# The difference in microparameters of discourse of people with and without non-fluent aphasia

Packages

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.4
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
## Warning: package 'tibble' was built under R version 4.0.4
```

```
## Warning: package 'tidyr' was built under R version 4.0.4
```

```
## Warning: package 'readr' was built under R version 4.0.4
```

```
## Warning: package 'purrr' was built under R version 4.0.4
```

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
## Warning: package 'stringr' was built under R version 4.0.4
```

```
## Warning: package 'forcats' was built under R version 4.0.4
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
library(readr)
```

###Introduction

**Hypothesis**

Within the neurolinguistics domain such a problem as a violation of discourse in case of brain damage is being researched. The goal of the current study is to check if there's a difference in the microlinguistic parameters of the discourse of those who's speech is normal and those who have some deviations (non-fluent aphasia). The speech of people with non-fluent aphasia is supposed to be slower, so the hypothesis is that their pauses are in general longer. One more question is whether there are some correlations between variables, for example, different types of pauses should correlate with each other.

**Data collection**

Previous research

Data collection were obtained from the undergraduate project i have done. The initial data is the audio recordings from the neurolinguistic experiment. The subjects were asked to describe two drawings, if possible, try to compose a short story. Audio recordings of their responses served as material for further annotation and analysis of samples of oral discourse.

The records were first preprocessed in Audacity (the free audio editor). The emptiness, noise, the experimenter's speech at the beginning and end of the file were cut off. 500ms of silence were created at the beginning and at the end of the file. Noise and clicks were removed. Then, cleaned data were uploaded in ELAN (tool for annotating audio files). The duration of pauses, parts of speech were marked during this step. The last step was calculate the average values for each record in EXCEL.

**Data description**

The final dataset consists of the following variables:

rec_id``: Record ID. mean_abs : The average length of an absolute pause (silence) mean_fill : The average length of a filled pause mean_in : The average length of a pause inside the edu (elementary discourse unit
The type of speech on the record: normal or with aphasia

In the previous research the data were tested by t-test only. In the current project some exploratory analysis, additional testing and visualization are performed.

Let's load data first:

```
micro <- read.csv("https://raw.githubusercontent.com/dashapetrova/ma_da/main/project/micro_data_2.csv")
head(micro)
```

```
##              rec_id mean_abs mean_fill mean_in mean_out num_pos num_other
## 1      A-10-bike.eaf   977.22    667.30  772.15   989.17      21         5
## 2   A-10-robbery.eaf  1217.63    592.50  892.26  1385.75      20        14
## 3     A-14-bike1.eaf  3456.92    306.35 2291.76  3185.61      22         6
## 4 A-14-robbery-1.eaf  6055.78    651.10 2863.00  7792.69      18         0
## 5      A-47-bike.eaf   684.83    503.00  433.75   812.89      30        15
## 6   A-50-robbery.eaf  2005.94    367.50 1562.50  1891.36      11         2
##      type
## 1 aphasia
## 2 aphasia
## 3 aphasia
## 4 aphasia
## 5 aphasia
## 6 aphasia
```

## Descriptive statistics

Here's the data summary:

