Detection of a Specific Language Impairment (SLI) in children's speech

Project Description

Specific Language Impairment is a condition that affects roughly 7% of 5-year old children. It is characterized by a lack of language ability in comparison to your peers but with no obvious mental or physical disability. Children with SLI are more likely to speak in short sentences and often have difficulty with the morphemes -ed, -s, be, and do. The main goal of this project is to research how SLI diagnosis correlates with certain speech characteristics, gender, and age of a child.

```
FALSE -- Attaching packages ---------------------------------------------------------------------------------
FALSE v ggplot2 3.3.1     v purrr   0.3.4
FALSE v tibble  3.0.1     v dplyr   1.0.0
FALSE v tidyr   1.1.0     v stringr 1.4.0
FALSE v readr   1.3.1     v forcats 0.5.0
FALSE -- Conflicts ------------------------------------------------------------------------------------------
FALSE x dplyr::filter() masks stats::filter()
FALSE x dplyr::lag()    masks stats::lag()
FALSE Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
FALSE (as 'lib' is unspecified)
FALSE Registered S3 method overwritten by 'GGally':
FALSE   method from
FALSE   +.gg   ggplot2
FALSE Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
FALSE (as 'lib' is unspecified)
```

```
sli_data <- read.csv('https://raw.githubusercontent.com/lnpetrova/Statistics/master/sli5.csv')
```

Variables

There are linguistic and non-linguistic variables. Linguistic ones:

- childTNW | Total Number of Words;
- childTNS | Total Number of Sentences;
- ndos | Number of Do's;
- repetition | Number of Repetitions;
- fillers | Number of Fillers;
- r2iverbs| Ratio of raw to inflected verbs | Children with SLI often have difficulty with the morphemes -ed, -s, be, and do. This results in the use of raw verbs

Non-linguistic ones:

- Y | Target

- group | Information about SLI for graphs
- age_years
- sex

Description of corpora

*ENNI

Each child was presented with two wordless picture stories with one more complicated than the other. The examiner held the book and turned the page after the child appeared to be finished telling the story for a particular picture.

*Gillam

The Gillam dataset is based on a tool for narrative assessment known as "The Test of Narrative Language (TNL). The TNL consists of four storytelling tasks, the first is a recall of a script based story, the rest being wordless picture books. The TNL appears to be an intermediary in difficulty compared to ENNI.

Preprocessing

```
str(sli_data)
```

```
## 'data.frame':    1044 obs. of  12 variables:
##  $ X          : int  118 119 120 121 122 123 124 125 126 127 ...
##  $ Y          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sex        : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 2 2 1 ...
##  $ age_years  : num  4.92 5 4.67 4.17 4.75 ...
##  $ corpus     : Factor w/ 2 levels "ENNI","Gillam": 1 1 1 1 1 1 1 1 1 1 ...
##  $ group      : Factor w/ 2 levels "SLI","TD": 1 1 1 1 1 1 1 1 1 1 ...
##  $ child_TNW  : int  212 468 212 292 611 321 658 261 227 436 ...
##  $ child_TNS  : int  66 67 67 90 100 64 97 62 58 70 ...
##  $ r_2_i_verbs: num  5.14 1.37 0.6 0.6 1 ...
##  $ n_dos      : int  0 1 3 1 1 0 1 0 0 0 ...
##  $ repetition : int  7 47 13 16 21 13 109 1 3 5 ...
##  $ fillers    : int  1 9 3 13 8 5 20 1 3 2 ...
```

The column X with some indexes is not needed here, so we can delete it.

```
sli_data$X <- NULL
```

Firstly, it's nessecery to round all the numeric values in the dataset and rename some columns for convinience.

```
sli_data <- mutate_if(sli_data, is.numeric, round, digits = 2)
```

Most of the variables were renamed for convinience.

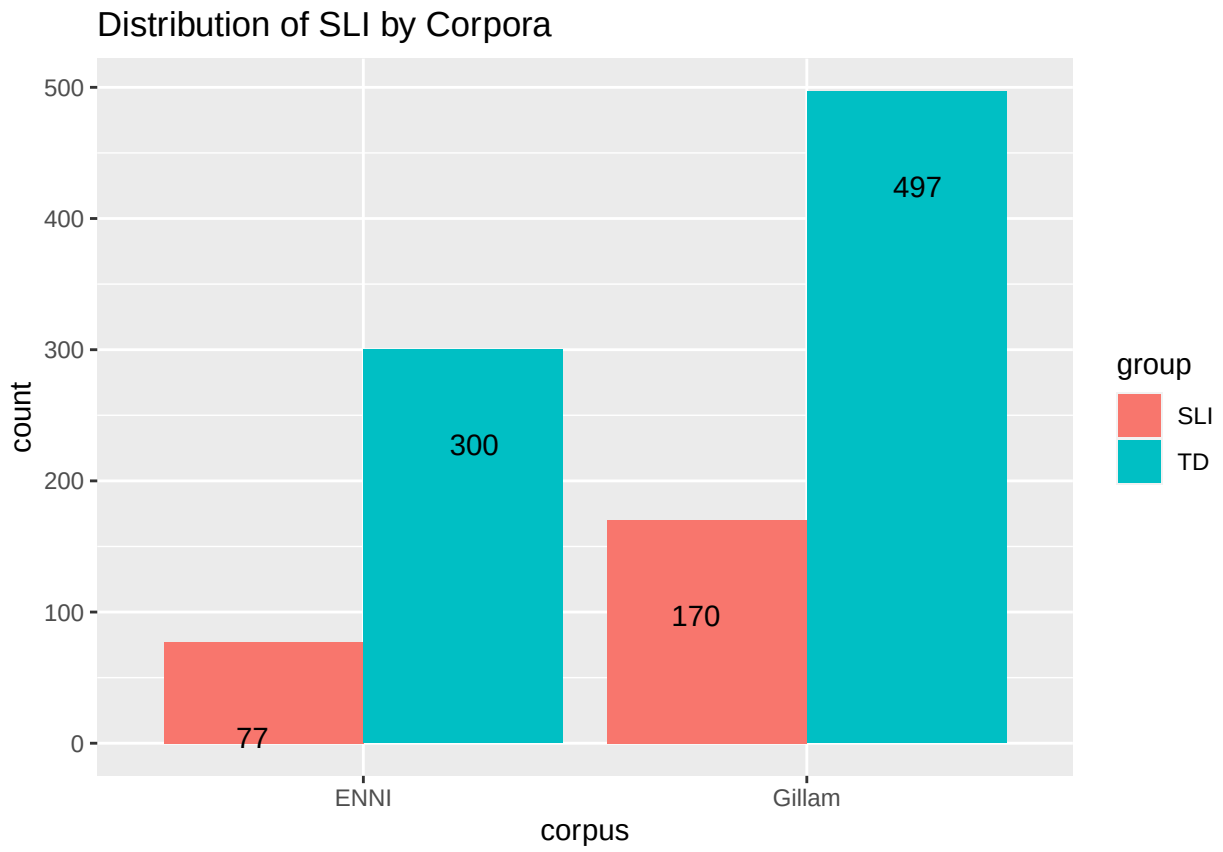
```
names(sli_data)[1] <- 'target'
names(sli_data)[3] <- 'age'
names(sli_data)[6] <- 'n_words'
names(sli_data)[7] <- 'n_sents'
names(sli_data)[8] <- 'verb_ratio'
```

```
str(sli_data)
```

```
## 'data.frame':    1044 obs. of  11 variables:
##  $ target    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 2 2 1 ...
##  $ age       : num  4.92 5 4.67 4.17 4.75 4.67 4.75 4.67 4.5 4.75 ...
##  $ corpus    : Factor w/ 2 levels "ENNI","Gillam": 1 1 1 1 1 1 1 1 1 1 ...
##  $ group     : Factor w/ 2 levels "SLI","TD": 1 1 1 1 1 1 1 1 1 1 ...
##  $ n_words   : num  212 468 212 292 611 321 658 261 227 436 ...
##  $ n_sents   : num  66 67 67 90 100 64 97 62 58 70 ...
##  $ verb_ratio: num  5.14 1.37 0.6 0.6 1 0.11 0.27 0.18 0.38 1.03 ...
##  $ n_dos     : num  0 1 3 1 1 0 1 0 0 0 ...
##  $ repetition: num  7 47 13 16 21 13 109 1 3 5 ...
##  $ fillers   : num  1 9 3 13 8 5 20 1 3 2 ...
```

