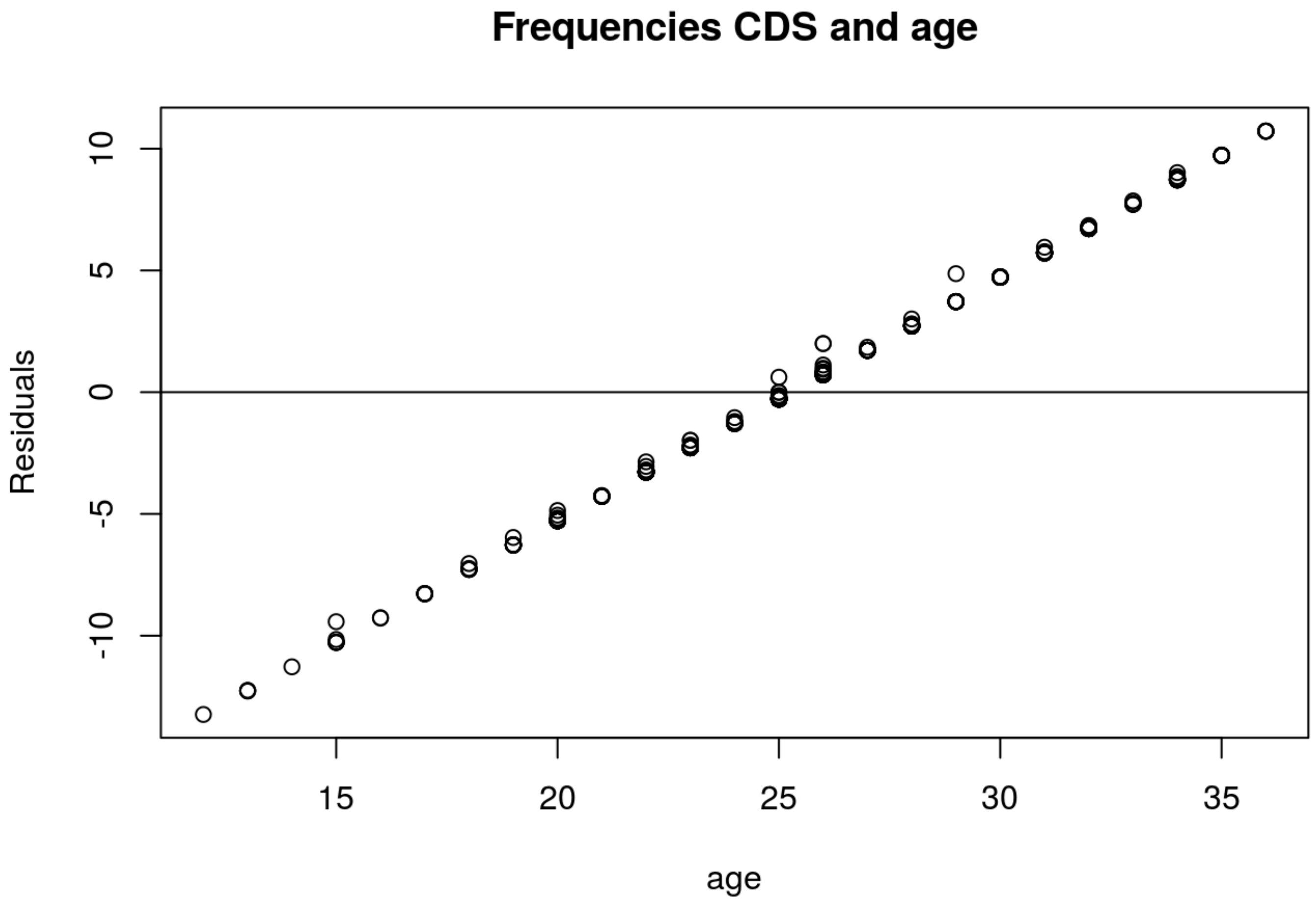
```
## `geom_smooth()` using formula 'y ~ x'
```