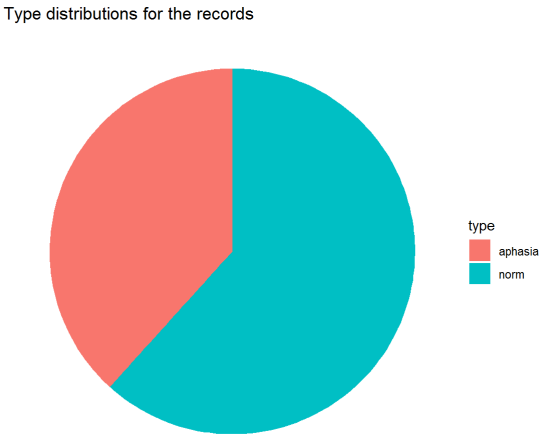
```
summary(micro)
```

```
##     rec_id            mean_abs       mean_fill        mean_in
##  Length:47          Min.   : 255.8   Min.   :   0.0   Min.   : 135.5
##  Class :character   1st Qu.: 659.4   1st Qu.:   0.0   1st Qu.: 448.5
##  Mode  :character   Median : 914.6   Median : 333.0   Median : 692.0
##                     Mean   :1139.4   Mean   : 322.0   Mean   : 797.7
##                     3rd Qu.:1194.3   3rd Qu.: 496.4   3rd Qu.: 956.3
##                     Max.   :6055.8   Max.   :1054.0   Max.   :2863.0
##     mean_out         num_pos        num_other         type
##  Min.   : 269.0   Min.   : 7.00   Min.   : 0.00   Length:47
##  1st Qu.: 673.7   1st Qu.:18.00   1st Qu.: 8.00   Class :character
##  Median :1058.6   Median :22.00   Median :14.00   Mode  :character
##  Mean   :1366.3   Mean   :28.85   Mean   :16.26
##  3rd Qu.:1386.0   3rd Qu.:34.50   3rd Qu.:22.50
##  Max.   :7792.7   Max.   :91.00   Max.   :59.00
```

According to the differences between min and max values, the range of the variables is big enough. But it may be caused by some extreme values, so that is why the distribution of the values will be plotted in the future.

Here's the pie chart that shows the distribution of record types.

```
ggplot(data = micro) +
  geom_bar(aes(x = "", fill = type)) +
  coord_polar(theta = "y") +
  theme_void() +
  labs(title = "Type distributions for the records")
```




Let's group the data by the type of the record and look at the mean values of variables.

```
mean_x <- function(x) mean(x)
micro %>%
  group_by(type) %>%
  summarise(across(colnames(micro)[2:7], mean_x))
```

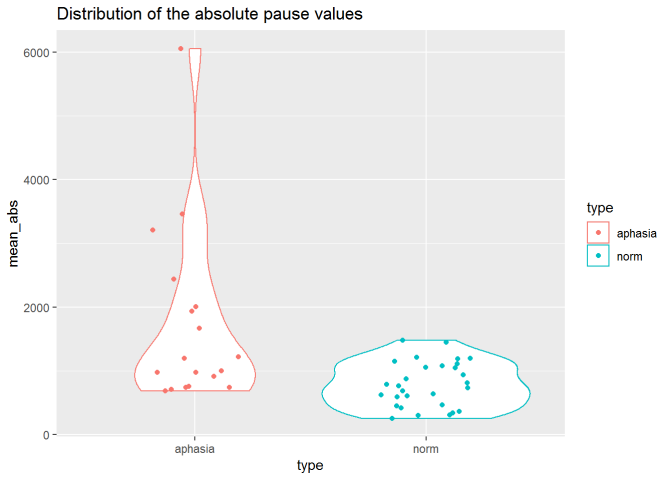
```
## # A tibble: 2 x 7
##   type    mean_abs mean_fill mean_in mean_out num_pos num_other
## * <chr>      <dbl>     <dbl>   <dbl>    <dbl>   <dbl>     <dbl>
## 1 aphasia   1704.      448.   1087.    2114.   22        11.4
## 2 norm       789.      244.    618.     902.   33.1      19.2
```

It is seen that the average values of some parameters (mean_abs, mean_in, mean_out, mean_fill) differ markedly between the two groups, while some other (num_pos, num_other) do not.

Next step is to choose all numerical variables and to look at its distribution in two groups.