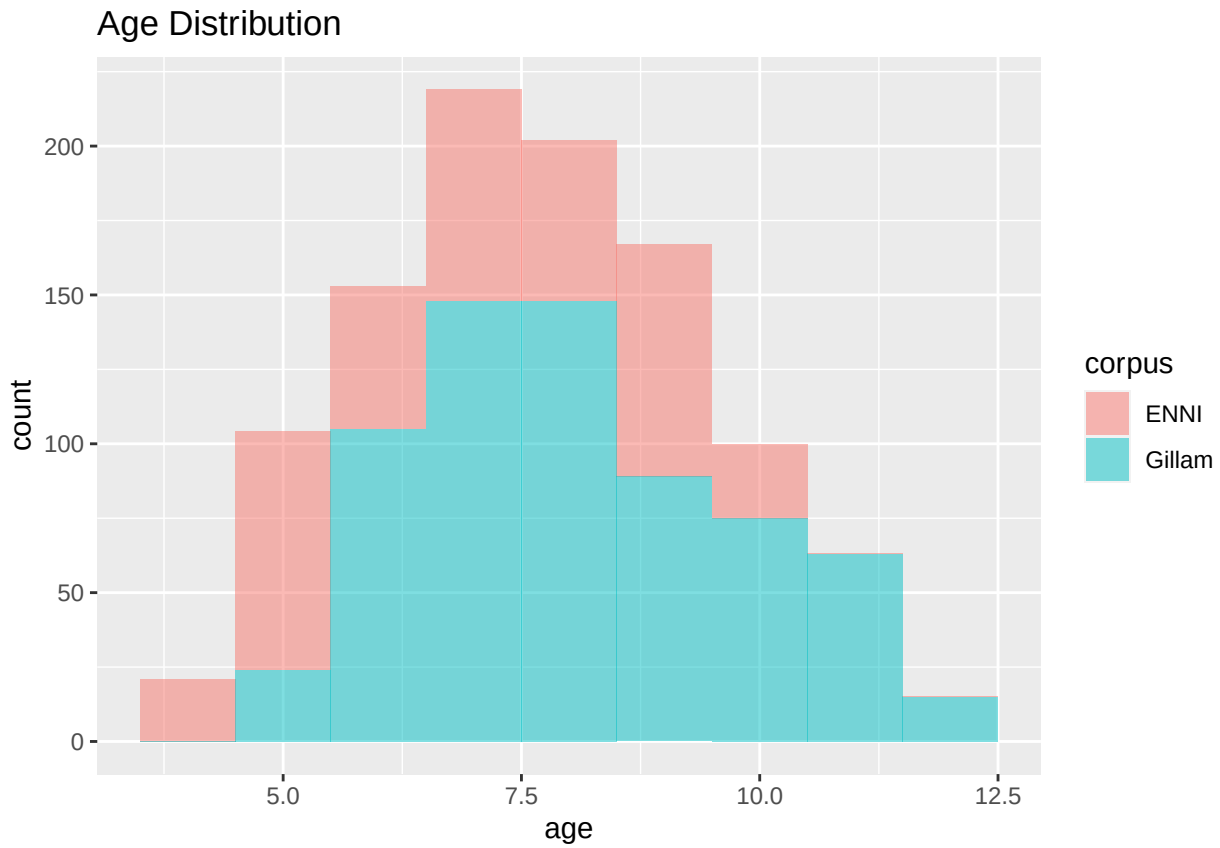
Visualisation of the data

```
sli_data %>%
    ggplot(aes(x=corpus, fill=group)) +
    geom_bar(position=position_dodge()) +
    labs(title=paste("Distribution of SLI by Corpora")) +
    geom_text(stat="count", aes(y = ..count.., label=..count..),
              position = position_dodge(1), vjust=5)
```

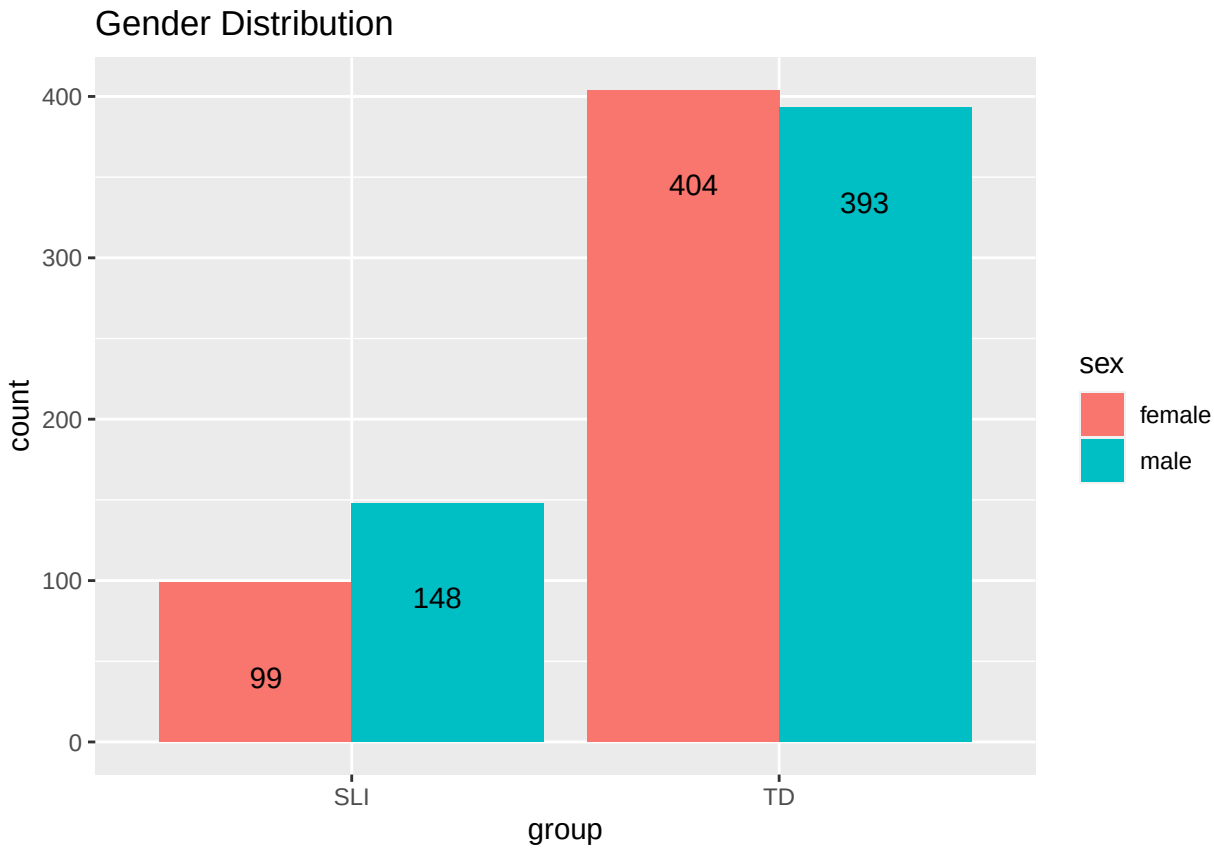



```
sli_data %>%
  ggplot(aes(x=age, fill=corpus)) +
  geom_histogram(binwidth=1, alpha=.5)+
```

```
  labs(title=paste("Age Distribution"))
```

