Russian comparative function words

Kirill Konovalov, Petr Filchenkov

12/06/2020

Hypotheses

The aim of this research is to analyze the representation of such Russian comparative function words as “будто”, “словно”, “а-ля”, “аки”, “как”, “подобно” in the Russian National Corpus.

We have several hypotheses to be tested. Some of them are based on the idea that each of these words occurs in texts of a specific genre:

1. Such conjunctions as “словно”, “аки” and “подобно” are supposed to occur mostly in belles-lettres texts;
2. “как” and “будто” seem to be stylistically neutral and should have even distribution in texts of all styles;
3. “а-ля” is expected to occur in the texts of the colloquial style.

The other group of hypotheses is devoted to the date of the texts:

1. We expect the word “аки” to appear in the texts written earlier than the XX century more often than the other words.
2. There is a correlation between the historical period when the text was written and the usage of a particular part of speech after the comparative words.

Research design

To test the hypotheses we need such statistical information regarding the comparative words as the style of the text where they occur, the date, and the next word part-of-speech. In terms of statistics, the null hypothesis runs as follows:

1. the usage of the words under consideration does not correlate with the style or date;
2. the usage of a particular part of speech after the comparative words does not correlate with the historical period when the text was written

As for statistical tools, we will use:

- exploratory data analysis to observe the general statistical information regarding the data;
- the binomial test to check the assumptions regarding the usage of the words “словно”, “аки” and “подобно” in belles-lettres texts, the usage of the words ‘как’ and ‘будто’ in all styles and the usage of “а-ля” in the texts of colloquial style;
- the binomial test to compare the probabilities of different part-of-speech usage in the context of comparative words in different historical periods;
- the chi-squared test to see whether there is a statistically significant correlation between the usage of a particular comparative function word and the date when the text was written, the usage of comparative words and left (or right) context or functional style ;
- the logistic regression model to determine how the particular year and each comparative word frequency are correlated

Data collection method

The data for this research is based on the Russian National Corpus that includes several smaller corpora. We collected the information about the words of the study from the written corpus. The documents were downloaded as xlsx files and preprocessed. We used python to randomly pick 250 words from each document and combined them in one csv dataset. Then we deleted the unnecessary columns and analysed left and right context words and added POS tags for them. Moreover, since there were so many variables in columns “Type”, “Created”, “Left_context_POS” and “Right_context_POS”, we decided to decrease their number. For the “Type” column we divided the genres of the documents into 5 groups of Russian functional styles: scientific, official, publicistic, belles-lettres, and colloquial. The dates from the “Created” column were organized into decades. The number of part of speech tags was also decreased from 24 to 10 by combining some of them into groups.

Description of the data

```
library(tidyverse)
```

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

```
## -- Attaching packages ---------------------------------------------------------------------------- tidyvers
e 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ tidyverse_conf
licts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
urlfile="https://raw.githubusercontent.com/a-la-r/data/master/df_250_wr.csv"
full_data<-read.csv(url(urlfile), sep = ";", encoding='UTF-8')
data<- subset(full_data, select = c(Center, Created, Type, Left_context_POS, Right_context_POS))
```

Let's look at the contents of the dataset. It has the following columns:

Center - the analyzed word Created - year of creation of the source document Type - functional style of the Russian language Left_context_POS - part of speech of the first word to the left of the analyzed one Right_context_POS - part of speech of the first word to the right of the analyzed one

```
head(data)
```

```
##   Center Created             Type Left_context_POS Right_context_POS
## 1    как    1940   художественный             VERB              NOUN
## 2    как    1910 публицистический             NPRO              NOUN
## 3    как    1720          деловой             NOUN              ADVB
## 4    как    1780          деловой             NPRO              TRSH
## 5    как    1960 публицистический             VERB              NOUN
## 6    как    1900 публицистический             VERB              NPRO
```

Exploratory data analysis, descriptive statistics and visualization

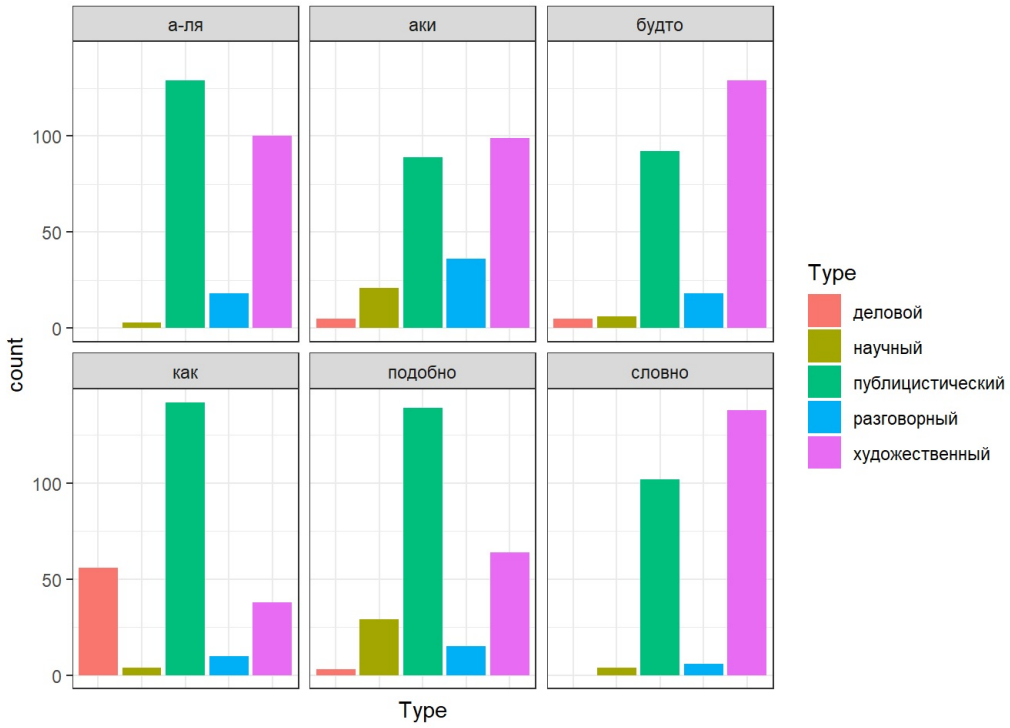
In this section, we are going to look at some statistics and several graphs to get a better understanding of the data and its distribution.

```
summary(data)
```

```
##      Center       Created                   Type     Left_context_POS
##  а-ля   :250   Min.   :1700   деловой         : 69   NOUN   :603
##  аки    :250   1st Qu.:1860   научный         : 67   AUX    :295
##  будто  :250   Median :1912   публицистический:693   VERB   :199
##  как    :250   Mean   :1907   разговорный     :103   ADJ    :126
##  подобно:250   3rd Qu.:1980   художественный  :568   NPRO   : 92
##  словно :250   Max.   :2017                          ADVB   : 81
##                                                      (Other):104
##  Right_context_POS
##  NOUN    :527
##  AUX     :305
##  ADJ     :258
##  VERB    :135
##  NPRO    :101
##  PRT_GRND: 73
##  (Other) :101
```