Note majority of residuals in the picture below.

```
linreg2.resid = resid(linreg2)

plot(main_data$AoA, linreg2.resid,
     ylab="Residuals", xlab="age",
     main="Frequencies CDS and age")
abline(0, 0)
```

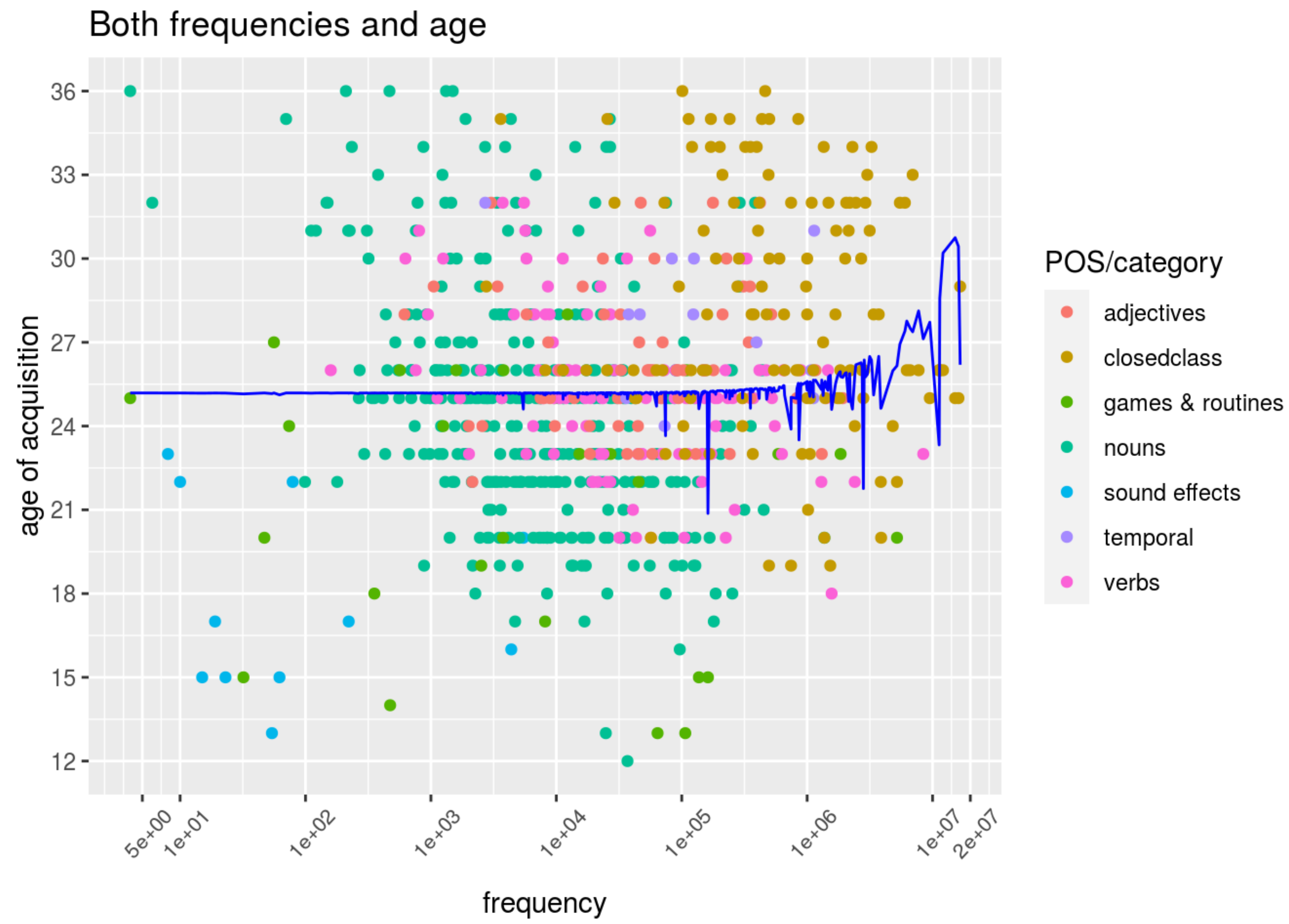


Model with both frequencies

```
linreg3 <- lm(data = main_data, AoA ~ Freq + CDS_freq)
main_data$linreg3 <- predict(linreg3)
summary(linreg3)
```

```
##
## Call:
## lm(formula = AoA ~ Freq + CDS_freq, data = main_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9753  -2.9778  -0.1862   2.7877  10.8161
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.519e+01  1.710e-01 147.324  < 2e-16 ***
## Freq         4.154e-07  1.272e-07   3.265  0.00115 **
## CDS_freq    -1.332e-03  5.222e-04  -2.550  0.01099 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.306 on 683 degrees of freedom
## Multiple R-squared:  0.01595,    Adjusted R-squared:  0.01307
## F-statistic: 5.536 on 2 and 683 DF,  p-value: 0.004119
```

```
ggplot(data = main_data, aes(Freq + CDS_freq, AoA, color=Lex_cat)
)+
  geom_point()+
  geom_line(aes(Freq + CDS_freq, linreg3), color = "blue")+
  labs(x = "frequency",
       y = "age of acquisition",
       title = "Both frequencies and age")+ scale_colour_discrete
("POS/category") +
  theme(axis.text.x = element_text(size=8, angle=47, margin = mar
gin(t=10)))+
  scale_y_continuous(breaks=seq(0,60,3))+
  scale_x_continuous(trans = log_trans(), breaks=c(1,5, 10, 100,
1e3, 1e4, 1e5, 1e6, 1e7, 2e7))
```

