Applying statistical methods to studies of Russian translations of Mayne Reid’s novel ‘The Boy Tar’

```
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(pROC)

## Warning: package 'pROC' was built under R version 3.6.3

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(ggplot2)
library(DescTools)

## Warning: package 'DescTools' was built under R version 3.6.3
```

Data (general description)

The project consists of 3 parts joined with the data - the Russian translations of Mayne Reid’s novel ‘The Boy Tar’. The text has a long translation history (see the sheme https://github.com/ElizavetaNosova/R_project_data/blob/master/translation_history.png, green means ‘full text is available’, orange means ‘some fragments of the text are available’, triangle means ‘the text is a shorten adapted version’). Different translations of the same text into the same language seems to be a suitable data to analyze the tendencies which depend both on the writer’s stylistical choice and the meaning of the text as if we analyze translations only the first factor is relevant. The data was chosen because there are a lot of versions of the text, it is quite long and, if full text is available, it can be splitted into chapters.Also I had some corresponding fragments of other texts retyped.

Pieces of data chosen for each task, the hypothesis and methods will be described in corresponding parts.

Part 1. Features with the same function. Metatext and appeals to the addressee

If some features have the same function, we can expect that they are correlated: either the translator dectises (or increases) the number of both features or replaces one feature with another. As the example of the isofunctional features, I have chosen metatext (all the comments the narrator makes about his speech) and appeals to the addressee. Fragments like 'you remember X' and 'I have already said about X' both organizes the speech structure. Dataset I use to analyze it is manually counted numbers of samples for full versions of the text. In long samples each sentence is counted separately. If there is both metatext and an appeal to the addressee in the sentence, it was counted twice (it does not happen enough often to make the features strongly correlated). This principle of samples detecting is easy to formalize, also it shows if the fragment was shorten. Also some old translations are the low-quality texts, and some text are edited versions of another ones, so we expend both random and deliberate changes to be presented in the texts. Some texts we translated from French translations (2 different ones). It would be difficult to estimate its impact on the changes, so I decided to analyze only Russian texts.

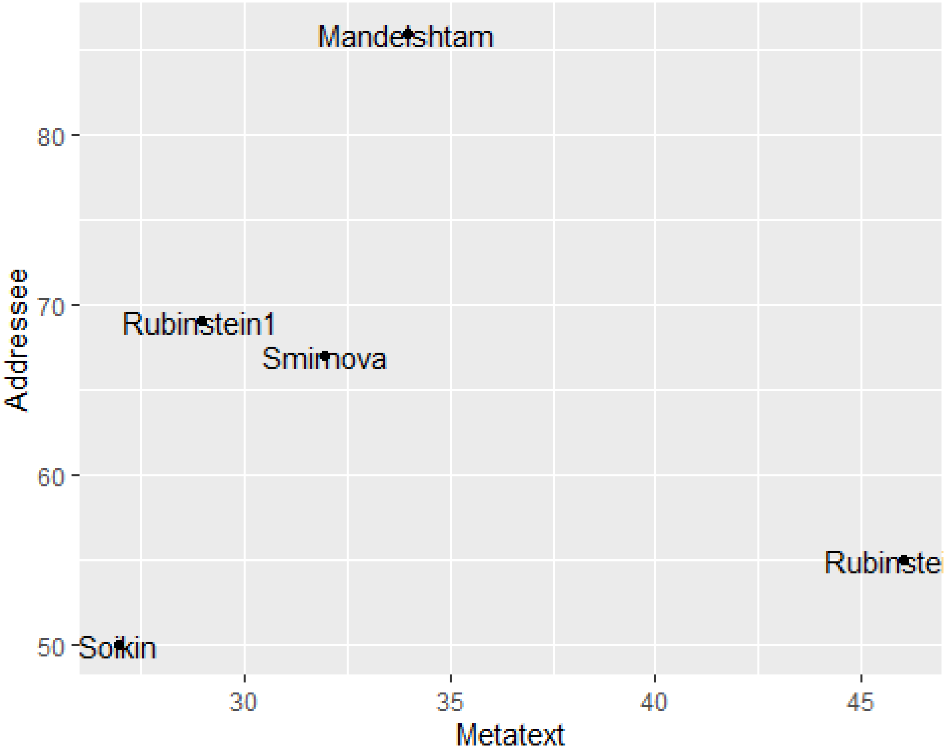
```
Part1_data <- read_csv('https://raw.githubusercontent.com/ElizavetaNosova/R_proj
ect_data/master/Metatext_addressee.csv')

## Parsed with column specification:
## cols(
##   Text = col_character(),
##   Metatext = col_double(),
##   Addressee = col_double()
## )
```