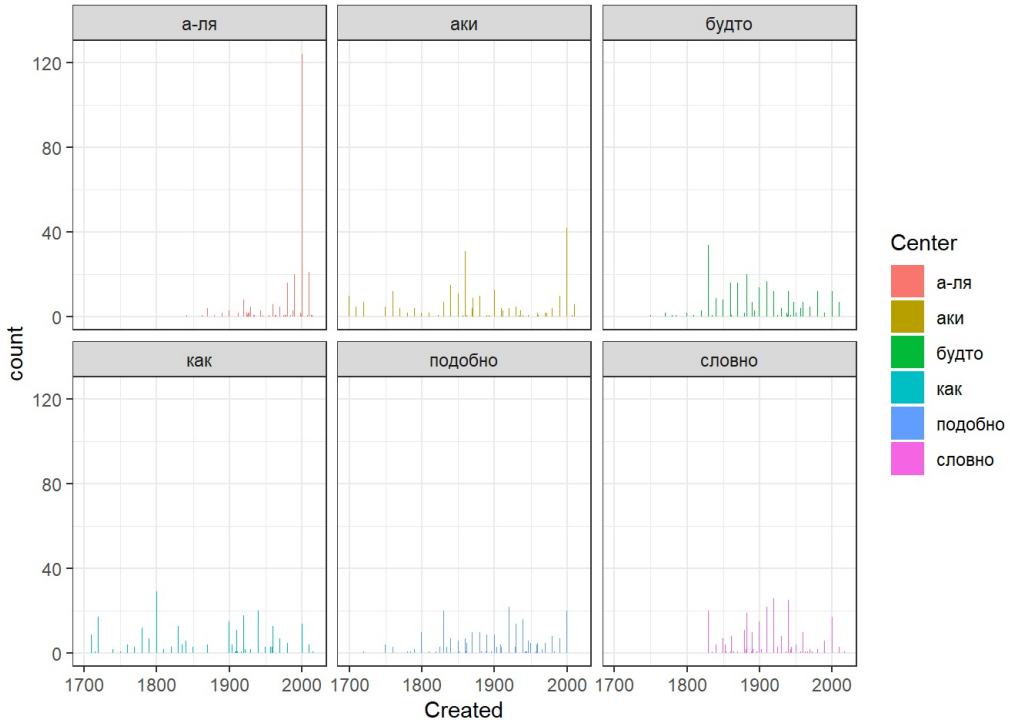
From this summary, we can conclude that all the texts are distributed between the following years: 1700 and 2017, the median year is 1912. Later, in some experiments, we will divide the year distribution into 2 halves. The majority of texts belong to publicistic and belles-lettres style. Nouns, auxiliary words, verbs, and adjectives occur more often in the context of the comparative words.

```
data %>%
  ggplot(aes(Type, fill = Type)) +
  geom_bar() +
  theme_bw() +
  facet_wrap(~Center) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



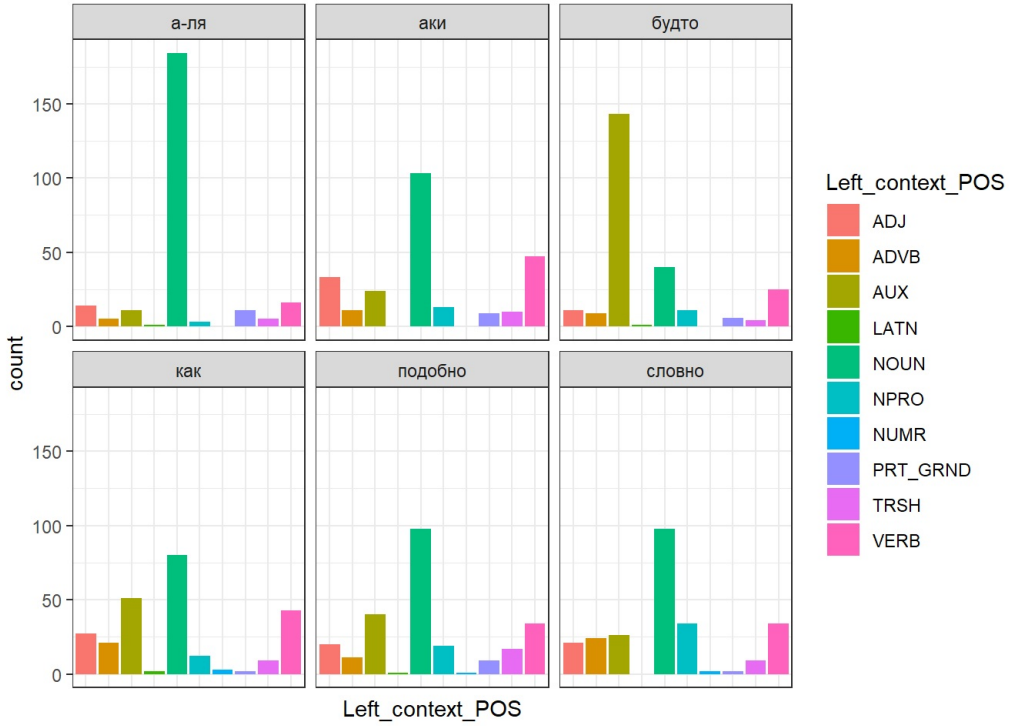
In these plots we can see that the distribution of the analyzed words according to the functional styles has some similarities as well as differences. The most used styles among the words are belles-lettres and publicistic. The least used are official and scientific. However, the proportions of the styles in each word vary. The word “а-ля” is most often used in publicistic documents and a bit less in belles-lettres texts. There are very few occurrences of it in colloquial and scientific texts whereas it is not used in official texts at all. The word “словно” is distributed similarly, however, it appears more in fiction rather than in publicistic texts. “Будто” and “аки” follow the same trend but they also occur in official texts. “Подобно” is seen in all the styles as well and mostly in publicistic documents. Although the word “как” is primarily used in publicistic texts, it is also represented in the official style more often than any other word.

