Barbarisms in the Russian web-discourse

Kirill Milintsevich, Ivan Rodin
2017

Introduction

Very often we notice that Russian-speakers use English words that have their equivalent in Russian in their everyday discourse. In this research, we want to know if the size of the city that a speaker resides in influences their use of the English barbarisms. Our hypothesis is that people from bigger cities use more barbarisms than people from the smaller ones. It is based on the assumption that bigger cities are usually more urbanized hence people living there have more contact with the English language. Also, we studied the influence of the gender on the use of the English barbarisms.

The data

We gathered and automatically analysed approx. 2.5 mln. commentaries from VK.com social network. With the help of precompiled dictionary of the barbarisms, we tagged each user with 1 if they used a barbarism in their comment and with 0 if otherwise. Our data contains only the users that have a city mentioned in their profiles.

```
library(tidyverse)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.3.3
```

```
library(dplyr)
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.3.3
```

```
barbar <- read.csv("https://goo.gl/ggeuW5")
```

Manipulating the data

Let's divide our cities into small(population < 100000) and big(population > 100000). Also, we'd like to throw out very small cities, with the population smaller than 10000 inhabitants, due to the small amount of commentators from those ones and the fact that they cannot be called "cities" (A.A.Perederiy. City Classification and Typology. Link: http://www.mstu.edu.ru/science/conferences/11ntk/materials/section8/section8_35.html).

Statistics

The time has come to look at some numbers. First, we'd like to look at the means and SDs of the frequencies of the barbarisms' usage in small and big cities.

```
group_by(barbar_by_status, status) %>%
  summarise(
    count=n(),
    mean = mean(barbar_frequency, na.rm=TRUE),
    sd = sd(barbar_frequency, na.rm=TRUE)
  )
```

```
## # A tibble: 2 Γ— 4
##   status count       mean         sd
##    <chr> <int>      <dbl>      <dbl>
## 1    Big   159 0.06946605 0.04508912
## 2  Small   641 0.15896334 0.09912085
```

Surprisingly, users from small cities tend to use almost twice as more barbarisms that users from big cities.

We also want to conduct a Shapiro test to see whether our frequencies are normally distributed:

```
with(barbar_by_status, shapiro.test(barbar_frequency[status == "Small"
```

```
##
##  Shapiro-Wilk normality test
##
## data:  barbar_frequency[status == "Small"]
## W = 0.81651, p-value < 2.2e-16
```

```
with(barbar_by_status, shapiro.test(barbar_frequency[status == "Big"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  barbar_frequency[status == "Big"]
## W = 0.37315, p-value < 2.2e-16
```

We got the p-value way less that our alpha = 0.05, which means that frequencies are not normally distributed. In that case, let's use unpaired two-samples Wilcoxon test to find out whether the means are significantly different from each other.

```
wilcox.test(barbar_frequency ~ status, data = barbar_by_status, exact =
```