The hypothesis: the is a correlation between isofunctional structures, such as metatext and appellations to the addressee (they are izofunctional only for this novel)

To prove or obtain the zero hyphothesis, I need to do a correlation test. The number of this structures does not change linearly: there is a 'baseline' number - the number of such constructions in the sourse text, and the translator can decrease or increase it. The rank correlation (Spearman's correlation) is suitable for this task.

```
cor.test(Part1_data$Metatext, Part1_data$Addressee, method = 'spearman')

## 
##  Spearman's rank correlation rho
## 
## data:  Part1_data$Metatext and Part1_data$Addressee
## S = 14, p-value = 0.6833
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho 
## 0.3
```

```r
ggplot(Part1_data, aes(x=Metatext, y=Addressee)) + geom_point() + geom_text(labe
l=Part1_data$Text)
```



According to the plot there correlation really doesn’t exist. Maybe number of these constructions is changed independently, maybe both logically possible tendencies (to increase or decrease both / to substimate) are presented - the only interpretation is no correlation was found.

Part 2. Are readability metrics equal?

I expected old translations, which were made quickly, to have higher readability metrics then Soviet ones. The only available version of full translation is Sytin’s translation, edited by Smirnova in 2010-s. The changes she made are not serious enough to change readability metrics. Also I had full text of the latest version of Rubinstein translation. Zero hypothesis: Readability metrics of Sytin’s (=Smirnova’s) text are significantly higher when the corresponding metrics of Rubinshrain’s text.

4 readability metrics are chosen:

- mean sentence length,

- mean word length (in words),

- % of participles

- all the tokens they can be parsed as participles with Pymorphy2 were counted and divided into chapter length,

- % of verbal noun

- were found by lemma’s final [Микк. Оптимизация сложности учебного текста. 1981], the number divided into chapter length. Verbal nouns with zero suffix were not counted.

```
Rubinstein_readability = read_csv('https://raw.githubusercontent.com/ElizavetaNosova/R_project_data/master/Rubinstein_readability.csv')
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   chapter_text = col_character(),
##   mean_sent_len = col_double(),
##   mean_word_len = col_double(),
##   participles = col_double(),
##   verbal_nouns = col_double()
## )
```

```
Sytin_readability = read_csv('https://raw.githubusercontent.com/ElizavetaNosova/R_project_data/master/Smirnova_readability.csv')
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

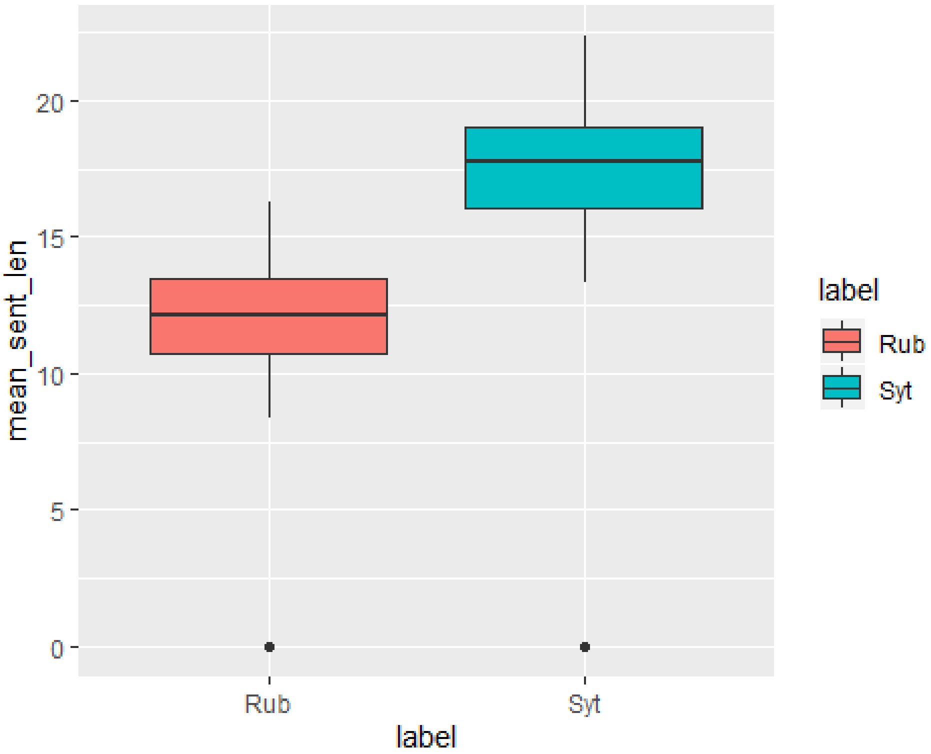
```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   chapter_text = col_character(),
##   mean_sent_len = col_double(),
##   mean_word_len = col_double(),
##   participles = col_double(),
##   verbal_nouns = col_double()
## )
```