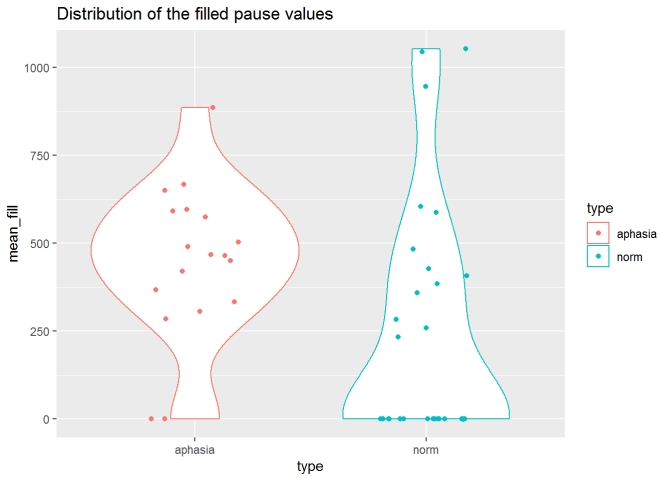
```
scores <- micro %>%
  select(mean_abs, mean_fill, mean_in, mean_out, num_pos, num_other)
```

```
micro %>%
  ggplot(aes(type, mean_abs, color=type)) +
  geom_violin() +
  geom_jitter(width=0.2) +
  labs(title = "Distribution of the absolute pause values")
```



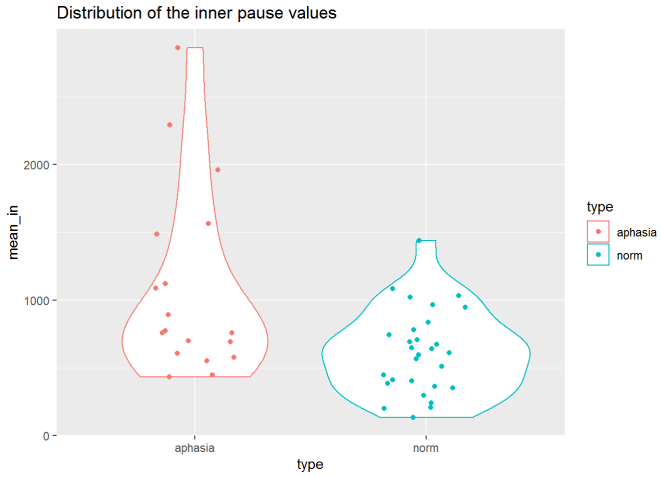

It is seen that the aphasia data is more diverse and have some extreme values, while values of the norm type are lower in general and they are less spread.

```
micro %>%
  ggplot(aes(type, mean_fill, color=type)) +
  geom_violin() +
  geom_jitter(width=0.2) +
  labs(title = "Distribution of the filled pause values")
```



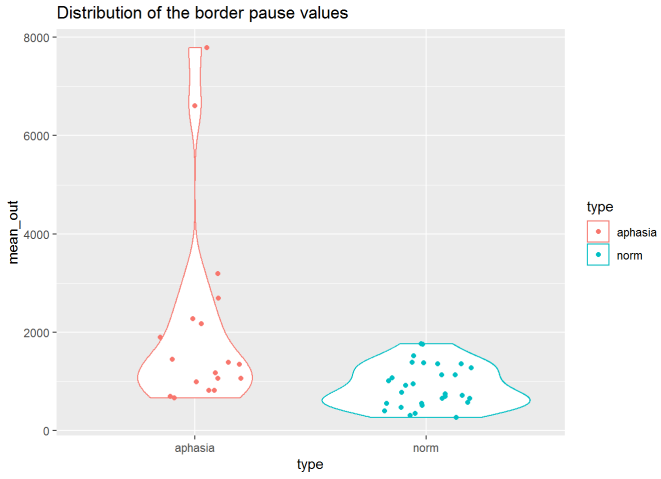

Here we can see that the range of the data is more similar, however the biggest parts of the groups are still in the different areas.

```
micro %>%
  ggplot(aes(type, mean_in, color=type)) +
  geom_violin() +
  geom_jitter(width=0.2) +
  labs(title = "Distribution of the inner pause values")
```




This plot looks similar to the first one: the aphasia values is more diverse and bigger, while values of the norm type are lower and less spread.

```
micro %>%
  ggplot(aes(type, mean_out, color=type)) +
  geom_violin() +
  geom_jitter(width=0.2) +
  labs(title = "Distribution of the border pause values")
```