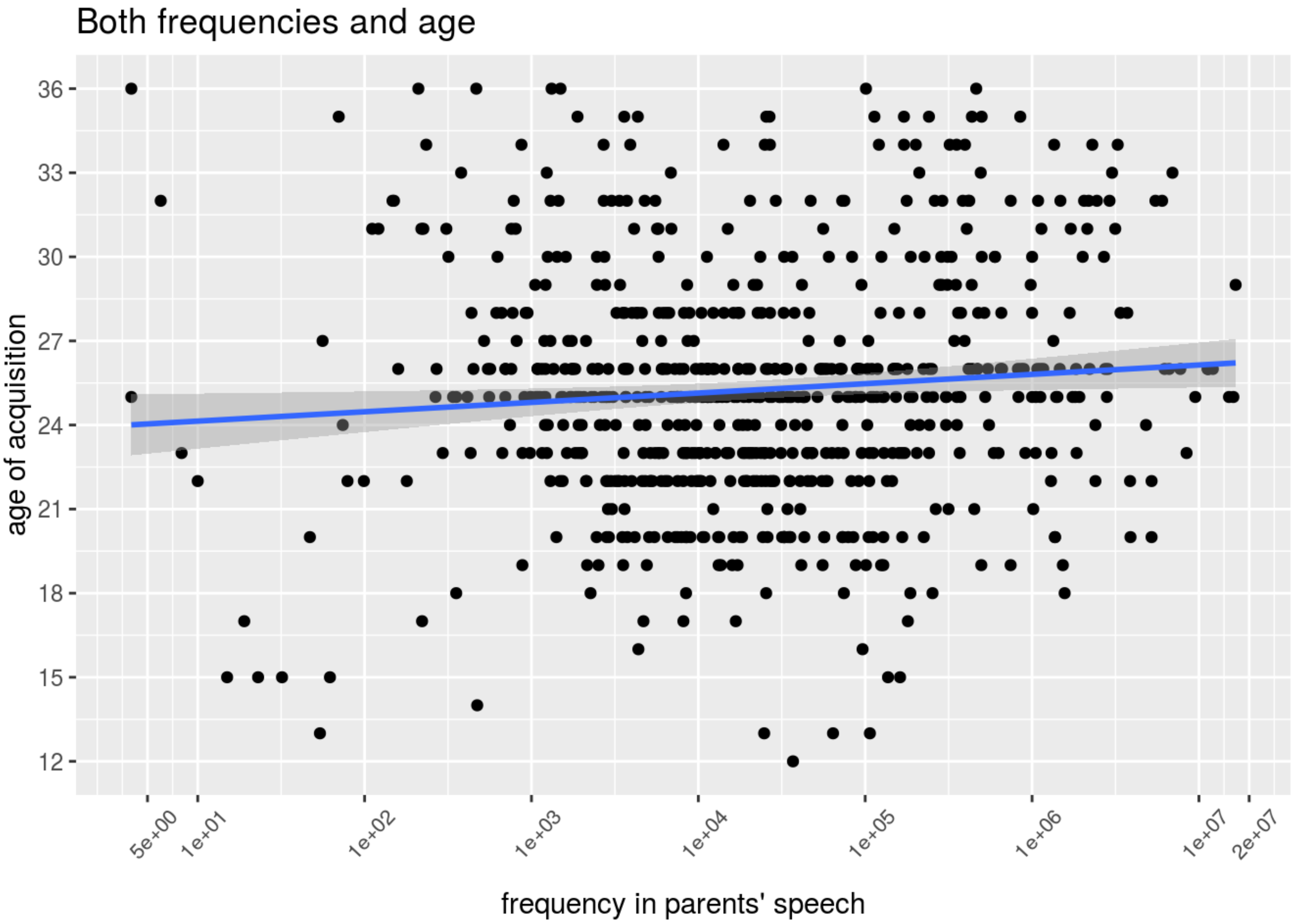


The picture expresses that combination of both frequencies leads to overfitting and errors in usage of linear regression. *683 degrees of freedom* is important factor for this result too.

The possible graph with normal regression line and residuals' picture are given below.