```
Rubinstein_readability$label <- c(rep('Rub', nrow(Rubinstein_readability)))
Sytin_readability$label <- c(rep('Syt', nrow(Sytin_readability)))
Sytin_readability$dummy_label <- c(rep(0, nrow(Sytin_readability)))
Rubinstein_readability$dummy_label <- c(rep(1, nrow(Rubinstein_readability)))
```

```
readability <- data.frame(mean_sent_len =   c(Rubinstein_readability$mean_sent_len, Sytin_readability$mean_sent_len), mean_word_len = c(Rubinstein_readability$mean_word_len, Sytin_readability$mean_word_len), participles = c(Rubinstein_readability$participles, Sytin_readability$participles), verbal_nouns = c(Rubinstein_readability$verbal_nouns, Sytin_readability$verbal_nouns), label = c(Rubinstein_readability$label, Sytin_readability$label), dummy_label = c(Rubinstein_readability$dummy_label, Sytin_readability$dummy_label))
```

Firstly I’m going to do t-tests. They are not paired, as the text is slitted into chapters in different ways.

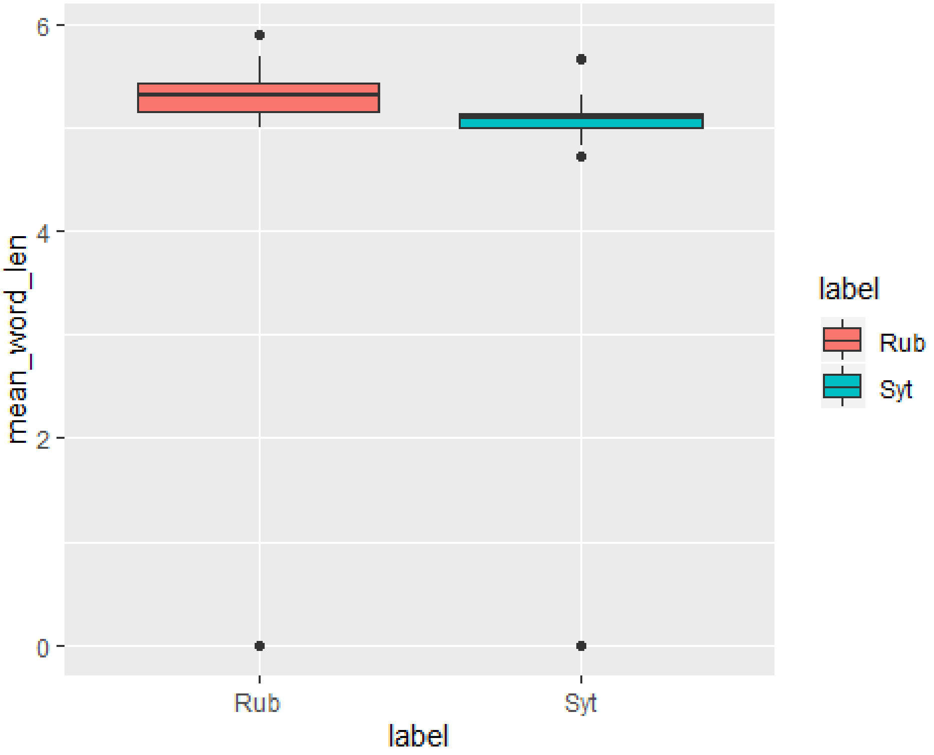
```
t.test(Rubinstein_readability$mean_sent_len, Sytin_readability$mean_sent_len, al
ternative = "less")