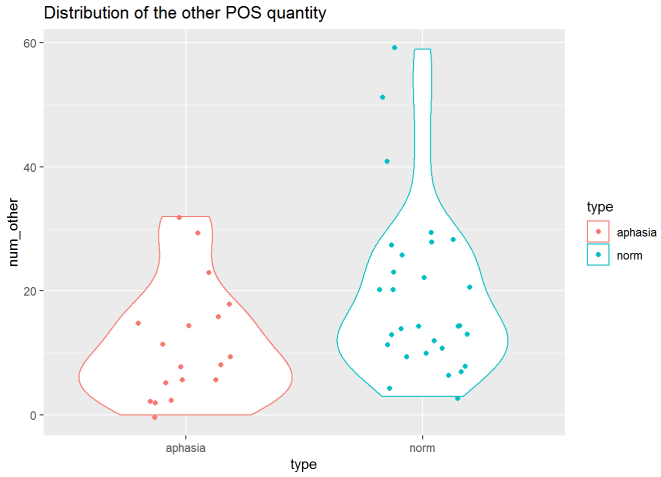



This plot has the same tendency, however the aphasia data is less diverse, there are just few extreme values.

```
micro %>%
  ggplot(aes(type, num_pos, color=type)) +
  geom_violin() +
  geom_jitter(width=0.2) +
  labs(title = "Distribution of the main POS quantity")
```


Here we can see the opposite trend, the values of the normal type is more varied and bigger.

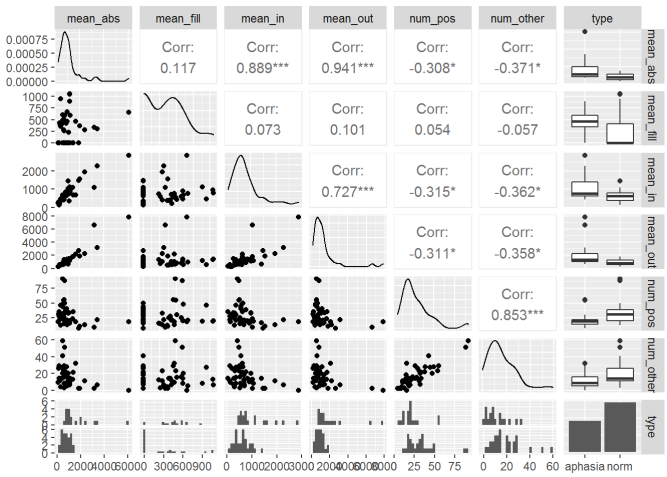
```
micro %>%
  ggplot(aes(type, num_other, color=type)) +
  geom_violin() +
  geom_jitter(width=0.2) +
  labs(title = "Distribution of the other POS quantity")
```




And the last plot is similar to the previous one, except there are no extreme values in the aphasia group.

Let's create a basic scatterplot matrix, a graph that includes several scatterplots, one for each pair of variables, and look if there is any correlation between variables.

```
more_data <- micro %>%
  select(colnames(micro)[2:8])
ggpairs(more_data)
```




*According to the plot and correlation coefficients, all variables are more or less correlated. Some of the greatest correlation are the following: + There is a positive correlation between the length of the absolute pause and the length of the pause inside/ouyside edus, so the bigger duration of the pause itself, the bigger duration of the absolute pause. It's interesting that there is no tendency like that if we consider the filled pauses. + There is also positive correlation between mean values of inner pauses and pauses on the border, so it is possible to assume that is if a person has long pauses in the edus, he will probably have long pauses between edus too. + Another positive correlation is between the number of different parts of speech: the more of notional POS, the more of functional POS