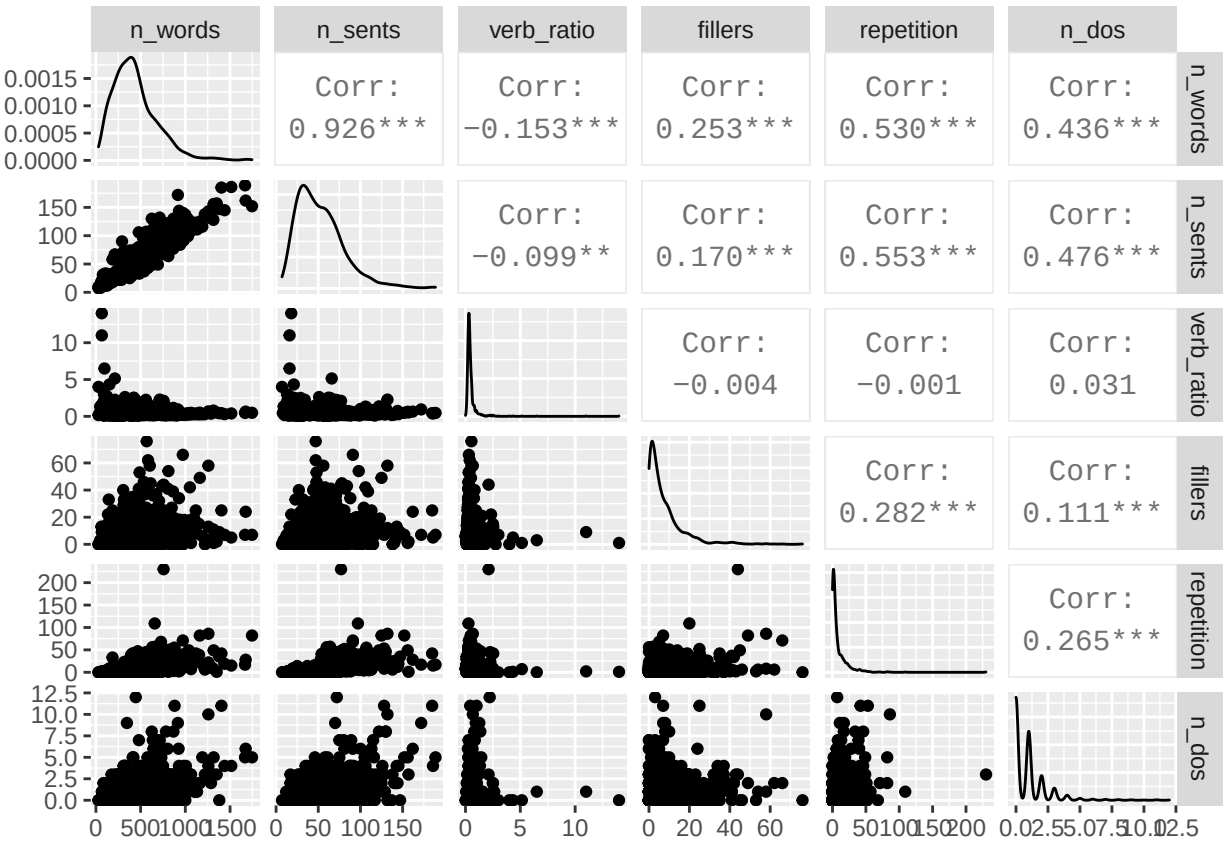



```
sli_data %>%
    ggplot(aes(x=group, fill=sex)) +
    geom_bar(position=position_dodge()) +
    labs(title=paste("Gender Distribution")) +
    geom_text(stat="count", aes(y = ..count.., label=..count..),
              position = position_dodge(0.8), vjust=5)
```



Considering the chosen variables, we can claim that some of them correlate. So let's look at the correlation of all linguistic variables before conducting tests.

```
features <- sli_data %>% select(n_words, n_sents, verb_ratio, fillers, repetition, n_dos)
ggpairs(features)
```



As we can see, the correlation between n_words and n_sents tends to 1, so association between these two variables is really strong. This situation is not desirable for the next steps, especially, for training models. As the correlation is so high, we can change it to a new variable which describes average length of sentences in the speech of a kid.

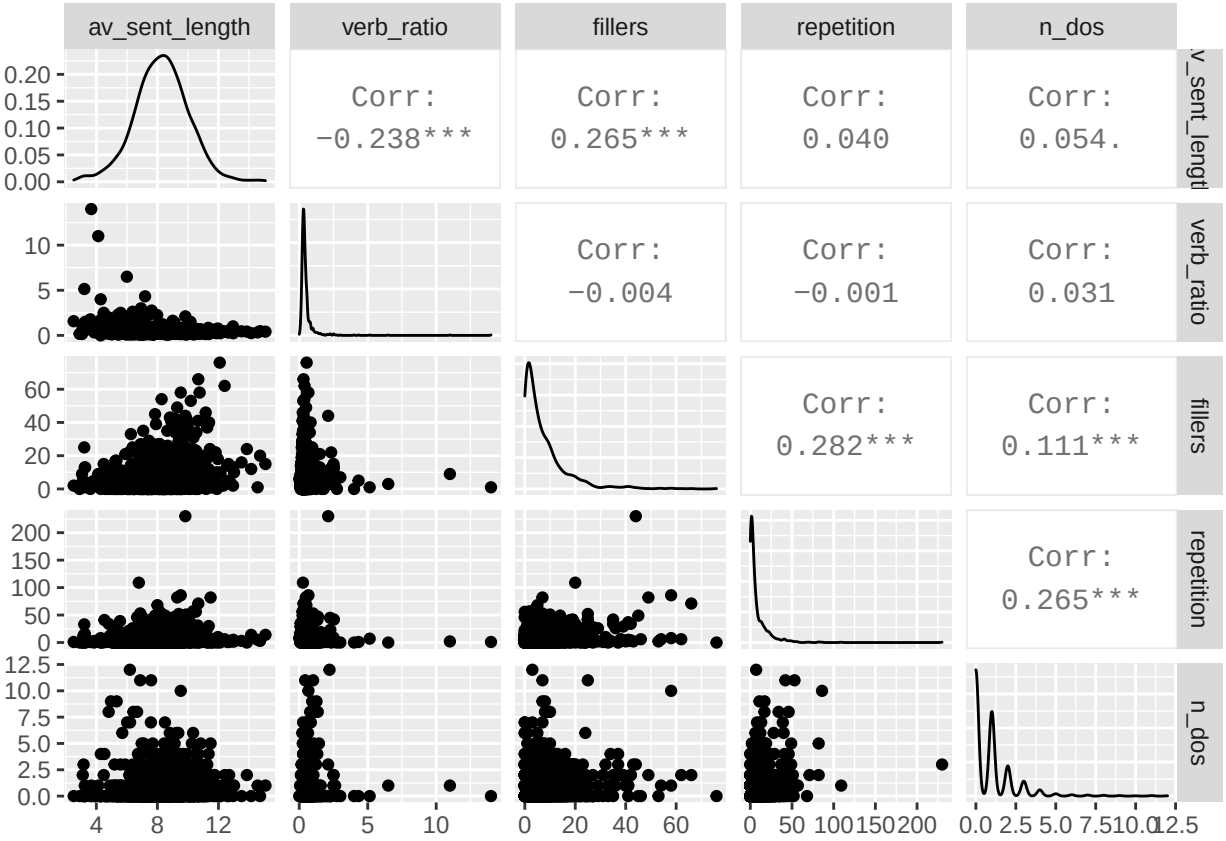
```
cor.test(sli_data$n_words, sli_data$n_sents)
```

```
##
##  Pearson's product-moment correlation
##
## data:  sli_data$n_words and sli_data$n_sents
## t = 79.15, df = 1042, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9167872 0.9341467
## sample estimates:
##       cor
## 0.9259546
```

```
sli_data <- mutate(sli_data, av_sent_length = n_words/n_sents)
```

Let's look at the correlation between variables now.

```
features <- sli_data %>% select(av_sent_length, verb_ratio, fillers, repetition, n_dos)
ggpairs(features)
```



Now we can see that correlations between all the variables are not so high. Let's conduct some tests to check statistical significance.