##
##  Welch Two Sample t-test
##
## data:  Rubinstein_readability$mean_sent_len and Sytin_readability$mean_sent_l
en
## t = -8.0695, df = 56.592, p-value = 2.766e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.164206
## sample estimates:
## mean of x mean of y
##  11.96388  17.21659

ggplot(data = readability, aes(y = mean_sent_len, x = label, fill = label, group
= label)) +
  geom_boxplot()
```



```
t.test(Rubinstein_readability$mean_word_len, Sytin_readability$mean_word_len, al
ternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Rubinstein_readability$mean_word_len and Sytin_readability$mean_word_l
en
## t = 1.6742, df = 67.405, p-value = 0.9506
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.5240215
## sample estimates:
## mean of x mean of y
##  5.219662  4.957150
```

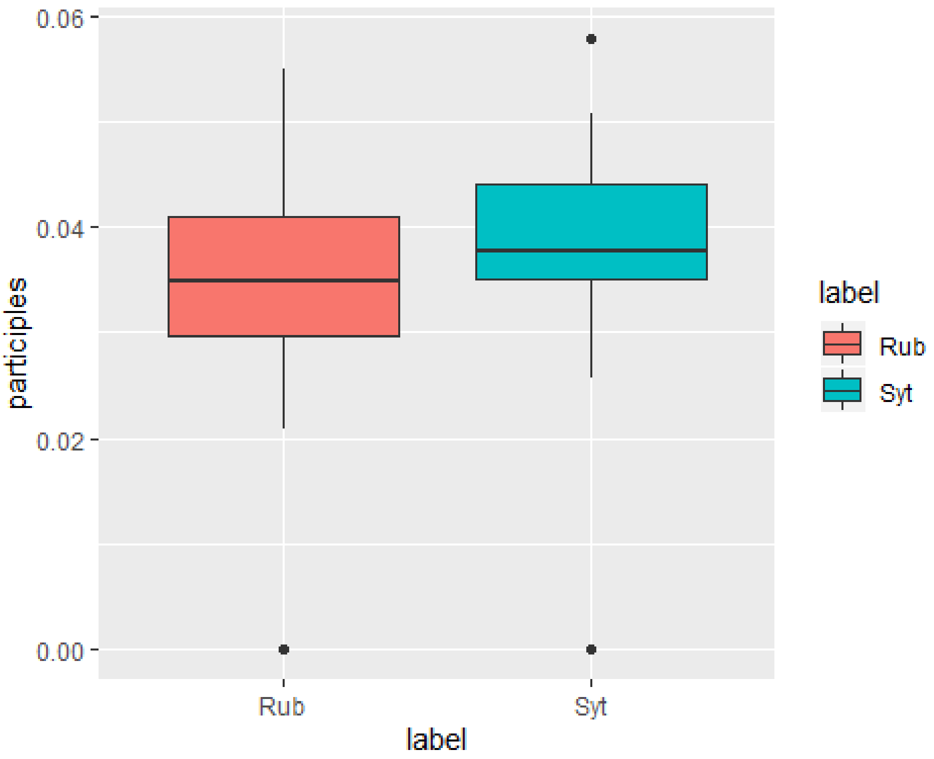
```
ggplot(data = readability, aes(y = mean_word_len, x = label, fill = label, group
= label)) +
  geom_boxplot()
```