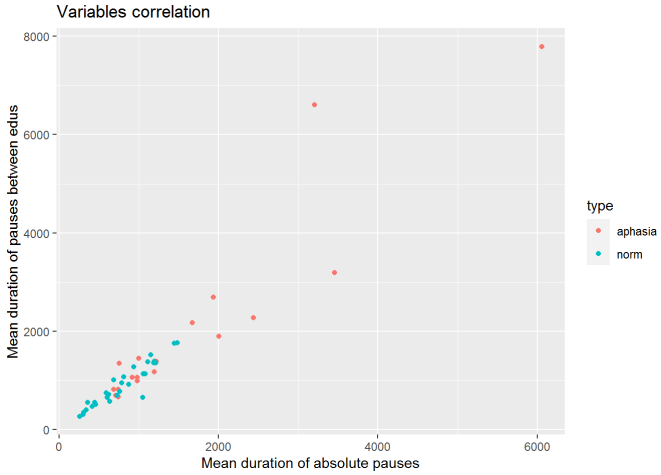
Let's choose these pairs of variables mentioned above and proceed to formal testing. We will check whether these values are indeed associated.

1. Duration of the absolute pause VS Duration of the pause between two edus

```
cor.test(micro$mean_abs, micro$mean_out)
```

```
##
##  Pearson's product-moment correlation
##
## data:  micro$mean_abs and micro$mean_out
## t = 18.71, df = 45, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8964957 0.9670744
## sample estimates:
##       cor
## 0.9413244
```

```
ggplot(data = micro, aes(x = mean_abs, y = mean_out, color = type)) +
  geom_point() +
  labs(x = "Mean duration of absolute pauses",
       y = "Mean duration of pauses between edus",
       title = "Variables correlation")
```

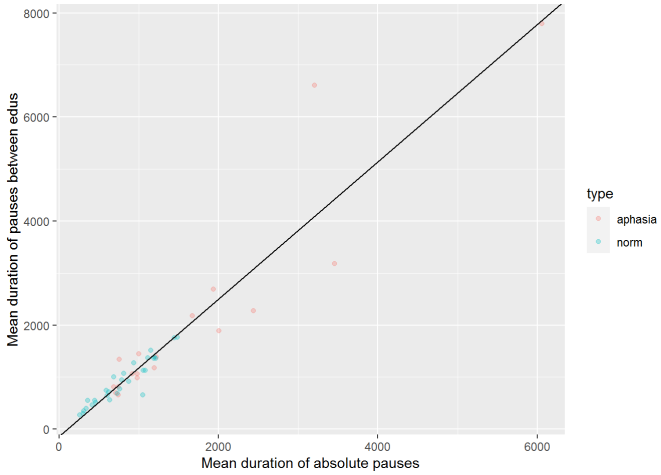



Let's now draw a regression line on top of the scatterplot.

```
model <- lm(mean_out ~ mean_abs, data = micro)
model
```

```
##
## Call:
## lm(formula = mean_out ~ mean_abs, data = micro)
##
## Coefficients:
## (Intercept)     mean_abs
##     -137.99         1.32
```

```
ggplot(data = micro, aes(x = mean_abs, y = mean_out, color=type))+
  geom_point(alpha = 0.3)+
  geom_abline(slope = model$coefficients[2], intercept = model$coefficients[1])+
  labs(x = "Mean duration of absolute pauses",
       y = "Mean duration of pauses between edus")
```

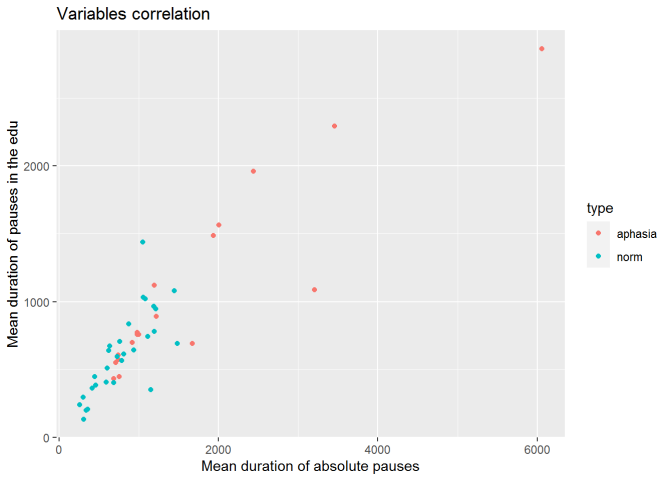



2. Duration of the absolute pause VS Duration of the pause inside the edu

```
cor.test(micro$mean_abs, micro$mean_in)
```

```
##
##  Pearson's product-moment correlation
##
## data:  micro$mean_abs and micro$mean_in
## t = 13.011, df = 45, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8078426 0.9368605
## sample estimates:
##       cor
## 0.8888168
```