```
data %>%
  ggplot(aes(Created, fill = Center)) +
  geom_bar() +
  theme_bw() +
  facet_wrap(~Center)
```



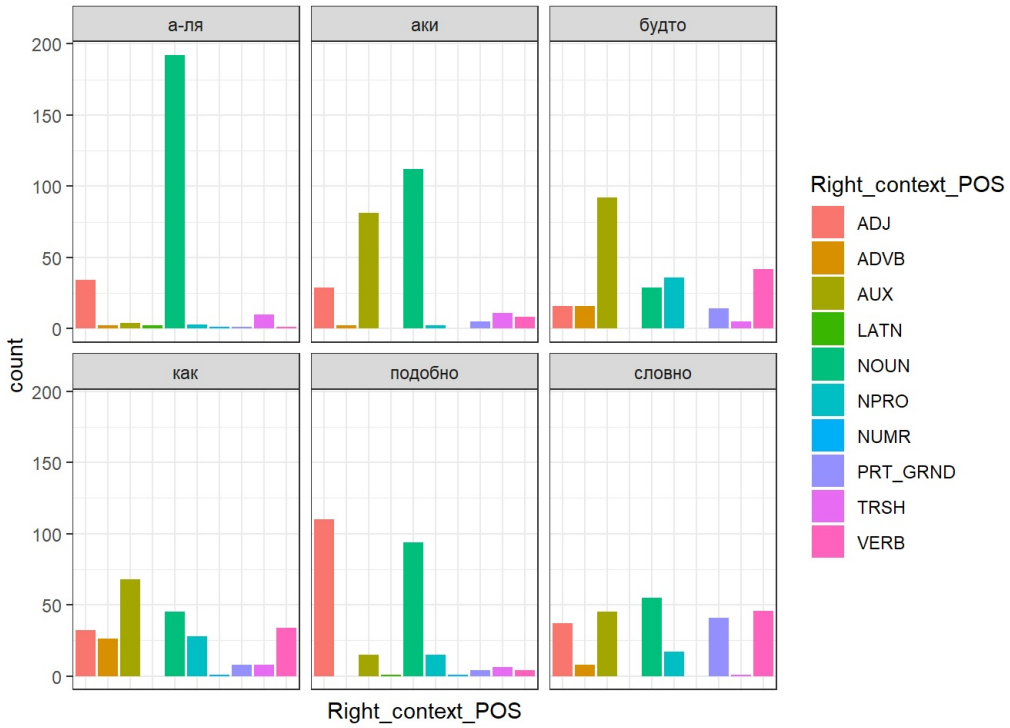
As we can see such words as “как” and “аки” are almost evenly distributed along the timeline. “Будто”, “подобно” and “словно” tend to be used more often after the 1800s. The word “а-ля” is not represented in the data before the second half of the 19th century and it became very popular closer to the year 2000.

```
data %>%
  ggplot(aes(Left_context_POS, fill = Left_context_POS)) +
  geom_bar() +
  theme_bw() +
  facet_wrap(~Center) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



The bar charts shown above demonstrate that in 5 out of 6 cases the most popular part of speech to the left of the analyzed word is a noun. The word “будто” most commonly has auxiliary words in the left context. This group includes conjunctions, prepositions and particles.

```
data %>%
  ggplot(aes(Right_context_POS, fill = Right_context_POS)) +
  geom_bar() +
  theme_bw() +
  facet_wrap(~Center) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



The distribution of the parts of speech of the first word in the right context is very diverse. In case of “а-ля” there are mostly only nouns and few adjectives. The word “аки” has two almost equal groups with nouns and auxiliary words (conjunctions, prepositions, particles). As for “подобно”, it has two groups with adjectives and nouns, which are commonly used after it. Auxiliary words dominate in the right context after “будто”. In the

case of “как” and “словно” there is a variety of parts speech used. In both examples we usually observe nouns, adjectives, verbs and auxiliary words. After “словно” we also have a number of participles and gerunds.

Results of applying statistical tests and modeling

First of all, we are going to start with several binomial tests. Secondly, we will look at the results of the chi-squared test and the application of the linear regression model.

Binomial tests

Test 1

In this test, we are going to check the probability of the words “как” and “будто” usage in 5 samples (one sample for each functional style)

H0: the chosen words appear in each sample with the probability different from ⅓

H1: the chosen words appear in each sample with the probability ⅓

```
kak_budto <- data
kak_budto$Center <- gsub("будто", 1, kak_budto$Center)
kak_budto$Center <- gsub("словно", 0, kak_budto$Center)
kak_budto$Center <- gsub("аки", 0, kak_budto$Center)
kak_budto$Center <- gsub("а-ля", 0, kak_budto$Center)
kak_budto$Center <- gsub("как", 1, kak_budto$Center)
kak_budto$Center <- gsub("подобно", 0, kak_budto$Center)


belles_lettres2 <- kak_budto[kak_budto$Type=="художественный",]

total <- nrow(belles_lettres2)
kak_budto <- nrow(belles_lettres2[belles_lettres2$Center==1,]) # number of successes


binom.test(kak_budto, total)
```

```
##
##  Exact binomial test
##
## data:  kak_budto and total
## number of successes = 167, number of trials = 568, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2568208 0.3333656
## sample estimates:
## probability of success
##              0.2940141
```