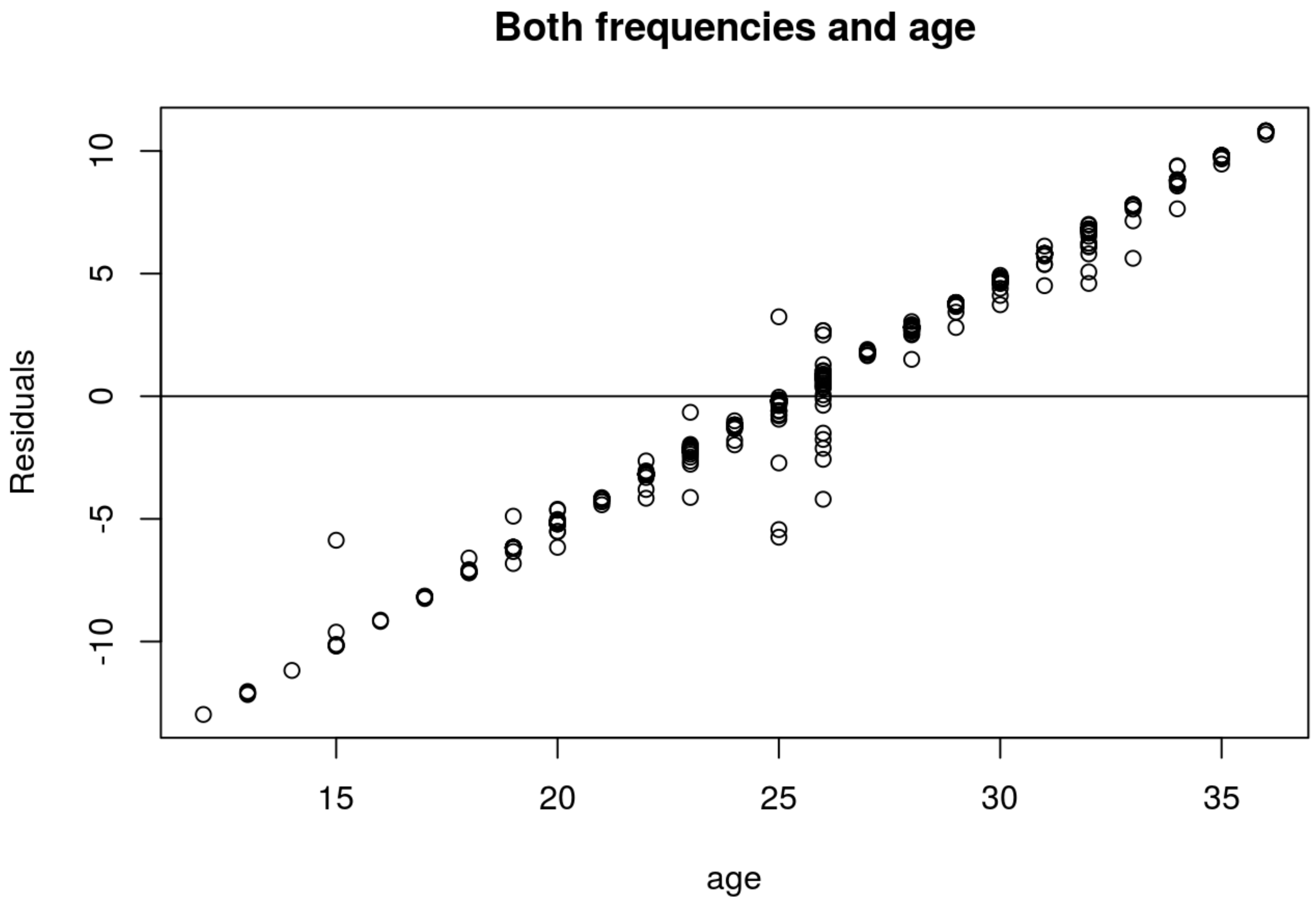
```
ggplot(data = main_data, aes(x = CDS_freq + Freq, y = AoA)) +
  geom_point() +
  labs(x = "frequency in parents' speech",
       y = "age of acquisition",
       title = "Both frequencies and age") +
  geom_smooth(method=lm)+
   theme(axis.text.x = element_text(size=8, angle=47, margin = ma
rgin(t=10)))+
  scale_y_continuous(breaks=seq(0,60,3))+
  scale_x_continuous(trans = log_trans(), breaks=c(1,5, 10, 100,
1e3, 1e4, 1e5, 1e6, 1e7, 2e7))
```

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



```
linreg3.resid = resid(linreg3)

plot(main_data$AoA, linreg3.resid,
     ylab="Residuals", xlab="age",
     main="Both frequencies and age")
abline(0, 0)
```



This information outlines that it is necessary to do future work on studing connection between frequencies and ages with other predictive model. Other possible variant for future work on this problem is collecting more data on this topic.

Explanation of the results

ANOVA test

Possible explanation of *ANOVA* results is that some words are acquired by children in phrases (for instance, ‘kommer og tar deg’ - ‘gonna get you’, the name of the game, includes different parts of speech in one name). In general, results prove that some words children repeat and understand later than others. For this paper it is important to ensure that there appears to be statistically significant connection between part of speech of the word and probability of its earlier or later acquisition. Generally speaking, the author asserts that nouns have higher probability to be memorized earlier than any other POS. Some verbs are learnt with nouns in collocations, but verbs’ acquisition commences later. Clossed class items, such as pronouns and prepositions, are reproduced later then words from opened class categories.

Correlation test and linear regression

The results on CDS test could be explained by the fact that children are able to memorize words accidentally. From child directed speech infants and toddlers might reproduce not the most or the least frequent words, but random ones.