Testing Hypothesis

T-tests for numerical variables

1.Verb ratio

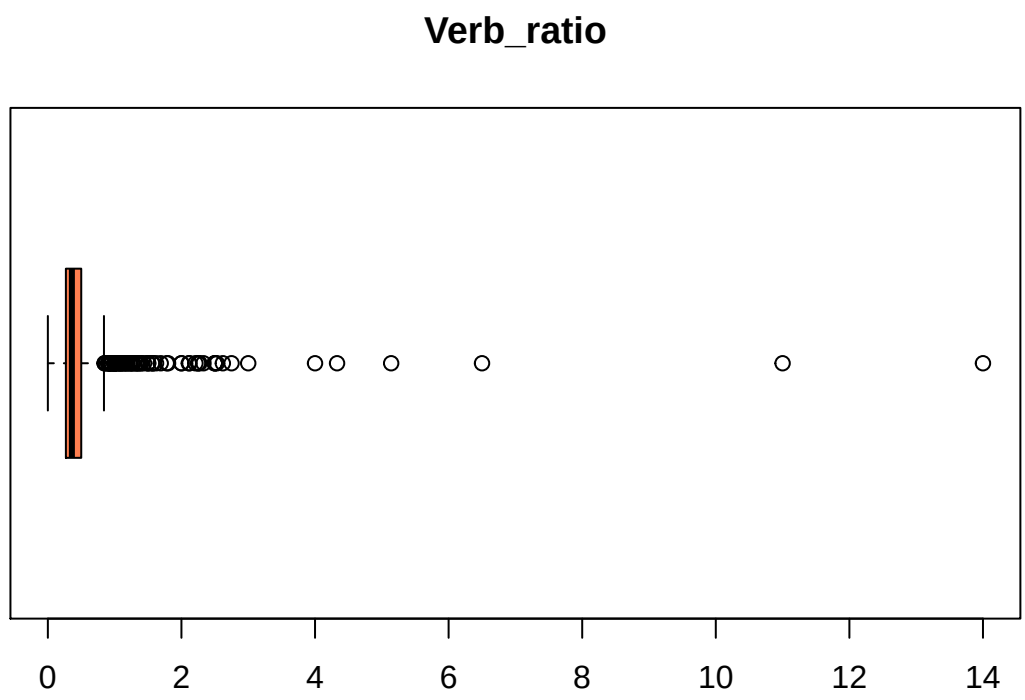
H0: There is no relationship between verb ration in children's speech and SLI.

H1:There is a statistically significant relationship between verb ratio in children's speech and SLI.

```
t.test(sli_data$verb_ratio ~ sli_data$target)
```

```
##
##  Welch Two Sample t-test
##
## data:  sli_data$verb_ratio by sli_data$target
## t = -5.2472, df = 251.17, p-value = 3.285e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6046240 -0.2746155
## sample estimates:
## mean in group 0 mean in group 1
##       0.3896110       0.8292308
```

```r
boxplot(sli_data$verb_ratio,
            ylab = " ",
            xlab = " ",
            main = "Verb_ratio",
            col = "coral", horizontal = TRUE,
            data = sli_data)
```



p-value $< 0{,}05$, so we can reject H0. There is a statistically significant relationship between verb ratio in children's speech and SLI.

2.Fillers

H0: There is no relationship between number of fillers in children's speech and SLI.

H1:There is a statistically significant relationship between number of fillers in children's speech and SLI.

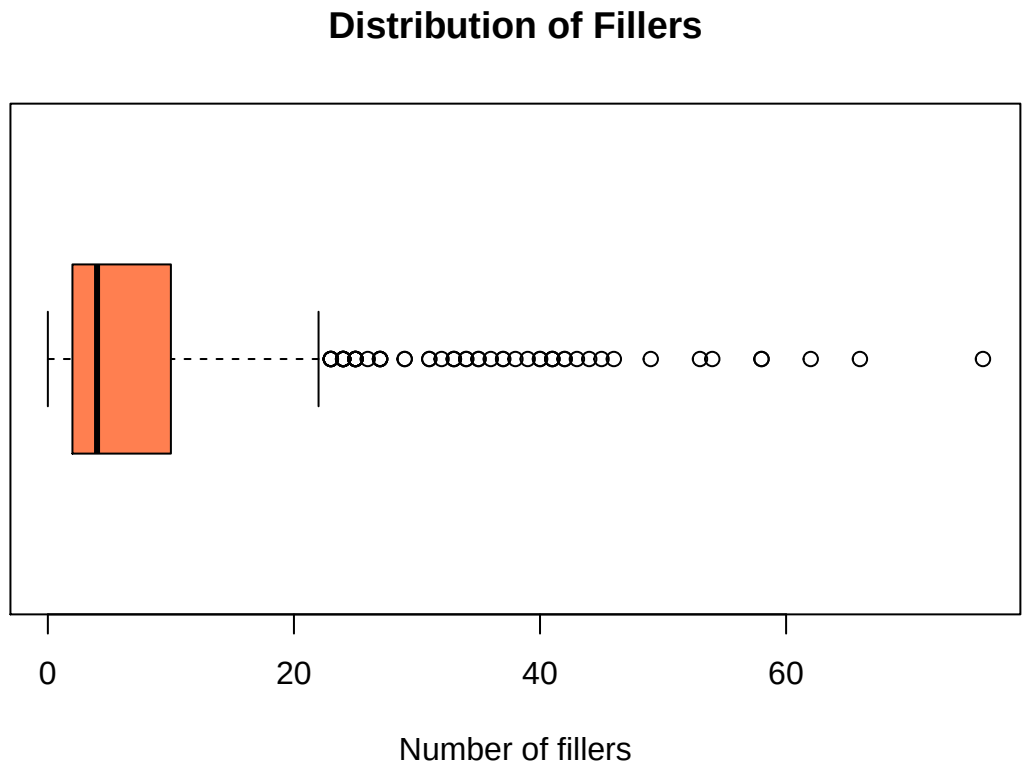
```r
t.test(sli_data$fillers ~ sli_data$target)
```

```
##
##  Welch Two Sample t-test
##
## data:  sli_data$fillers by sli_data$target
## t = -0.65354, df = 375.02, p-value = 0.5138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.8512700  0.9276445
## sample estimates:
## mean in group 0 mean in group 1
##        7.396487        7.858300
```

p-value $> 0{,}05$, so we can't reject H0. It means that there is no relationship between number of fillers in

children's speech and SLI. According to the boxplot, we see that there are many outliers, and T-test can be wrong in this case.

```
boxplot(sli_data$fillers,
            ylab = " ",
            xlab = "Number of fillers",
            main = "Distribution of Fillers",
            col = "coral", horizontal = TRUE,
            data = sli_data)
```