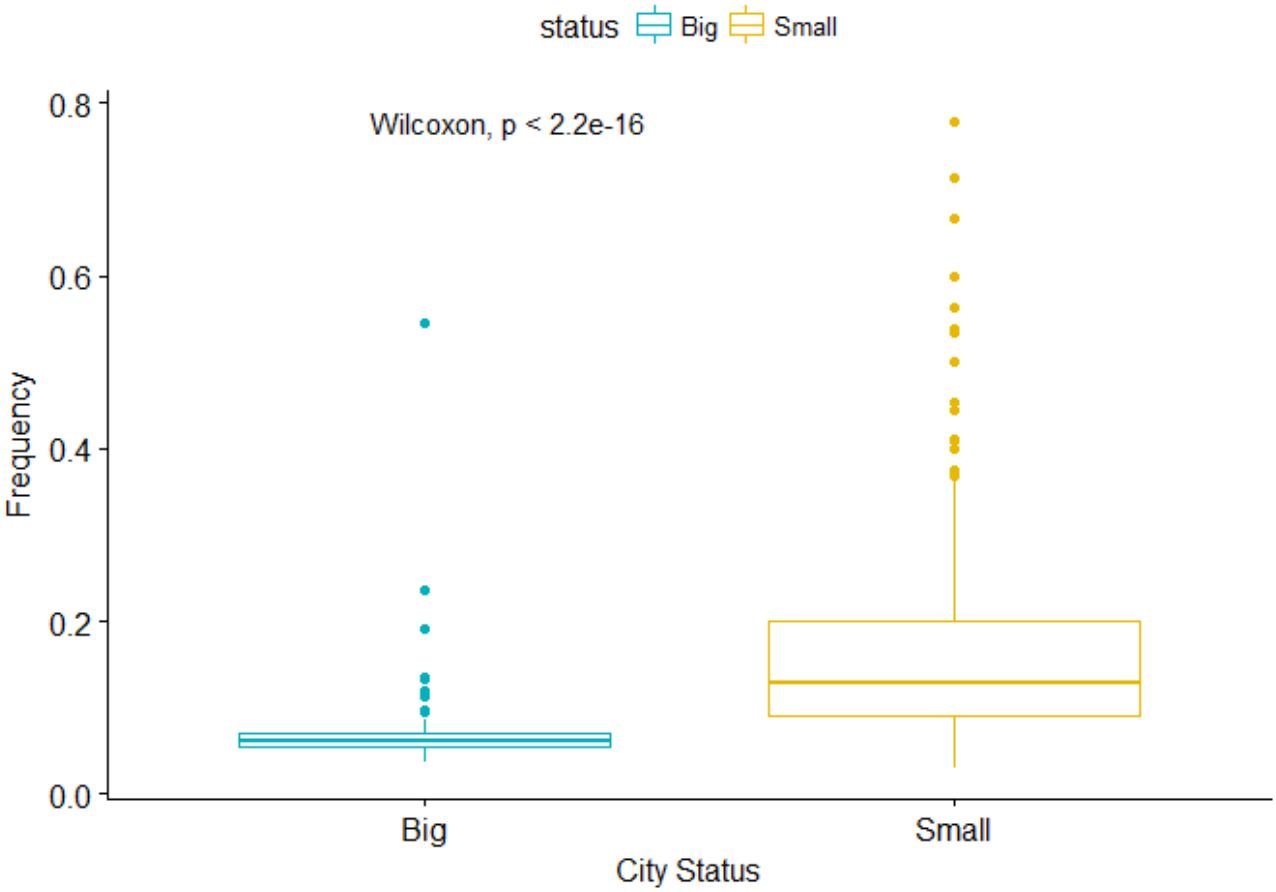
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  barbar_frequency by status
## W = 9996.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

The p-value of the test is < 2.2e-16, which is less than the significance level alpha =

0.05. We can conclude that small cities' median barbarism frequency is significantly different from big cities' median frequency with a p-value < 2.2e-16.

Plotting the data:

```
p <- ggboxplot(barbar_by_status, x = "status", y = "barbar_frequency",
               color = "status", palette = c("#00AFBB", "#E7B800"),
               ylab = "Frequency", xlab = "City Status")
p + stat_compare_means()
```



Clean our data

Previously, we took the data from every city, that got into our scope, i.e. that a city with only one registered user might be present there, and if that user left a comment with a barbarism in it, his city would get 1.0 probability of it's inhabitant leaving a fancy comment. That's, obviously, not very good. Let's try to conduct the same analysis on the filtered data, where only cities with at least 200 registered users were allowed in. Now, we also divide our cities into more categories, which are "Small",