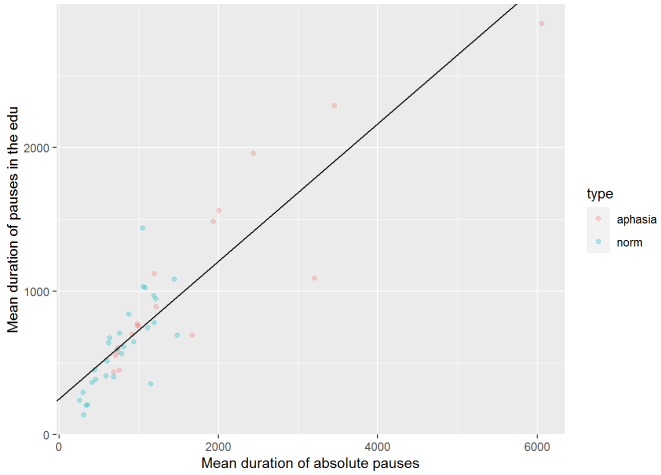
```
ggplot(data = micro, aes(x = mean_abs, y = mean_in, color = type)) +
  geom_point() +
  labs(x = "Mean duration of absolute pauses",
       y = "Mean duration of pauses in the edu",
       title = "Variables correlation")
```




```
model <- lm(mean_in ~ mean_abs, data = micro)
model
```

```
##
## Call:
## lm(formula = mean_in ~ mean_abs, data = micro)
##
## Coefficients:
## (Intercept)     mean_abs
##    251.4182       0.4794
```

```
ggplot(data = micro, aes(x = mean_abs, y = mean_in, color=type))+
  geom_point(alpha = 0.3)+
  geom_abline(slope = model$coefficients[2], intercept = model$coefficients[1])+
  labs(x = "Mean duration of absolute pauses",
       y = "Mean duration of pauses in the edu")
```

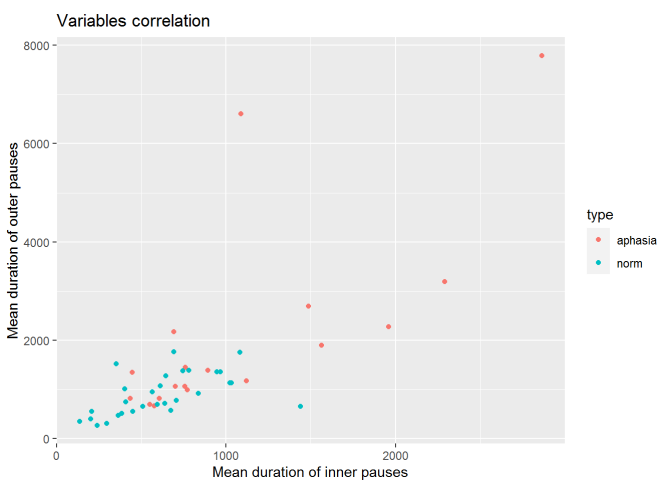



3. Duration of the pause inside the edu VS Duration of the pause between two edus

```
cor.test(micro$mean_in, micro$mean_out)
```

```
##
##  Pearson's product-moment correlation
##
## data:  micro$mean_in and micro$mean_out
## t = 7.0984, df = 45, p-value = 7.283e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5555952 0.8388827
## sample estimates:
##       cor
## 0.7268019
```