```
t.test(Rubinstein_readability$participles, Sytin_readability$participles, altern
ative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Rubinstein_readability$participles and Sytin_readability$participles
## t = -1.7596, df = 78.16, p-value = 0.0412
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.000168801
## sample estimates:
##  mean of x  mean of y
## 0.03541531 0.03854228
```

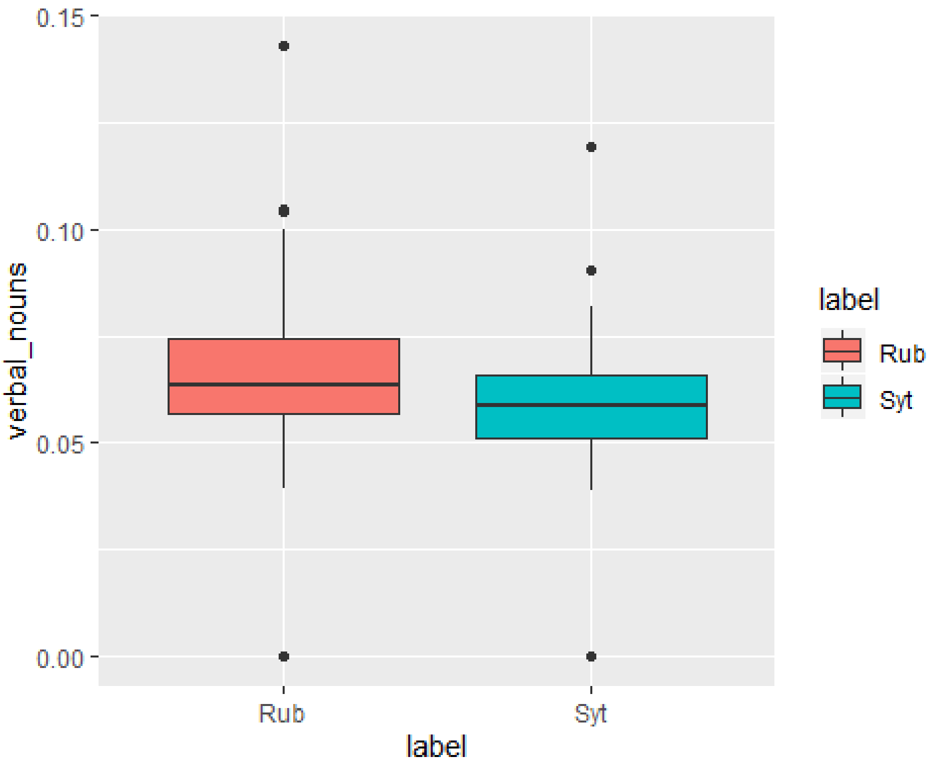
```
ggplot(data = readability, aes(y = participles, x = label, fill = label, group =
label)) +
  geom_boxplot()
```



```
t.test(Rubinstein_readability$verbal_nouns, Sytin_readability$verbal_nouns, alte
rnative = "less")
```

```
## 
##  Welch Two Sample t-test
## 
## data:  Rubinstein_readability$verbal_nouns and Sytin_readability$verbal_nouns
## t = 1.8062, df = 86.965, p-value = 0.9628
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.01241589
## sample estimates:
##  mean of x  mean of y 
## 0.06581363 0.05934869
```