3.Repetitions

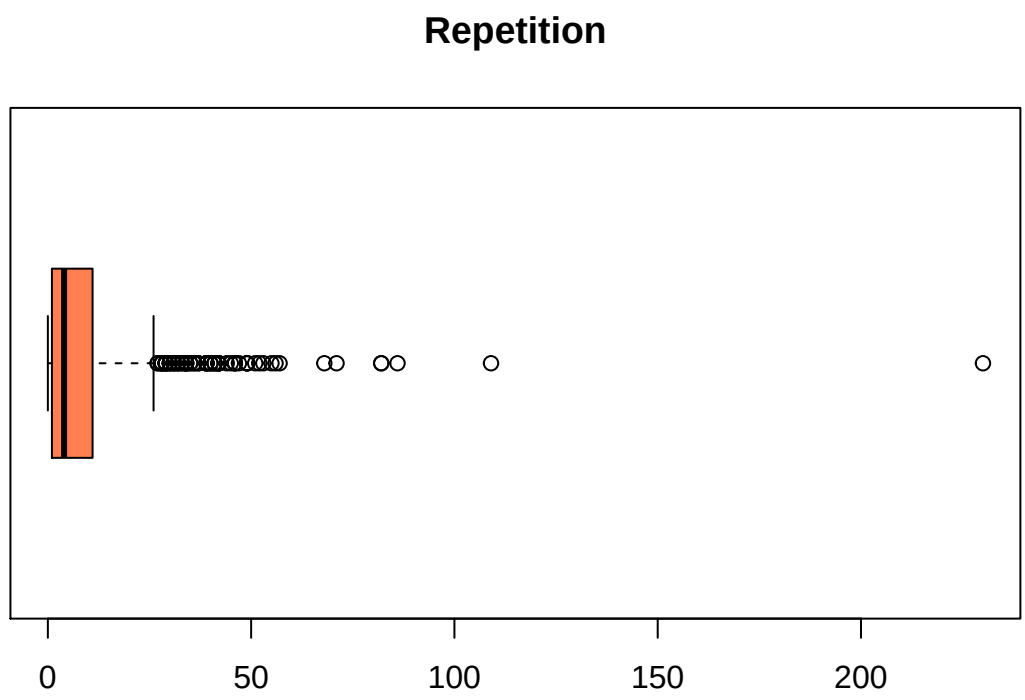
H0: There is no relationship between number of repetitions in children's speech and SLI.

H1:There is a statistically significant relationship between number of repetitions in children's speech and SLI.

```
t.test(sli_data$repetition ~ sli_data$target)
```

```
##
##  Welch Two Sample t-test
##
## data:  sli_data$repetition by sli_data$target
## t = -1.177, df = 289.18, p-value = 0.2402
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.233562  1.064959
## sample estimates:
## mean in group 0 mean in group 1
##        8.035132        9.619433
```

```r
boxplot(sli_data$repetition,
            ylab = " ",
            xlab = " ",
            main = "Repetition",
            col = "coral", horizontal = TRUE,
            data = sli_data)
```



p-value > 0,05, so we can't reject H0. It means that there is no relationship between number of repetitions in children's speech and SLI. Boxplot shows us that that there are not so many outliers, but still they can influence T-test result.

4.Number of dos

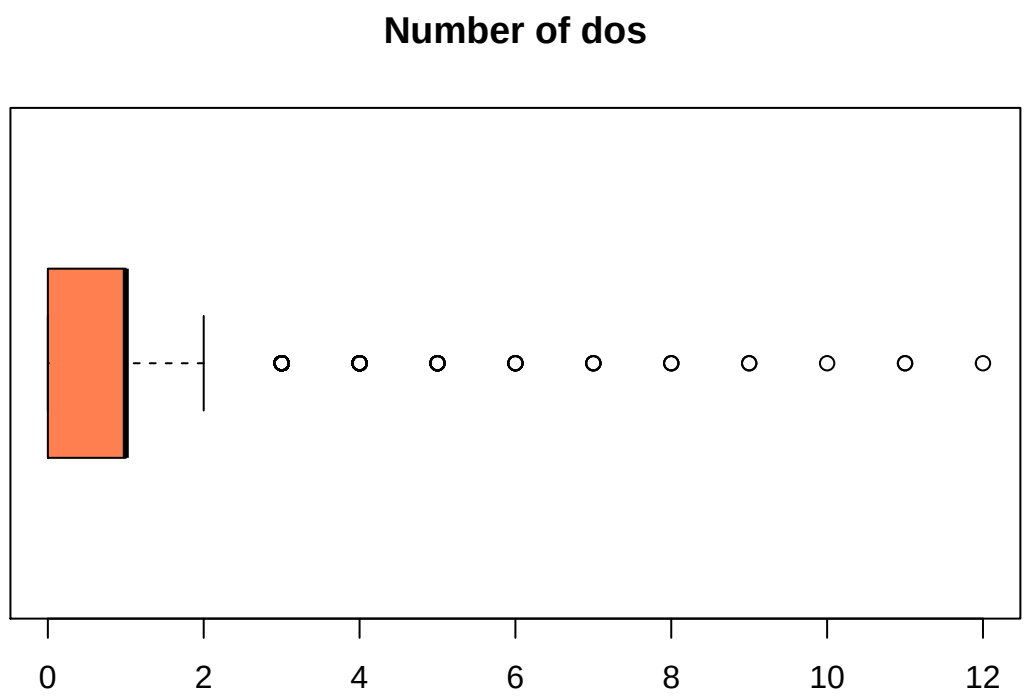
H0: There is no relationship between number of dos in children's speech and SLI.

H1:There is a statistically significant relationship between number of dos in children's speech and SLI.

```r
t.test(sli_data$n_dos ~ sli_data$target)
```

```
##
##  Welch Two Sample t-test
##
## data:  sli_data$n_dos by sli_data$target
## t = -1.2125, df = 315.19, p-value = 0.2262
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4095027  0.0972284
## sample estimates:
## mean in group 0 mean in group 1
##       0.9410289       1.0971660
```

```r
boxplot(sli_data$n_dos,
            ylab = " ",
            xlab = " ",
            main = "Number of dos",
            col = "coral", horizontal = TRUE,
            data = sli_data)
```



p-value > 0,05, so we can't reject H0. It means that there is no relationship between number of repetitions in children's speech and SLI. As the plot shows, values are not distributed normally. T-test can be wrong here. So, we'll check it with the help of models.

5.Average Sentence Length

H0: There is no relationship between an average sentence length in children's speech and SLI.

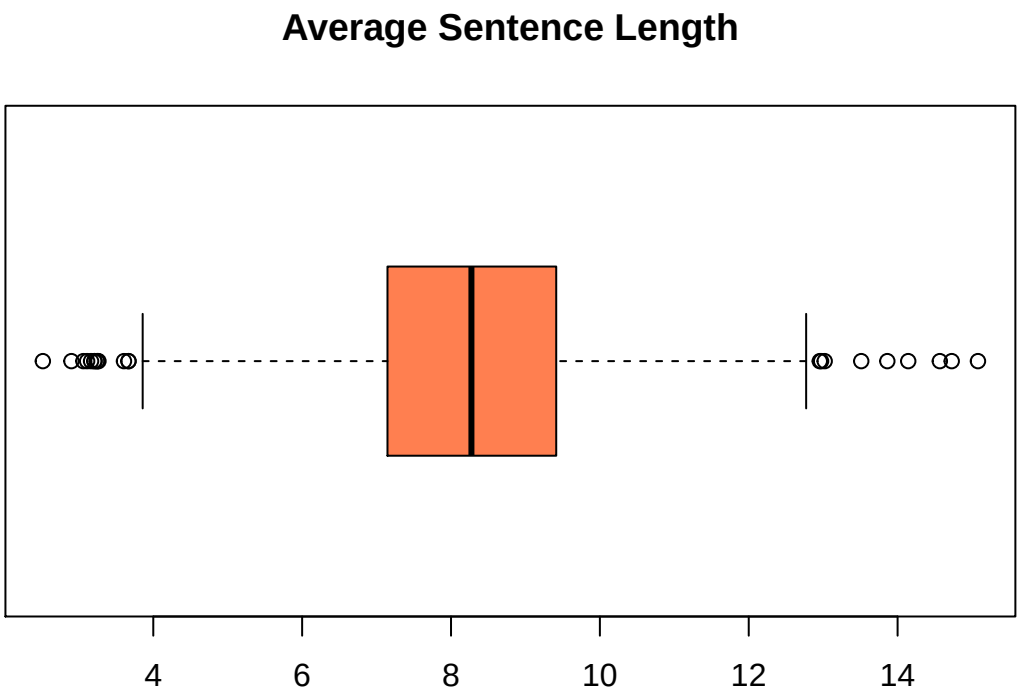
H1:There is a statistically significant relationship between an average sentence length in children's speech and SLI.

```r
t.test(sli_data$av_sent_length ~ sli_data$target)
```

```
##
##  Welch Two Sample t-test
##
## data:  sli_data$av_sent_length by sli_data$target
## t = 9.5649, df = 376.22, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9906458 1.5033432
## sample estimates:
## mean in group 0 mean in group 1
```

```
##        8.563735        7.316741
```

```
boxplot(sli_data$av_sent_length,
            ylab = " ",
            xlab = " ",
            main = "Average Sentence Length",
            col = "coral", horizontal = TRUE,
            data = sli_data)
```



p-value $< 0{,}05$, so we can reject H0. There is a statistically significant relationship between an average sentence length in children's speech and SLI. According to the boxplot, we see that there are not so many outliers, and it's appropriate to use T-test in this case.