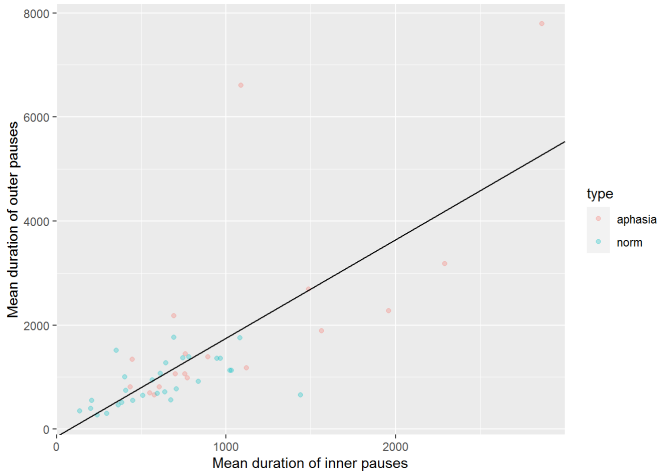
```
ggplot(data = micro, aes(x = mean_in, y = mean_out, color = type)) +
  geom_point() +
  labs(x = "Mean duration of inner pauses",
       y = "Mean duration of outer pauses",
       title = "Variables correlation")
```



```
model <- lm(mean_out ~ mean_in, data = micro)
model
```

```
##
## Call:
## lm(formula = mean_out ~ mean_in, data = micro)
##
## Coefficients:
## (Intercept)      mean_in
##     -141.17         1.89
```

```
ggplot(data = micro, aes(x = mean_in, y = mean_out, color=type))+
  geom_point(alpha = 0.3)+
  geom_abline(slope = model$coefficients[2], intercept = model$coefficients[1])+
  labs(x = "Mean duration of inner pauses",
       y = "Mean duration of outer pauses")
```

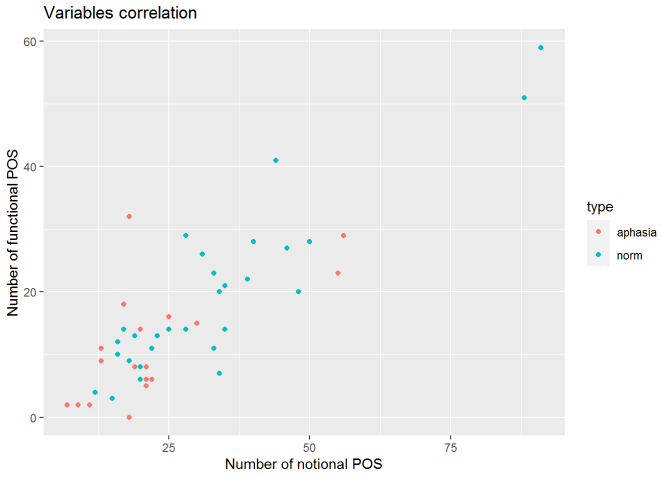



4. Number of notional POS VS number of functional POS

```
cor.test(micro$num_pos, micro$num_other)
```

```
##
##  Pearson's product-moment correlation
##
## data:  micro$num_pos and micro$num_other
## t = 10.953, df = 45, p-value = 2.779e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7490482 0.9157010
## sample estimates:
##       cor
## 0.8527804
```

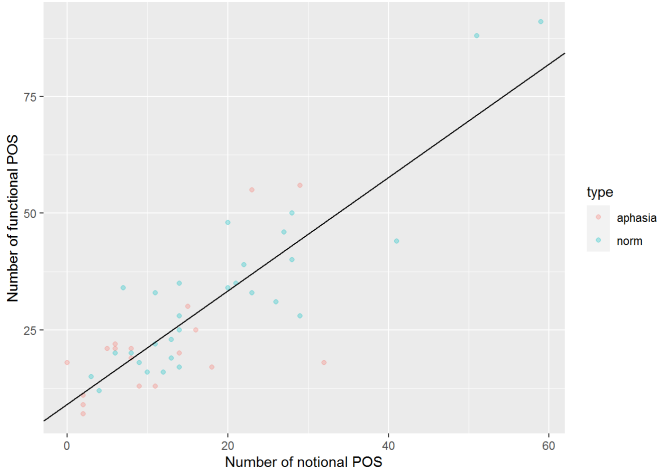
```
ggplot(data = micro, aes(x = num_pos, y = num_other, color = type)) +
  geom_point() +
  labs(x = "Number of notional POS",
       y = "Number of functional POS",
       title = "Variables correlation")
```




```
model <- lm(num_pos ~ num_other, data = micro)
model
```

```
##
## Call:
## lm(formula = num_pos ~ num_other, data = micro)
##
## Coefficients:
## (Intercept)    num_other
##       9.107        1.215
```

```
ggplot(data = micro, aes(x = num_other, y = num_pos, color=type))+
  geom_point(alpha = 0.3)+
  geom_abline(slope = model$coefficients[2], intercept = model$coefficients[1])+
  labs(x = "Number of notional POS",
       y = "Number of functional POS")
```