```
ggplot(data = readability, aes(y = verbal_nouns, x = label, fill = label, group
= label)) +
  geom_boxplot()
```



For mean sentence length and number of participles we can obtain the zero hypothesis. The boxplots shows, that two other metrics - mean word length and number of verbal nouns - are higher for Rubinstein's translation (label 0)

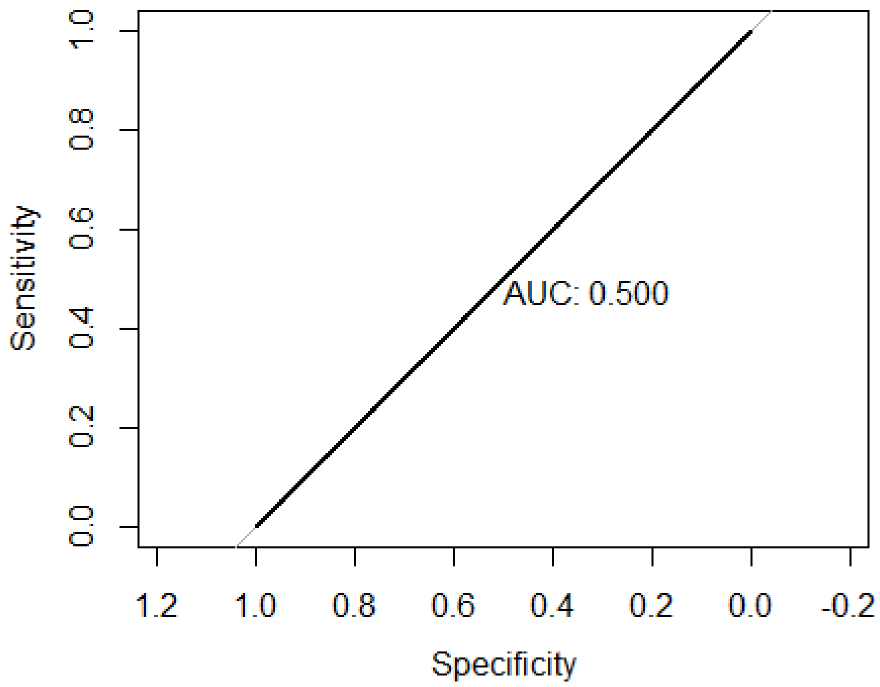
Also I used logisted regression to check if the readability metrics can be used to differenciate the translations Althowgh prediction of the model I will get cannot be used for any senseful task, I splitted data into train and test subsets to estimate how good the prediction would be (if we can predict which text is a fragment from using some metrics, the metrics are really different)

```
test.index <- c(10, 20, 30, 40, 50, 60, 70, 80, 90,100)
readability.train <- readability[-test.index, ]
readability.test <- readability[test.index, ]
```

```
readability_regression <- glm(dummy_label ~ mean_sent_len + mean_word_len + part
iciples + verbal_nouns, data=readability.train, family = binomial())

summary(readability_regression)

##
## Call:
## glm(formula = dummy_label ~ mean_sent_len + mean_word_len + participles +
##     verbal_nouns, family = binomial(), data = readability.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.81016  -0.01715   0.00429   0.03857   1.73117
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.2142     1.4171  -0.151  0.87983
## mean_sent_len   -2.3003     0.8963  -2.566  0.01027 *
## mean_word_len    8.6121     3.1808   2.708  0.00678 **
## participles   -244.3526   110.8701  -2.204  0.02753 *
## verbal_nouns    -4.4738    42.2471  -0.106  0.91567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 125.041  on 94  degrees of freedom
## Residual deviance:  21.982  on 90  degrees of freedom
## AIC: 31.982
##
## Number of Fisher Scoring iterations: 9
```

No readability metrics have *** importence, only one has ** importence

```
readability.test$predicted_label <- predict(readability_regression, readability.
test, type='response')
roc(readability.test$predicted_label,readability.test$dummy_label,
            plot=TRUE,
            print.auc=TRUE)

## Warning in roc.default(readability.test$predicted_label,
## readability.test$dummy_label, : 'response' has more than two levels. Consider
## setting 'levels' explicitly or using 'multiclass.roc' instead