```
kak_budto <- data
kak_budto$Center <- gsub("будто", 1, kak_budto$Center)
kak_budto$Center <- gsub("словно", 0, kak_budto$Center)
kak_budto$Center <- gsub("аки", 0, kak_budto$Center)
kak_budto$Center <- gsub("а-ля", 0, kak_budto$Center)
kak_budto$Center <- gsub("как", 1, kak_budto$Center)
kak_budto$Center <- gsub("подобно", 0, kak_budto$Center)

publ <- kak_budto[kak_budto$Type=="публицистический",]

total <- nrow(publ)
kak_budto <- nrow(publ[publ$Center==1,]) # number of successes


binom.test(kak_budto, total)
```

```
##
##  Exact binomial test
##
## data:  kak_budto and total
## number of successes = 234, number of trials = 693, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3024865 0.3742278
## sample estimates:
## probability of success
##              0.3376623
```

```
kak_budto <- data
kak_budto$Center <- gsub("будто", 1, kak_budto$Center)
kak_budto$Center <- gsub("словно", 0, kak_budto$Center)
kak_budto$Center <- gsub("аки", 0, kak_budto$Center)
kak_budto$Center <- gsub("а-ля", 0, kak_budto$Center)
kak_budto$Center <- gsub("как", 1, kak_budto$Center)
kak_budto$Center <- gsub("подобно", 0, kak_budto$Center)

coll <- kak_budto[kak_budto$Type=="разговорный",]

total <- nrow(coll)
kak_budto <- nrow(coll[coll$Center==1,]) # number of successes


binom.test(kak_budto, total)
```

```
##
##  Exact binomial test
##
## data:  kak_budto and total
## number of successes = 28, number of trials = 103, p-value = 4.028e-06
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1888431 0.3684001
## sample estimates:
## probability of success
##              0.2718447
```

```
kak_budto <- data
kak_budto$Center <- gsub("будто", 1, kak_budto$Center)
kak_budto$Center <- gsub("словно", 0, kak_budto$Center)
kak_budto$Center <- gsub("аки", 0, kak_budto$Center)
kak_budto$Center <- gsub("а-ля", 0, kak_budto$Center)
kak_budto$Center <- gsub("как", 1, kak_budto$Center)
kak_budto$Center <- gsub("подобно", 0, kak_budto$Center)

official <- kak_budto[kak_budto$Type=="деловой",]

total <- nrow(official)
kak_budto <- nrow(official[official$Center==1,]) # number of successes


binom.test(kak_budto, total)
```

```
##
##  Exact binomial test
##
## data:  kak_budto and total
## number of successes = 61, number of trials = 69, p-value = 3.243e-11
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.7842675 0.9485934
## sample estimates:
## probability of success
##               0.884058
```

```
kak_budto <- data
kak_budto$Center <- gsub("будто", 1, kak_budto$Center)
kak_budto$Center <- gsub("словно", 0, kak_budto$Center)
kak_budto$Center <- gsub("аки", 0, kak_budto$Center)
kak_budto$Center <- gsub("а-ля", 0, kak_budto$Center)
kak_budto$Center <- gsub("как", 1, kak_budto$Center)
kak_budto$Center <- gsub("подобно", 0, kak_budto$Center)

sci <- kak_budto[kak_budto$Type=="научный",]

total <- nrow(sci )
kak_budto <- nrow(sci[sci$Center==1,]) # number of successes


binom.test(kak_budto, total)
```

```
##
##  Exact binomial test
##
## data:  kak_budto and total
## number of successes = 10, number of trials = 67, p-value = 4.042e-09
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.07396468 0.25740245
## sample estimates:
## probability of success
##              0.1492537
```

We can see that the null hypothesis can be rejected only for the publicistic style experiment. The tests on belles-lettres and colloquial samples have performed the results close to hypothesis 1, while the experiments on official and scientific styles have completely different (from ⅓) results.

Test two

We are going to test if the words “подобно”, “словно” and “аки” occur more often in belles-lettres texts than the words “а-ля”, “как” and “будто”.

H0: the chosen words appear in belles-lettres texts with the probability 1/2

H1: the chosen words appear in belles-lettres texts more often

```
podobno_slovno <- data
podobno_slovno$Center <- gsub("будто", 0, podobno_slovno$Center)
podobno_slovno$Center <- gsub("словно", 1, podobno_slovno$Center)
podobno_slovno$Center <- gsub("аки", 1, podobno_slovno$Center)
podobno_slovno$Center <- gsub("а-ля", 0, podobno_slovno$Center)
podobno_slovno$Center <- gsub("как", 0, podobno_slovno$Center)
podobno_slovno$Center <- gsub("подобно", 1, podobno_slovno$Center)

belles_lettres2 <- podobno_slovno[podobno_slovno$Type=="художественный",]

total <- nrow(belles_lettres2)
podobno_slovno <- nrow(belles_lettres2[belles_lettres2$Center==1,]) # number of successes

binom.test(podobno_slovno, total)
```

```
##
##  Exact binomial test
##
## data:  podobno_slovno and total
## number of successes = 301, number of trials = 568, p-value = 0.1661
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4879392 0.5716072
## sample estimates:
## probability of success
##              0.5299296
```

From the result of this test it is seen that the probability of success equals 0.5299296 and p-value is 2.454e-13, which is less than the significance level of 5%. We reject the null hypothesis and can claim that hypothesis one is correct, which means that the words “будто”, “словно” and “аки” do occur more often in belles-lettres texts.

Test three

Here we examine the idea that the word “а-ля” occurs in the colloquial texts more commonly than the rest of the researched words.

H0: the chosen word appears in the colloquial texts with the probability 1/6

H1: the chosen word appears in the colloquial texts more often

```
ala_coll <- data
ala_coll$Center <- gsub("будто", 0, ala_coll$Center)
ala_coll$Center <- gsub("словно", 0, ala_coll$Center)
ala_coll$Center <- gsub("аки", 0, ala_coll$Center)
ala_coll$Center <- gsub("а-ля", 1, ala_coll$Center)
ala_coll$Center <- gsub("как", 0, ala_coll$Center)
ala_coll$Center <- gsub("подобно", 0, ala_coll$Center)

colloquial <- ala_coll[ala_coll$Type=="разговорный",]

total <- nrow(colloquial)
nbudto_slovno <- nrow(colloquial[colloquial$Center==1,]) # number of successes

binom.test(nbudto_slovno, total)
```

```
##
##  Exact binomial test
##
## data:  nbudto_slovno and total
## number of successes = 18, number of trials = 103, p-value = 1.363e-11
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1069879 0.2620572
## sample estimates:
## probability of success
##              0.1747573
```

The probability of success is 0.1747573 and the p-value equals 1.363e-11, so we reject the null hypothesis and can say that “а-ля” is used more in the colloquial style.

Test four

In the fourth test, we want to check if the word “аки” appears in the old texts before the 1900s more commonly than other words.

H0: the chosen word appears in the old texts with the probability 1/6

H1: the chosen word appears in old texts more often