Again, as we saw, these variables seem to be positively associated.

## Inferential statistics

*Substantial hypothesis*:

Two types of speech should differ in the parameters. Explanation: for people with non-fluent aphasia it is hard to process and generate speech, so the pauses and fillers are supposed to be greater in comparison with those of the normal speech.

*Statistical hypotheses*:

*H0*: there is no difference between two type of speech considering some microparameters (the true correlation coefficient R is 0.)

*H1*: there is a difference between two type of speech considering some microparameters (the true correlation coefficient R is not 0.)

Let's convert the types into numerical values and perform two sample independent t-test for each variable. The independent t-test is chosen because the records are made separately, there is no matching between two samples and the number of records of two types also differs.

```
micro$type[which(micro$type == "norm")] = 1
micro$type[which(micro$type == "aphasia")] = 2
```

```
t.test(micro$mean_abs ~ micro$type, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  micro$mean_abs by micro$type
## t = -2.7562, df = 18.411, p-value = 0.01281
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1612.1695  -218.7683
## sample estimates:
## mean in group 1 mean in group 2
##        788.7928       1704.2617
```

```
t.test(micro$mean_fill ~ micro$type, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  micro$mean_fill by micro$type
## t = -2.5177, df = 44.835, p-value = 0.01545
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -366.34637  -40.69314
## sample estimates:
## mean in group 1 mean in group 2
##        244.0141        447.5339
```

```
t.test(micro$mean_in ~ micro$type, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  micro$mean_in by micro$type
## t = -2.7238, df = 21.387, p-value = 0.01259
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -825.6449 -111.1798
## sample estimates:
## mean in group 1 mean in group 2
##        618.2966       1086.7089
```

```
t.test(micro$mean_out ~ micro$type, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  micro$mean_out by micro$type
## t = -2.5426, df = 18.01, p-value = 0.02041
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2212.4694  -210.4882
## sample estimates:
## mean in group 1 mean in group 2
##        902.3407       2113.8194
```

```
t.test(micro$num_pos ~ micro$type, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  micro$num_pos by micro$type
## t = 2.3501, df = 44.046, p-value = 0.02331
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.581768 20.625128
## sample estimates:
## mean in group 1 mean in group 2
##        33.10345        22.00000
```

```
t.test(micro$num_other ~ micro$type, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  micro$num_other by micro$type
## t = 2.3584, df = 44.237, p-value = 0.02284
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.135014 14.458855
## sample estimates:
## mean in group 1 mean in group 2
##        19.24138        11.44444
```

P-values of all variables here are approximately 0, so at the 5% significance level we reject *H0* about the absence of the difference in microparameters. Thus, we can conclude that microparameters of discourse and the type of speech differ: even though people with no brain disorder may have some speech failures too, people diagnosed with non-fluent aphasia tend to have more serious discourse violations considering several microparameters.

The P-values that were obtained in the previous research were a bit higher in general but still nearly the same level, so the current results confirm previous conclusions.

**Conclusions:** 1. microparameters of discourse and the type of speech are different: Even though people with no brain disorder may have some speech failures too, people diagnosed with non-fluent aphasia tend to have more serious discourse violations considering several microparameters. 2. there are positive and negative correlations between the variables