## Setting levels: control = 5.3151025672537e-08, case = 9.12736755341886e-05

## Setting direction: controls < cases
```



```
##
## Call:
## roc.default(response = readability.test$predicted_label, predictor = readabil
ity.test$dummy_label,     plot = TRUE, print.auc = TRUE)
##
## Data: readability.test$dummy_label in 1 controls (readability.test$predicted_
label 5.3151025672537e-08) < 1 cases (readability.test$predicted_label 9.1273675
5341886e-05).
## Area under the curve: 0.5
```

According to ROC, prediction result is not better than random guessing. Although there are some tendencies in these metrics proportion, they cannot be used to predict which text is the chapter from. Although, I am not sure if any two children fiction texts can be differentiated using readability metrics.

Part 3. Adapted translations studied

There are two text versions, adapted for young children whet the corresponding full versions, which are shorter, when their source texts (full Russian translations). I am going to check some hypothesis. As available data format differs (it is described below).

3a. Sytin vs 2 parts of Lyalitskaya adaptation I have only retyped corresponding fragments (63), which seems to be edited by Lyalitskaya. Hypothesis: as Lyalitskaya tries to make text shorter, she uses less conjunctions then Sytin does.

Preprocessing and method: I used Python library difflib to get the list of deleted and added words for each pair of text fragments. After that I counted added and deleted adjectives. As the samples are examples of changes (not examples of deleted or recreated fragments), I suggested, that, if my hypothesis is correct, number of deleted conjunctions (counted for each pair of text fragments separately) would be significantly larger, then the number of added oned. If we do a paired test, we can be sure, that the replaced conjunctions (1 added, 1 deleted) do not affect the result

```
sytin.vs.lyalitskaya <- read_csv('https://raw.githubusercontent.com/ElizavetaNos
ova/R_project_data/master/Sytin_Lyalitskaya_conj.csv')
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   S = col_character(),
##   L = col_character(),
##   dif = col_character(),
##   delited = col_character(),
##   added = col_character(),
##   delited_poses = col_character(),
##   added_poses = col_character(),
##   added_conj = col_double(),
##   delited_conj = col_double()
## )
```

```
MeanDiffCI(sytin.vs.lyalitskaya$added_conj, sytin.vs.lyalitskaya$delited_conj)
```

```
##   meandiff     lwr.ci     upr.ci
## -1.1269841 -1.5811539 -0.6728143
```

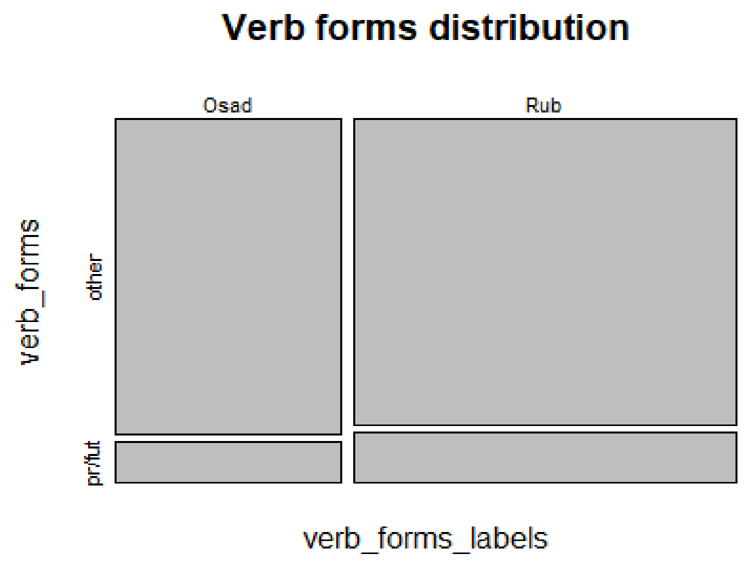
According to the test result, Lyalitskaya reduces amount of conjunctions.

3b. There is one more pair ‘full translation and adaptation’: Osadchuck has prepared audioadaptation using Rubinstein’s translation (last version). Both text are available; but the main method she used to change text is deleting fragments of the text and some word changes. As full texts are available, it is possible to detect all the samples of using or not using something and analyze its distribution with chi-squared test.

For both hypothesis samples were counted using Python, and at this R notebook I reproduce the datasets which I would get if I wrote down the category of each observation point.

Hypothesis 3b-1. Osadchuck prefers present and future verb forms to make the point of view shift. Texts have been parsed with pymorhy, after that present/future and other verb forms (participles included) have been counted.