```
aki <- data
aki$Center <- gsub("будто", 0, aki$Center)
aki$Center <- gsub("словно", 0, aki$Center)
aki$Center <- gsub("аки", 1, aki$Center)
aki$Center <- gsub("а-ля", 0, aki$Center)
aki$Center <- gsub("как", 0, aki$Center)
aki$Center <- gsub("подобно", 0, aki$Center)

old_texts <- aki[aki$Created<=1900,]
total <- nrow(old_texts)
naki <- nrow(old_texts[old_texts$Center==1,]) # number of successes

binom.test(naki, total)
```

```
##
##  Exact binomial test
##
## data:  naki and total
## number of successes = 158, number of trials = 674, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2029181 0.2682697
## sample estimates:
## probability of success
##              0.2344214
```

```
new_texts <- aki[aki$Created>=1900,]
total <- nrow(new_texts)
naki <- nrow(new_texts[new_texts$Right_context_POS==1,])

binom.test(naki, total)
```

```
##
##  Exact binomial test
##
## data:  naki and total
## number of successes = 0, number of trials = 895, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.000000000 0.004113171
## sample estimates:
## probability of success
##                      0
```

The probability of success is 0.2344214 and the p-value < 2.2e-16, so we reject the null hypothesis and the test proves that “аки” is more common for old texts.

Test five

In our fifth test, we try to prove whether there is a dependence between the usage of some particular part of speech after the comparative words and time period of their usage.

H0: the part of speech after the comparative words appears in the texts before and after 1900 with the same probability

H1: the part of speech appear more often after the comparative words after 1900

In this case, we check the usage of nouns after them before and after 1900.

Nouns after comparative words before XX

```
nouns <- data
nouns$Right_context_POS <- gsub("NOUN", 1, nouns$Right_context_POS)
nouns$Right_context_POS[nouns$Right_context_POS != 1] <- 0

old_texts <- nouns[nouns$Created<=1900,]
total <- nrow(old_texts)
nnouns <- nrow(old_texts[old_texts$Right_context_POS==1,])

binom.test(nnouns, total)
```

```
##
##  Exact binomial test
##
## data:  nnouns and total
## number of successes = 184, number of trials = 674, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2396782 0.3083175
## sample estimates:
## probability of success
##               0.272997
```

Nouns after comparative words in XX-XXI

```
new_texts <- nouns[nouns$Created>=1900,]
total <- nrow(new_texts)
nnouns <- nrow(new_texts[new_texts$Right_context_POS==1,])

binom.test(nnouns, total)
```

```
##
##  Exact binomial test
##
## data:  nnouns and total
## number of successes = 358, number of trials = 895, p-value = 2.384e-09
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3677204 0.4329401
## sample estimates:
## probability of success
##                    0.4
```

The probability of noun usage before 1900 is 0.272997, while after 1900 it is 0.4. So we can conclude that nouns are more likely to be used after comparative words in XX-XXI centuries.

Here we test the usage of auxiliary words before and after 1900.

```
aux <- data
aux$Right_context_POS <- gsub("AUX", 1, aux$Right_context_POS)
aux$Right_context_POS[aux$Right_context_POS != 1] <- 0

old_texts <- aux[aux$Created<=1900,]
total <- nrow(old_texts)
naux <- nrow(old_texts[old_texts$Right_context_POS==1,])

binom.test(naux, total)
```

```
##
##  Exact binomial test
##
## data:  naux and total
## number of successes = 180, number of trials = 674, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.234001 0.302178
## sample estimates:
## probability of success
##              0.2670623
```

Auxiliary words after comparative words in XX-XXI

```
new_texts <- aux[aux$Created>=1900,]
total <- nrow(new_texts)
naux <- nrow(new_texts[new_texts$Right_context_POS==1,])

binom.test(naux, total)
```

```
##
##  Exact binomial test
##
## data:  naux and total
## number of successes = 140, number of trials = 895, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1332293 0.1819066
## sample estimates:
## probability of success
##              0.1564246
```

The probability of auxiliary word usage before 1900 is 0.2670623, while after 1900 it is 0.1564246. In other words, auxiliary words occur after comparative words more often in the past.

The same test has been performed for adjectives

Adjectives after comparative words before XX

```
adj <- data
adj$Right_context_POS <- gsub("ADJ", 1, adj$Right_context_POS)
adj$Right_context_POS[adj$Right_context_POS != 1] <- 0

old_texts <- adj[adj$Created<=1900,]
total <- nrow(old_texts)
nadj <- nrow(old_texts[old_texts$Right_context_POS==1,])

binom.test(nadj, total)
```

```
##
##  Exact binomial test
##
## data:  nadj and total
## number of successes = 116, number of trials = 674, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1443541 0.2027651
## sample estimates:
## probability of success
##              0.1721068
```

Adjectives after comparative words in XX-XXI

```
new_texts <- adj[adj$Created>=1900,]
total <- nrow(new_texts)
nadj <- nrow(new_texts[new_texts$Right_context_POS==1,])

binom.test(nadj, total)
```

```
##
##  Exact binomial test
##
## data:  nadj and total
## number of successes = 151, number of trials = 895, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1447439 0.1948893
## sample estimates:
## probability of success
##              0.1687151
```

The probability of adjective usage before 1900 is 0.1721068, while after 1900 it is 0.1687151. It means that adjectives appear in both historical periods almost with the same probability.

Thus, all three tests for each word have demonstrated different results. After comparative words nouns are more likely to be used in XX-XXI centuries, auxiliary words occur more often in the past and adjectives appear in the past and in the present with similar probability.

Test six

In this test, we check whether there is a dependence between the usage of some particular part of speech before the comparative words and time period of their usage.

H0: the part of speech before the comparative words appears in the texts before and after 1900 with the same probability

H1: the part of speech appear more often before the comparative words after 1900

In this case, we check the usage of nouns after them before and after 1900

Comparative words after nouns before XX