Tests for non-linguistic variables

6.Age

H0: There is no relationship between an age of a kid and SLI.

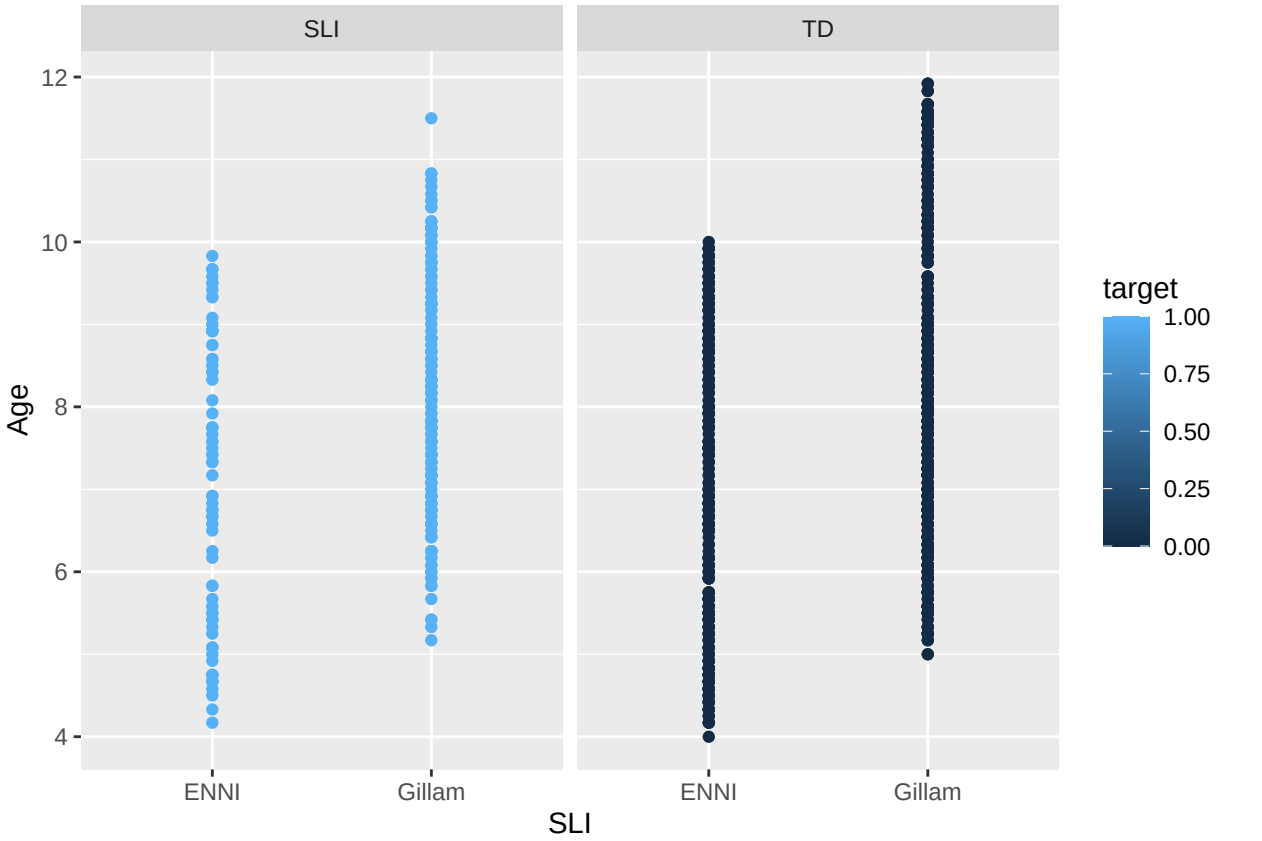
H1:There is a statistically significant relationship between an age of a kid in children's speech and SLI.

```
t.test(sli_data$age ~ sli_data$target)
```

```
##
##  Welch Two Sample t-test
##
## data:  sli_data$age by sli_data$target
## t = -0.017972, df = 477.33, p-value = 0.9857
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2365755  0.2322873
```

```
## sample estimates:
## mean in group 0 mean in group 1
##        7.742836        7.744980
```

```
ggplot(sli_data, aes(x = corpus, y = age, color = target))+
geom_point()+
facet_wrap(~group)+
labs(x = "SLI", y = "Age")
```



p-value $>$ 0,05, so we can't reject H0. It means that there is no relationship between an age of a kid and SLI.

7.Sex

H0: There is no relationship between a sex of a kid and SLI.

H1:There is a statistically significant relationship between a sex of a kid in children's speech and SLI.

```
tab1 <- table(sli_data$sex, sli_data$group)
tab1
```

```
##
##          SLI  TD
##   female  99 404
##   male   148 393
```

Values are pretty high, so it's approproate to use chi-square distribution for rejecting or accepting hypotheses.

```
chisq.test(sli_data$sex, sli_data$group)
```

```
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sli_data$sex and sli_data$group
## X-squared = 8.081, df = 1, p-value = 0.004473
```

p-value < 0.05, we can reject H0

```
cramersV(tab1)
```

```
## [1] 0.0879795
```

However, Cramer's test shows that the diagnosis doesn't depend on a sex of a speaker so much.

First Conclusions

The first tests showed that not all of the variables are statistically important for SLI prediction. Using T-test, we figured out that the variable 'repetition' is not significant in the experiment. It's still not clear whether number of fillers and number of dos are relevant for the experiment because of abnormal distribution of the values that are not good in terms of evaluation with the use of T-test. As for non-linguistic features, the age variable is not statistically important, and a sex of a speaker is into doubt.

Models

Model1 with all variables

Let's try to train logistic regression on all variables. Moreover, we'll be able to test our previous test results.

```
model1 <- glm(target~av_sent_length + fillers + repetition + verb_ratio + n_dos + age +
              + sex, data = sli_data, family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = target ~ av_sent_length + fillers + repetition +
##     verb_ratio + n_dos + age + +sex, family = "binomial", data = sli_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6036  -0.6579  -0.4745  -0.2519   2.9370
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.790484   0.525883  -1.503   0.1328
## av_sent_length -0.578533   0.064975  -8.904  < 2e-16 ***
## fillers         0.018860   0.009754   1.934   0.0532 .
## repetition      0.017828   0.007587   2.350   0.0188 *
## verb_ratio      1.936357   0.292642   6.617 3.67e-11 ***
## n_dos           0.015806   0.058636   0.270   0.7875
## age             0.368285   0.056968   6.465 1.01e-10 ***
## sexmale         0.226691   0.167485   1.354   0.1759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1142.38  on 1043  degrees of freedom
## Residual deviance:  916.16  on 1036  degrees of freedom
## AIC: 932.16
##
## Number of Fisher Scoring iterations: 6
```

```
confint(model1)
```

```
## Waiting for profiling to be done...
```

```
##                        2.5 %      97.5 %
## (Intercept)    -1.8266866280  0.23723775
## av_sent_length -0.7088369675 -0.45387139
## fillers        -0.0008088054  0.03755342
## repetition      0.0032283610  0.03282302
## verb_ratio      1.3891333911  2.53420684
## n_dos          -0.0989099621  0.13151003
## age             0.2579278252  0.48149564
## sexmale        -0.1010118534  0.55624321
```

If a 95% confidence interval contains zero, this indicates that the corresponding effect is not significant. So, we can conclude that the corresponding effect of n_dos, fillers, sex are not significant for this model.