Other results are complicated. To understand why the test and the graphs demonstrate correlation between adults' speech and age we should perform additional researches. One of possible explanations is that it is more important which words infants and toddlers hear from adults' different talks than from talks directed to them. The second one is that adults do not speak among themselves about topics and ideas which they could use in child directed speech. According to this, we are able to guess that words from talks with children are less frequent in adults' speech. The third one is that Norwegian Web corpus is not suitable for evaluating adults' speech. It might be better to use colloquial or media corpuses for more precise results. So, the frequencies of words which adults use in CDS are less in adults' speech than in CDS.

Combination of both frequencies does not give any controversial results towards the previous ones. Linear regression might not be the most suitable method for studing this data because of dispersion in frequencies and other variations in them.

These statements could possibly become contradicted in future researches on this topic with enhanced data.

References

Papers

Hansen, P. (2016, December 06). What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development. First Language, 1–21. doi: 10.1177/0142723716679956

Garmann, N. G., Hansen, P., Simonsen, H. G., & Kristoffersen, K. E. (2019). The Phonology of Children's Early Words: Trends, Individual Variation, and Parents' Accommodation in Child-Directed Speech. Frontiers in Communication, 4. doi:10.3389/fcomm.2019.00010 (doi:10.3389/fcomm.2019.00010)

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. Journal of Child Language, 21(01), 85-123.

Garmann, N. G. (2016). Garmann-Norwegian. [CHILDES/PhonBank corpus]. Retrieved from http://childes.talkbank.org/data/PhonBank/Norwegian-Garmann.zip (http://childes.talkbank.org/data/PhonBank/Norwegian-Garmann.zip).

Garmann, N. G., Hansen, P., Simonsen, H. G., & Kristoffersen, K. E. (in press). Phonological characteristics of children's first words. In F. Chenu, S. Kern, & F. Gayraud (Eds.), Proceedings from the 3rd ELA conference. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Guevara, E. (2010). NoWaC: A large web-based corpus for Norwegian. Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, 1-7. Retrieved from http://www.aclweb.org/anthology/W10-1501 (http://www.aclweb.org/anthology/W10-1501).

Kristoffersen, K. E., & Simonsen, H. G. (2012). Tidlig språkutvikling hos norske barn. MacArthur-Bates foreldrerapportering for kommunikativ utvikling [Early language development among Norwegian children. MacArthur-Bates Communicative Development Inventories]. Oslo: Novus Forlag.

Lind, M., Simonsen, H. G., Hansen, P., Holm, E., & Mevik, B.-H. (2015). Norwegian Words: A lexical database for clinicians and researchers. Clinical Linguistics & Phonetics, 29(4), 276-290.

Simonsen, H. G. (1990). Barns fonologi: System og variasjon hos tre norske og ett samoisk barn [Children's phonology: System and variation in three Norwegian children and one Samoan]. (PhD thesis), University of Oslo.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23.

Garmann, N. G. (2016). Garmann-Norwegian. [CHILDES/PhonBank corpus]. Retrieved from http://childes.talkbank.org/data/PhonBank/Norwegian-Garmann.zip (http://childes.talkbank.org/data/PhonBank/Norwegian-Garmann.zip).

Garmann, N. G., Hansen, P., Simonsen, H. G., & Kristoffersen, K. E. (in press). Phonological characteristics of children's first words. In F. Chenu, S. Kern, & F. Gayraud (Eds.), Proceedings from the 3rd ELA conference. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Simonsen, H. G. (1990). Barns fonologi: System og variasjon hos tre norske og ett samoisk barn [Children's phonology: System and variation in three Norwegian children and one Samoan]. (PhD thesis), University of Oslo.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23.

Data

Hansen, P. (2016, December 06). Replication data for: What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development. Retrieved June 12, 2020, from https://dataverse.no/dataset.xhtml?persistentId=doi%3A10.18710%2FJEWIVW (https://dataverse.no/dataset.xhtml?persistentId=doi%3A10.18710%2FJEWIVW)

(2003). CHILDES. Retrieved June 12, 2020, from https://childes.talkbank.org/ (https://childes.talkbank.org/)