```
nouns <- data
nouns$Left_context_POS <- gsub("NOUN", 1, nouns$Left_context_POS)
nouns$Left_context_POS[nouns$Left_context_POS != 1] <- 0

old_texts <- nouns[nouns$Created<=1900,]
total <- nrow(old_texts)
nnouns <- nrow(old_texts[old_texts$Left_context_POS==1,])

binom.test(nnouns, total)
```

```
##
##  Exact binomial test
##
## data:  nnouns and total
## number of successes = 224, number of trials = 674, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2968409 0.3693237
## sample estimates:
## probability of success
##              0.3323442
```

Comparative words after nouns in XX-XXI

```
new_texts <- nouns[nouns$Created>=1900,]
total <- nrow(new_texts)
nnouns <- nrow(new_texts[new_texts$Left_context_POS==1,])

binom.test(nnouns, total)
```

```
##
##  Exact binomial test
##
## data:  nnouns and total
## number of successes = 396, number of trials = 895, p-value = 0.0006428
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4095953 0.4757008
## sample estimates:
## probability of success
##              0.4424581
```

The probability of noun usage before 1900 is 0.3323442, while after 1900 it is 0.4424581. So we can conclude that nouns are more likely to be used before comparative words in XX-XXI centuries.

Here we test the usage of auxiliary words before and after 1900

Comparative words after auxiliary words before XX

```
aux <- data
aux$Left_context_POS <- gsub("AUX", 1, aux$Left_context_POS)
aux$Left_context_POS[aux$Left_context_POS != 1] <- 0

old_texts <- aux[aux$Created<=1900,]
total <- nrow(old_texts)
naux <- nrow(old_texts[old_texts$Left_context_POS==1,])

binom.test(naux, total)
```

```
##
##  Exact binomial test
##
## data:  naux and total
## number of successes = 143, number of trials = 674, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1818748 0.2450018
## sample estimates:
## probability of success
##              0.2121662
```

Comparative words after auxiliary words in XX-XXI

```
new_texts <- aux[aux$Created>=1900,]
total <- nrow(new_texts)
naux <- nrow(new_texts[new_texts$Left_context_POS==1,])

binom.test(naux, total)
```

```
##
##  Exact binomial test
##
## data:  naux and total
## number of successes = 165, number of trials = 895, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1594664 0.2113456
## sample estimates:
## probability of success
##              0.1843575
```

The probability of auxiliary word usage before 1900 is 0.2121662, while after 1900 it is 0.1843575. In other words, auxiliary words occur before comparative words slightly more often in the past.

The same test has been performed for adjectives

Comparative words after adjectives before XX

```
adj <- data
adj$Left_context_POS <- gsub("ADJ", 1, adj$Left_context_POS)
adj$Left_context_POS[adj$Left_context_POS != 1] <- 0

old_texts <- adj[adj$Created<=1900,]
total <- nrow(old_texts)
nadj <- nrow(old_texts[old_texts$Left_context_POS==1,])

binom.test(nadj, total)
```

```
##
##  Exact binomial test
##
## data:  nadj and total
## number of successes = 63, number of trials = 674, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.07257385 0.11800652
## sample estimates:
## probability of success
##             0.09347181
```

Comparative words after adjectives in XX-XXI

```
new_texts <- adj[adj$Created>=1900,]
total <- nrow(new_texts)
nadj <- nrow(new_texts[new_texts$Left_context_POS==1,])

binom.test(nadj, total)
```

```
##
##  Exact binomial test
##
## data:  nadj and total
## number of successes = 71, number of trials = 895, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.06247423 0.09901291
## sample estimates:
## probability of success
##             0.07932961
```

The probability of adjective usage before 1900 is 0.09347181, while after 1900 it is 0.07932961. It means that adjectives appear before the comparative words a bit more often in the past.

Thus, nouns are more likely to be used before the comparative words in XX-XXI centuries, while auxiliary words and adjectives occur slightly more often in the past.

Chi-squared tests

In this test we are going to check the dependence between 2 variables: the historical period(before XX cent./XX-XXI cent.) and the usage of a particular function word

H0: the historical period and the usage of particular function words are independent variables

H1: There is a dependence between them

```
data <- data %>% mutate(Century = ifelse(Created >= 1900, "XX-XXI cent", "before XX cent."))

tab <- table(data$Center, data$Century)

prop.table(tab)
```

```
##
##           before XX cent. XX-XXI cent
##   а-ля         0.00600000  0.16066667
##   аки          0.09666667  0.07000000
##   будто        0.08533333  0.08133333
##   как          0.08000000  0.08666667
##   подобно      0.07266667  0.09400000
##   словно       0.06266667  0.10400000
```

```
chisq.test(data$Center, data$Century)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$Center and data$Century
## X-squared = 192.85, df = 5, p-value < 2.2e-16
```

P-value is less than 0.05, so we reject the null hypothesis about the independence of these two variables.

We also have conducted the chi-squared test for the usage of function words and functional styles.

```
chisq.test(data$Center, data$Type)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$Center and data$Type
## X-squared = 396.32, df = 20, p-value < 2.2e-16
```

And we have checked the dependence between the usage of function words and their context.

```
chisq.test(data$Center, data$Left_context_POS)
```

```
## Warning in chisq.test(data$Center, data$Left_context_POS): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$Center and data$Left_context_POS
## X-squared = 470.21, df = 45, p-value < 2.2e-16
```

Comparative words and left context

```
chisq.test(data$Center, data$Left_context_POS)
```

```
## Warning in chisq.test(data$Center, data$Left_context_POS): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$Center and data$Left_context_POS
## X-squared = 470.21, df = 45, p-value < 2.2e-16
```

In all these tests p-value<0.05 and it means that all of these variable pairs are dependent.

Linear Regression

In this experiment, using R tools we have gained the information regarding frequency of the comparative function words for each year. Then on the basis of the linear regression model, we try to predict the frequency of each word by the year.

H0: We can’t predict the frequency of each particular word by the year.

H1: Knowing the year, we can guess which word is more likely to be used.