```
verb_forms <- c(rep('pr/fut', 1552), rep('other', 9672), rep('pr/fut', 750), rep
('other', 5885))
verb_forms_labels <- c(rep('Rub', 1552+9672), rep('Osad', 750+5885))
verb_forms_df <- data.frame(verb_forms_labels, verb_forms)

verb_forms.table <- table(verb_forms_df)
verb_forms.table

##                   verb_forms
## verb_forms_labels other pr/fut
##              Osad  5885    750
##              Rub   9672   1552

chisq.test(verb_forms.table)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  verb_forms.table
## X-squared = 23.431, df = 1, p-value = 1.295e-06

mosaicplot(verb_forms.table, main='Verb forms distribution')
```

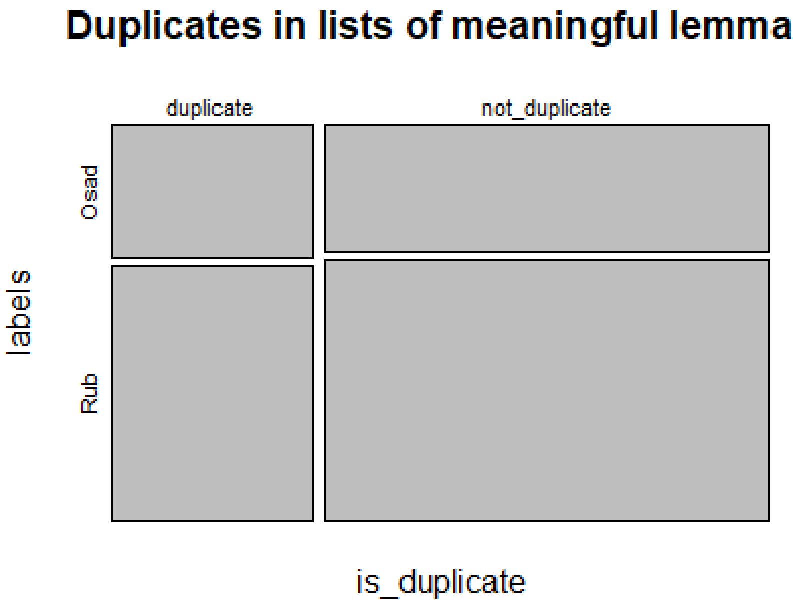


No significant difference in distribution was found

If I had detected a significant difference, I was going to select fragments which I presented in both text, to check, if the difference is caused by

Hypothesis 3b-2. There are some repetitions in full texts, and Osadchuck seems to lessen its number. I expected the proportion of meaningful lemmas (not conjunctions, prepositions, participles and pronouns), which appears at the chapter for the first time and not for the first time, to be different

```
is_duplicate <- c(rep('duplicate', 119699), rep('not_duplicate', 272836), rep('d
uplicate', 63617),rep('not_duplicate', 134408))
labels <- c(rep('Rub', 119699+272836), rep('Osad', 63617+134408))
duplicates_df <- data.frame(is_duplicate , labels)

duplicates_table <- table(duplicates_df)
duplicates_table

##                 labels
## is_duplicate      Osad    Rub
##   duplicate      63617 119699
##   not_duplicate 134408 272836

chisq.test(duplicates_table)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  duplicates_table
## X-squared = 163.68, df = 1, p-value < 2.2e-16

mosaicplot(duplicates_table, main='Duplicates in lists of meaningful lemmas')
```



For this hypothesis also no significant difference was found.

Conclusion Interpretation of the results is presented in the main parts of the project. It seems that local changes in the texts editions do not change the distribution of the features. Maybe, I should have specified contexts which I analize working with full texts (especially of adaptations) as even if the changes are made purposefully they are not detected by tests as there are a lot of other context there the changes are not expected.