Model2 with 5 variables

As statistical tests and the previous model have shown that n_dos and sex are not significant for SLI detection, we'll not use them in model2.

```
##
## Call:
## glm(formula = target ~ av_sent_length + fillers + repetition +
##     verb_ratio + age, family = "binomial", data = sli_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6806  -0.6574  -0.4818  -0.2467   2.9007
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.663137   0.515142  -1.287  0.19799
## av_sent_length -0.583926   0.064732  -9.021  < 2e-16 ***
## fillers         0.019096   0.009788   1.951  0.05105 .
## repetition      0.018754   0.007245   2.589  0.00964 **
## verb_ratio      1.967885   0.289617   6.795 1.08e-11 ***
## age             0.372188   0.056741   6.559 5.40e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1142.38  on 1043  degrees of freedom
## Residual deviance:  918.06  on 1038  degrees of freedom
## AIC: 930.06
##
```

```
## Number of Fisher Scoring iterations: 6
```

```
confint(model2)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                         2.5 %      97.5 %
## (Intercept)    -1.6777457640  0.34409999
## av_sent_length -0.7137096328 -0.45969487
## fillers        -0.0006246525  0.03786409
## repetition      0.0046864144  0.03304817
## verb_ratio      1.4244556172  2.55796674
## age             0.2622940479  0.48497404
```

In this case, we can see that using of the variable 'fillers' is rather questionable that's why the variable in model3 is not fixed. We'll test whether it's necessery to use it for predicting SLI or not.

Mixed-effect Model3 with 5 variables

```
library(lme4)
```

```
FALSE Loading required package: Matrix
```

```
FALSE
FALSE Attaching package: 'Matrix'
```

```
FALSE The following objects are masked from 'package:tidyr':
FALSE
FALSE     expand, pack, unpack
```

```
model3 <- glmer(target~av_sent_length + repetition + verb_ratio + age + (1|fillers), data = sli_data, fa
summary(model3)
```

```
FALSE Generalized linear mixed model fit by maximum likelihood (Laplace
FALSE   Approximation) [glmerMod]
FALSE  Family: binomial  ( logit )
FALSE Formula: target ~ av_sent_length + repetition + verb_ratio + age + (1 |
FALSE     fillers)
FALSE    Data: sli_data
FALSE
FALSE      AIC      BIC   logLik deviance df.resid
FALSE    933.1    962.8   -460.6    921.1     1038
```

```
FALSE
FALSE Scaled residuals:
FALSE     Min      1Q  Median      3Q     Max
FALSE -6.1979 -0.4945 -0.3502 -0.1706  7.9104
FALSE
FALSE Random effects:
FALSE  Groups  Name        Variance Std.Dev.
FALSE  fillers (Intercept) 0.03777  0.1944
FALSE Number of obs: 1044, groups:  fillers, 52
FALSE
FALSE Fixed effects:
FALSE                 Estimate Std. Error z value Pr(>|z|)
FALSE (Intercept)    -0.803766   0.521931  -1.540  0.12356
FALSE av_sent_length -0.562876   0.063763  -8.828  < 2e-16 ***
FALSE repetition      0.022874   0.006962   3.286  0.00102 **
FALSE verb_ratio      2.017124   0.292622   6.893 5.45e-12 ***
FALSE age             0.380123   0.057080   6.659 2.75e-11 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE Correlation of Fixed Effects:
FALSE             (Intr) av_sn_ repttn vrb_rt
FALSE av_snt_lngt -0.482
FALSE repetition  -0.184 -0.212
FALSE verb_ratio  -0.442  0.040  0.064
FALSE age         -0.418 -0.542  0.289  0.155
```

```
confint(model3)
```

```
FALSE Computing profile confidence intervals ...
```

```
FALSE                       2.5 %      97.5 %
FALSE .sig01          0.00000000  0.55717340
FALSE (Intercept)    -1.83350163  0.22079295
FALSE av_sent_length -0.69098294 -0.44064874
FALSE repetition      0.00925941  0.03659266
FALSE verb_ratio      1.46792861  2.61343530
FALSE age             0.26969672  0.49382409
```

```
anova(model3, model2)
```

```
FALSE Data: sli_data
FALSE Models:
FALSE model3: target ~ av_sent_length + repetition + verb_ratio + age + (1 |
FALSE model3:     fillers)
FALSE model2: target ~ av_sent_length + fillers + repetition + verb_ratio +
FALSE model2:     age
FALSE        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
FALSE model3    6 933.13 962.83 -460.56   921.13
FALSE model2    6 930.06 959.77 -459.03   918.06 3.0652  0  < 2.2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we use the variable 'fillers' with random effect, AIC is getting higher that means the model works worse than Model2 without random effect.

Model4 without "fillers"

```
##
## Call:
## glm(formula = target ~ av_sent_length + repetition + verb_ratio +
##     age, family = "binomial", data = sli_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6987  -0.6618  -0.4879  -0.2495   2.8314
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.819626   0.507862  -1.614 0.106554
## av_sent_length -0.558700   0.062987  -8.870  < 2e-16 ***
## repetition      0.023250   0.006866   3.386 0.000708 ***
## verb_ratio      2.012877   0.290288   6.934 4.09e-12 ***
## age             0.377330   0.056560   6.671 2.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1142.38  on 1043  degrees of freedom
## Residual deviance:  921.67  on 1039  degrees of freedom
## AIC: 931.67
##
## Number of Fisher Scoring iterations: 6
```

```
confint(model4)
```

```
FALSE Waiting for profiling to be done...
```

```
FALSE                       2.5 %      97.5 %
FALSE (Intercept)    -1.820758026  0.17249795
FALSE av_sent_length -0.684961426 -0.43779857
FALSE repetition      0.009823603  0.03679602
FALSE verb_ratio      1.467913264  2.60408175
FALSE age             0.267802755  0.48977005
```

In this model, there are only significant variables.

Let's compare Model4 with Model3 and Model2.

```
anova(model3, model4)
```

```
FALSE Data: sli_data
FALSE Models:
FALSE model4: target ~ av_sent_length + repetition + verb_ratio + age
FALSE model3: target ~ av_sent_length + repetition + verb_ratio + age + (1 |
FALSE model3:     fillers)
FALSE        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
FALSE model4    5 931.67 956.42 -460.84   921.67
FALSE model3    6 933.13 962.83 -460.56   921.13 0.5436  1     0.4609
```

As AIC of Model4 is lower, it copes better than a mixed-effect model. However, models are not significantly different.

```r
anova(model4, model2, test = "Chisq")
```

```
FALSE Analysis of Deviance Table
FALSE
FALSE Model 1: target ~ av_sent_length + repetition + verb_ratio + age
FALSE Model 2: target ~ av_sent_length + fillers + repetition + verb_ratio +
FALSE     age
FALSE   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
FALSE 1      1039     921.67
FALSE 2      1038     918.06  1   3.6089  0.05747 .
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p-value > 0.05, models are not significantly different.