```
years_distr <- t(table(data$Center, data$Created))
years_distr <- add_rownames(as.data.frame.matrix(years_distr))
```

```
## Warning: Deprecated, use tibble::rownames_to_column() instead.
```

```
colnames(years_distr) <-c("year", "ala", "aki", "budto", "kak", "podobno", "slovno")
years_distr$year <- as.numeric(years_distr$year)

years_distr
```

```
## # A tibble: 101 x 7
##     year   ala   aki budto   kak podobno slovno
##    <dbl> <int> <int> <int> <int>   <int>  <int>
##  1  1700     0    10     0     0       0      0
##  2  1710     0     5     0     9       0      0
##  3  1715     0     0     0     1       0      0
##  4  1720     0     7     0    17       1      0
##  5  1740     0     0     0     2       0      0
##  6  1750     0     5     1     1       4      0
##  7  1760     0    12     0     4       3      0
##  8  1770     0     4     2     3       0      0
##  9  1780     0     2     1    12       1      0
## 10  1785     0     0     1     0       1      0
## # ... with 91 more rows
```

```
fit1 <- lm(ala ~ year, data = years_distr)
summary(fit1)
```

```
##
## Call:
## lm(formula = ala ~ year, data = years_distr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -6.238  -3.606  -1.983   0.211 118.327
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -60.81474   31.05621  -1.958   0.0530 .
## year          0.03324    0.01630   2.040   0.0441 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.49 on 99 degrees of freedom
## Multiple R-squared:  0.04032,    Adjusted R-squared:  0.03063
## F-statistic:  4.16 on 1 and 99 DF,  p-value: 0.04406
```

```
fit2 <- lm(aki ~ year, data = years_distr)
summary(fit2)
```

```
##
## Call:
## lm(formula = aki ~ year, data = years_distr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.380 -2.292 -1.777  0.175 40.495
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.678080  14.665241   1.478    0.143
## year        -0.010087   0.007697  -1.310    0.193
##
## Residual standard error: 5.898 on 99 degrees of freedom
## Multiple R-squared:  0.01705,    Adjusted R-squared:  0.007122
## F-statistic: 1.717 on 1 and 99 DF,  p-value: 0.1931
```

```
fit3 <- lm(budto ~ year, data = years_distr)
summary(fit3)
```

```
##
## Call:
## lm(formula = budto ~ year, data = years_distr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9654 -2.4628 -2.3089 -0.7249 31.3472
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.053762  13.587167   0.519    0.605
## year        -0.002405   0.007131  -0.337    0.737
##
## Residual standard error: 5.464 on 99 degrees of freedom
## Multiple R-squared:  0.001148,   Adjusted R-squared:  -0.008942
## F-statistic: 0.1137 on 1 and 99 DF,  p-value: 0.7366
```

```
fit4 <- lm(kak ~ year, data = years_distr)
summary(fit4)
```

```
##
## Call:
## lm(formula = kak ~ year, data = years_distr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6300 -2.5962 -1.5436 -0.3082 24.9179
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.94429   12.53695   2.548   0.0124 *
## year        -0.01548    0.00658  -2.352   0.0206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.042 on 99 degrees of freedom
## Multiple R-squared:  0.05294,    Adjusted R-squared:  0.04337
## F-statistic: 5.534 on 1 and 99 DF,  p-value: 0.02063
```

```
fit5 <- lm(podobno ~ year, data = years_distr)
summary(fit5)
```

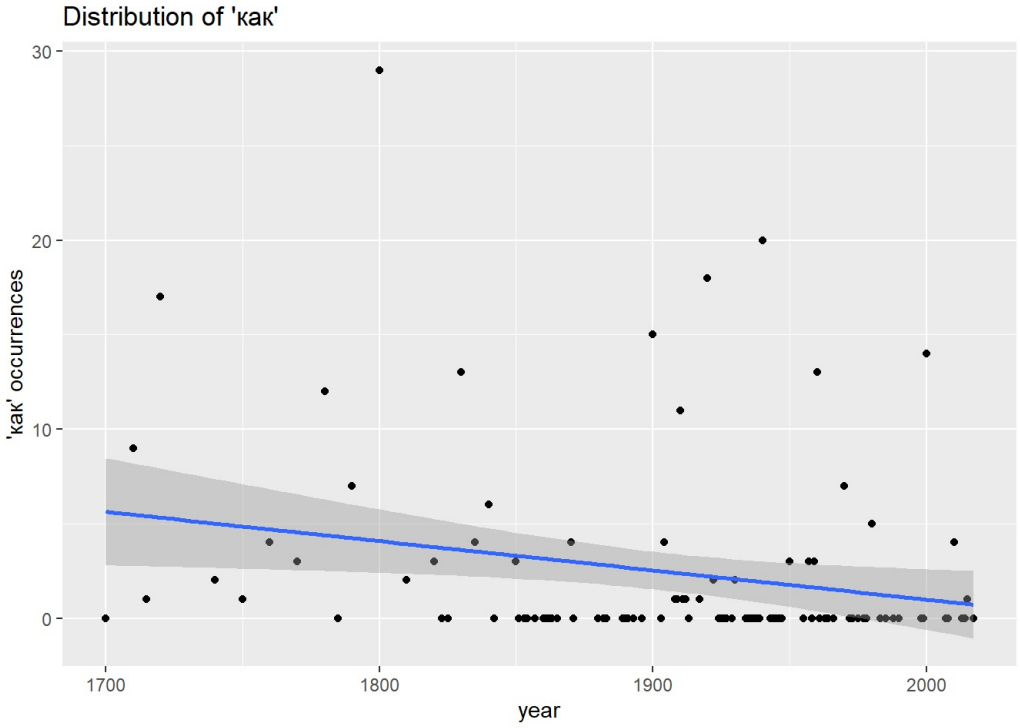
```
##
## Call:
## lm(formula = podobno ~ year, data = years_distr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7335 -2.4471 -2.3407  0.5237 19.5453
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.887677  11.414966   0.428    0.669
## year        -0.001267   0.005991  -0.212    0.833
##
## Residual standard error: 4.591 on 99 degrees of freedom
## Multiple R-squared:  0.0004517,  Adjusted R-squared:  -0.009645
## F-statistic: 0.04474 on 1 and 99 DF,  p-value: 0.8329
```