"Big", "Large", "Biggest", and "Million".

```
barbar_filtered <- read.csv("https://goo.gl/TekJTH")

barbar_small_f <- barbar_filtered[barbar_filtered$population < 100000 &
barbar_big_f <- barbar_filtered[barbar_filtered$population < 250000 & b
barbar_large_f <- barbar_filtered[barbar_filtered$population < 500000 &
barbar_biggest_f <- barbar_filtered[barbar_filtered$population < 10000(
barbar_million_f <- barbar_filtered[barbar_filtered$population > 10000(

barbar_small_f$status <- rep("Small", length(barbar_small_f[,1]))
barbar_big_f$status <- rep("Big", length(barbar_big_f[,1]))
barbar_large_f$status <- rep("Large", length(barbar_large_f[,1]))
barbar_biggest_f$status <- rep("Biggest", length(barbar_biggest_f[,1])
barbar_million_f$status <- rep("Million", length(barbar_million_f[,1])

barbar_by_status_detailed <- rbind(barbar_small_f[,c("barbar_frequency'
barbar_by_status_detailed$status <- ordered(barbar_by_status_detailed$s
                        levels = c("Small", "Big", "Large", "Biggest"
```

Now let's look at their means again.

```
group_by(barbar_by_status_detailed, status) %>%
  summarise(
    count=n(),
    mean = mean(barbar_frequency, na.rm=TRUE),
    sd = sd(barbar_frequency, na.rm=TRUE)
  )
```

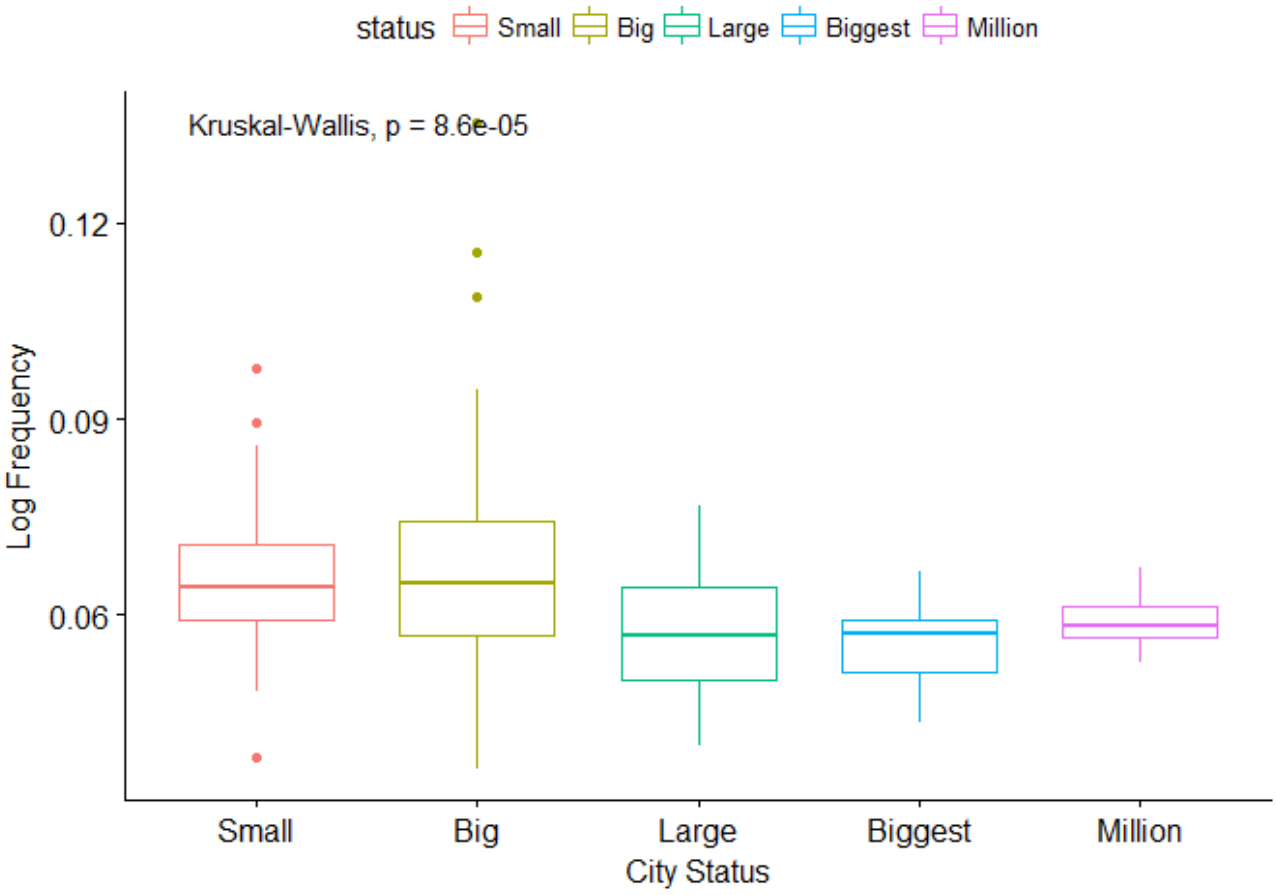
```
## # A tibble: 5 Γ— 4
##    status count       mean          sd
##     <ord> <int>      <dbl>       <dbl>
## 1   Small    39 0.06577130 0.012087895
## 2     Big    74 0.06694744 0.017007460
## 3   Large    39 0.05762430 0.009003499
## 4 Biggest    20 0.05571422 0.006576426
## 5 Million    13 0.05867399 0.004343862
```