Models for checking the variable 'repetition' and 'fillers'

As T-test showed that the variable 'repetition' is not significant, it should be checked one more time. The variable 'fillers' was also doubtful.

```
##
## Call:
## glm(formula = target ~ repetition, family = "binomial", data = sli_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9323  -0.7305  -0.7183  -0.7134   1.7281
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.238655   0.085560 -14.477   <2e-16 ***
## repetition   0.007687   0.004972   1.546    0.122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1142.4  on 1043  degrees of freedom
## Residual deviance: 1140.0  on 1042  degrees of freedom
## AIC: 1144
##
## Number of Fisher Scoring iterations: 4
```

```r
confint(model51)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) -1.409066973 -1.07312475
## repetition  -0.002212922  0.01770499
```

```
##
## Call:
## glm(formula = target ~ fillers, family = "binomial", data = sli_data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.8598  -0.7354  -0.7251  -0.7217   1.7163
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.212334   0.094226 -12.866   <2e-16 ***
## fillers      0.005362   0.007725   0.694    0.488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1142.4  on 1043  degrees of freedom
## Residual deviance: 1141.9  on 1042  degrees of freedom
## AIC: 1145.9
##
## Number of Fisher Scoring iterations: 4
```

```
confint(model52)
```

```
FALSE Waiting for profiling to be done...
```

```
FALSE                    2.5 %      97.5 %
FALSE (Intercept) -1.39895941 -1.02933341
FALSE fillers     -0.01026942  0.02016378
```

It turns out that "repetition" and "fillers" are not significant.

Model6 without "repetition" and "fillers"

```
model6 <- glm(target~av_sent_length + verb_ratio + age, data = sli_data, family = "binomial")
summary(model6)
```

```
##
## Call:
## glm(formula = target ~ av_sent_length + verb_ratio + age, family = "binomial",
##     data = sli_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8036  -0.6608  -0.4964  -0.2688   2.7233
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.58077    0.49374  -1.176    0.239
## av_sent_length -0.51520    0.06021  -8.556  < 2e-16 ***
## verb_ratio      2.03339    0.28242   7.200 6.03e-13 ***
## age             0.32629    0.05418   6.022 1.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1142.38  on 1043  degrees of freedom
## Residual deviance:  933.49  on 1040  degrees of freedom
```

```
## AIC: 941.49
##
## Number of Fisher Scoring iterations: 6
```

```
confint(model6)
```

```
FALSE Waiting for profiling to be done...
```

```
FALSE                     2.5 %     97.5 %
FALSE (Intercept)    -1.5523935  0.3854183
FALSE av_sent_length -0.6357180 -0.3994310
FALSE verb_ratio      1.5012962  2.6078815
FALSE age             0.2212928  0.4339190
```

Conclusions

So, tests provided with the use of models approved and clarified the results that we've got before testing numeric and categorical variables.

There are 3 significant variables:

*average sentence length

*verb_ratio

*age

Desicion Tree

```
install.packages('party')
```

```
FALSE Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
FALSE (as 'lib' is unspecified)
```

```
library(party)
```

```
FALSE Loading required package: grid
```

```
FALSE Loading required package: mvtnorm
```

```
FALSE Loading required package: modeltools
```

```
FALSE Loading required package: stats4
```

```
FALSE
FALSE Attaching package: 'modeltools'
```

```
FALSE The following object is masked from 'package:lme4':
FALSE
FALSE     refit
```

```
FALSE Loading required package: strucchange
```

```
FALSE Loading required package: zoo
```

```
FALSE
FALSE Attaching package: 'zoo'
```

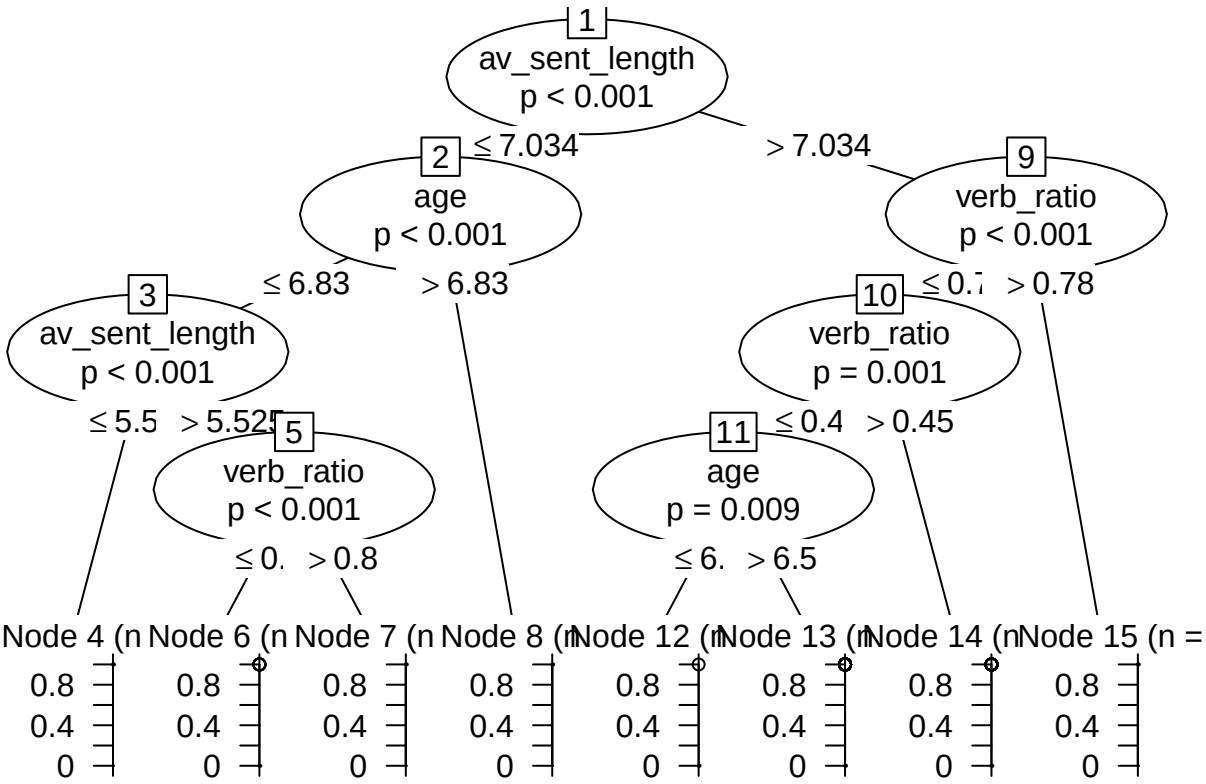
```
FALSE The following objects are masked from 'package:base':
FALSE
FALSE     as.Date, as.Date.numeric
```

```
FALSE Loading required package: sandwich
```

```
FALSE
FALSE Attaching package: 'strucchange'
```

```
FALSE The following object is masked from 'package:stringr':
FALSE
FALSE     boundary
```

```
tree = ctree(target~av_sent_length + fillers + repetition + verb_ratio + n_dos + age +
               + sex, data = sli_data)
plot(tree)
```



We've got the same significant variables with the use of desicion tree.

Random Forest

Let's see levels of feature importance with the use of Random forest Model.

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
fit <- randomForest(target~av_sent_length + fillers + repetition + verb_ratio + n_dos + age +
              + sex, data = sli_data)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
print(fit)
```

```
##
## Call:
##  randomForest(formula = target ~ av_sent_length + fillers + repetition +      verb_ratio + n_dos + a
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 0.1390415
##                     % Var explained: 23.02
```

```
importance(fit)
```

```
##                IncNodePurity
## av_sent_length     42.222544
## fillers            17.018655
## repetition         18.490789
## verb_ratio         44.655302
## n_dos               8.218997
## age                29.908864
## sex                 3.064426
```

As we can see, importance of linguistic variables is higher than non-linguistic ones.

Results

In this project, SLI diagnosis correlation with certain speech characteristics, gender, and age of a child was researched. In the process of the research, different statictical tests were provided: T-test, chi square test, Pearson correlation test. Moreover, logistic regression models were used for proving the results of statistical tests. It was found out that there were only 3 significant out of 8: age, average sentence length and verb ratio. Random Forest model shows that the most significant variables are linguistic ones (speech characteristics).