```
fit6 <- lm(slovno ~ year, data = years_distr)
summary(fit6)
```

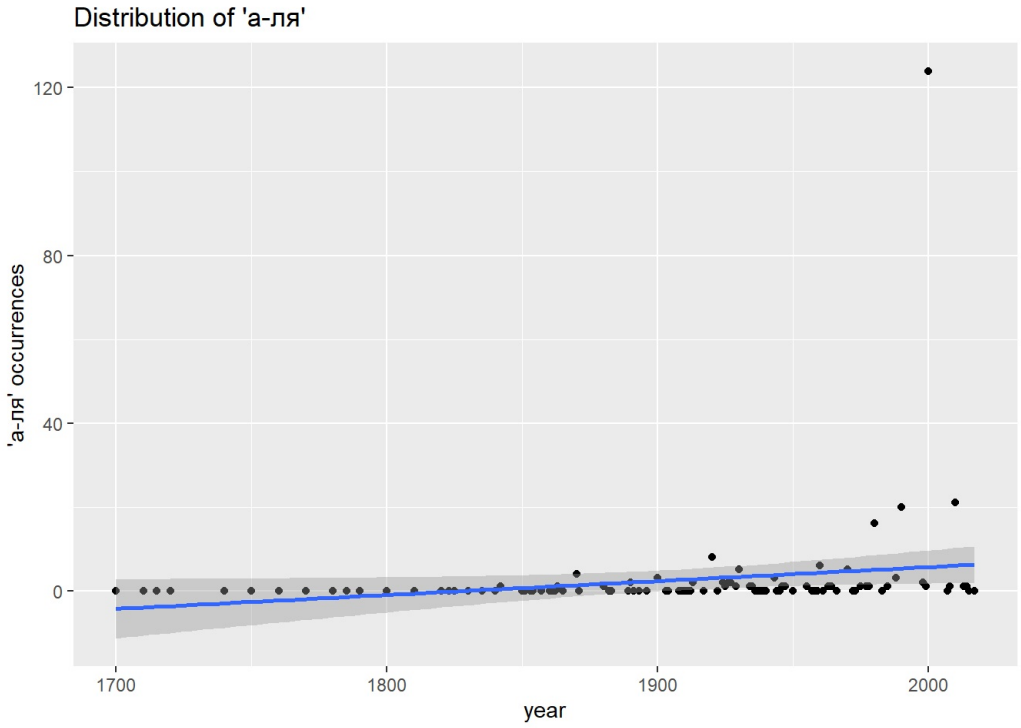
```
##
## Call:
## lm(formula = slovno ~ year, data = years_distr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8538 -2.5814 -2.1218 -0.7347 23.4696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.006891  13.870567  -0.289    0.773
## year         0.003405   0.007280   0.468    0.641
##
## Residual standard error: 5.578 on 99 degrees of freedom
## Multiple R-squared:  0.002205,   Adjusted R-squared:  -0.007874
## F-statistic: 0.2187 on 1 and 99 DF,  p-value: 0.641
```

We reject the null hypothesis only for such words as “а-ля” and “как”. The experiments with the rest of the words(“будто”, “словно”, “аки”, “подобно”) have demonstrated p-value>0.05 so we should accept the null hypothesis for them.

```
p <- ggplot(data = years_distr, aes(x = year, y = kak)) +
  geom_point() +
  labs(y = "'как' occurrences",
       x = "year",
       title = "Distribution of 'как'") +
  geom_smooth(method=lm)
p
```




```
p <- ggplot(data = years_distr, aes(x = year, y = ala)) +
  geom_point() +
  labs(x = "year",
       y = "'а-ля' occurrences",
       title = "Distribution of 'а-ля'") +
  geom_smooth(method=lm)
p
```



Conclusions

Having considered the results of the research, we can make the following linguistic conclusions.

Hypotheses about the usage of functional styles:

- “как” and “будто” are not evenly distributed in all styles
- “будто”, “словно” and “аки” are more frequent in belles-lettres texts
- “а-ля” has a tendency to appear in the colloquial style

The words “как” and “будто” were very likely to appear in official documents (especially the word “как”). This fact can prove the stylistic neutrality of the word “как”. However, their usage in scientific texts for some reason was very low. Probably, the data sample was not balanced enough for academic texts, as there were not many texts of this style. Only in publicistic texts, these two words could be used with probability ⅓. We have to admit that stylistic neutrality does not correlate with even distribution as different styles themselves are not neutral (e.g. in belles-lettres texts the authors are more likely to use expressions which are not typical, while official documents have to be very neutral). The hypothesis regarding “будто”, “словно” and “аки” usage in belles-lettres texts has been proved. But each of these words should be tested separately as well. The word “а-ля” indeed is very likely to appear in texts of colloquial style. The hypothesis actually was based on the observation that some people use it too often as a filler word.

Hypotheses concerning the date of the creation of the documents:

- “аки” is more typical for old texts written before the 1900s
- nouns in the right context are used more often in the text after the 1900s
- auxiliary words in the right context are used more in the past
- adjectives in the right context have almost the same probability of appearance in both old and modern documents
- nouns in the left context are used more in modern texts
- auxiliary words and adjectives in the left context occur more often in old texts
- there are dependencies between the usage of function words and a historical period, between the usage of function words and functional styles as well as between the usage of function words and their left and right context
- the usage of “а-ля” and “как” can’t be predicted from the perspective of the year
- the frequencies of the words “будто”, “словно”, “аки” and “подобно” can be predicted by the year

The result regarding the usage of “аки” in the past seems quite obvious as many dictionaries refer to this word as an archaism. The fact that auxiliary words were used more often in the past can be explained by the simplification of the language. Probably, earlier Russian speakers tended to use more complex syntactic structures. The usage of nouns after and before comparative words in modern times can be connected with the semantics of nouns and may be observed in greater detail. Probably, we are more likely to compare objects (not their features or actions) than our ancestors.

So, all of these results raise many other questions and suggestions to be tested in future researches.