Now, we can see, that Small and Big cities are almost equal and only a little bit more that Large, Biggest and Million cities. Also, the deviation reduced a lot, meaning that

our data is more "compact" now.

Visualizing our new data:

```
p <- ggboxplot(barbar_by_status_detailed, x = "status", y = "barbar_fr
               color = "status",
               ylab = "Log Frequency", xlab = "City Status")
p + stat_compare_means()
```



No ANOVA here

To find out whether we can use the ANOVA test to check the significance of differences between our means. Thus, our data have to meet the assumptions homogeneity and normality.

We use Levene's Test to check the homogeneity of variances.

```
leveneTest(barbar_frequency ~ status, data = barbar_by_status_detailed
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   4  4.9468 0.0008311 ***
##       180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value is less than 0.05, which means that we cannot assume the homogeneity of variances in the different city size groups.

To check our data for normality, we use the Shapiro-Wilk test on the ANOVA residuals.

```
res.aov <- aov(barbar_frequency ~ status, data = barbar_by_status_deta
aov_residuals <- residuals(object = res.aov)
shapiro.test(x = aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.92732, p-value = 5.563e-08
```

Again, our value is way less that our alpha = 0.05, so we cannot assume the normality of the data.

In this case, we cannot use the ANOVA test because our data failed tests for homogeneity and normality. As an alternative, we'll use Kruskal-Wallis rank sum test.

```
kruskal.test(barbar_frequency ~ status, data = barbar_by_status_detail
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  barbar_frequency by status
## Kruskal-Wallis chi-squared = 23.843, df = 4, p-value = 8.586e-05
```

Looking at the p-value, we can conclude that there is significant differences between the cities of different sizes. However, we have 5 categories, so we might want to look what categories exactly are significantly different. To do so, we use the pair-wise Wilcoxon test.

```
pairwise.wilcox.test(barbar_by_status_detailed$barbar_frequency, barba
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  barbar_by_status_detailed$barbar_frequency and barbar_by_sta
##
##         Small  Big    Large  Biggest
## Big     0.9591 -      -      -
## Large   0.0036 0.0036 -      -
## Biggest 0.0021 0.0036 0.6979 -
## Million 0.0312 0.0706 0.5643 0.3568
##
## P value adjustment method: BH
```

From this, we can see that there's no significant difference between Small and Big cities and Large, Biggest and Million cities. What we are left with is that the difference is significant between Small and Large, Small and Biggest and Big and Large and Big and Biggest cities.

Adding the rival features

Previously, we found out that people from the cities with the population less than 250,000 people are slightly more prone to using barbarisms in their commentaries. It is interesting to know whether there is any other factor that influences the use of barbarisms. We decided to look at the user's gender.

We have re-analysed our data and added a new feature of the user - their gender. Our hypothesis is that female users tend to use more barbarisms than male ones. We were rather categorical in determining if a user uses barbarisms or not. In our data if a user was detected in using a barbarism at least once, they are considered to be using barbarisms. The data are available here: https://goo.gl/938nJg (Warning! The

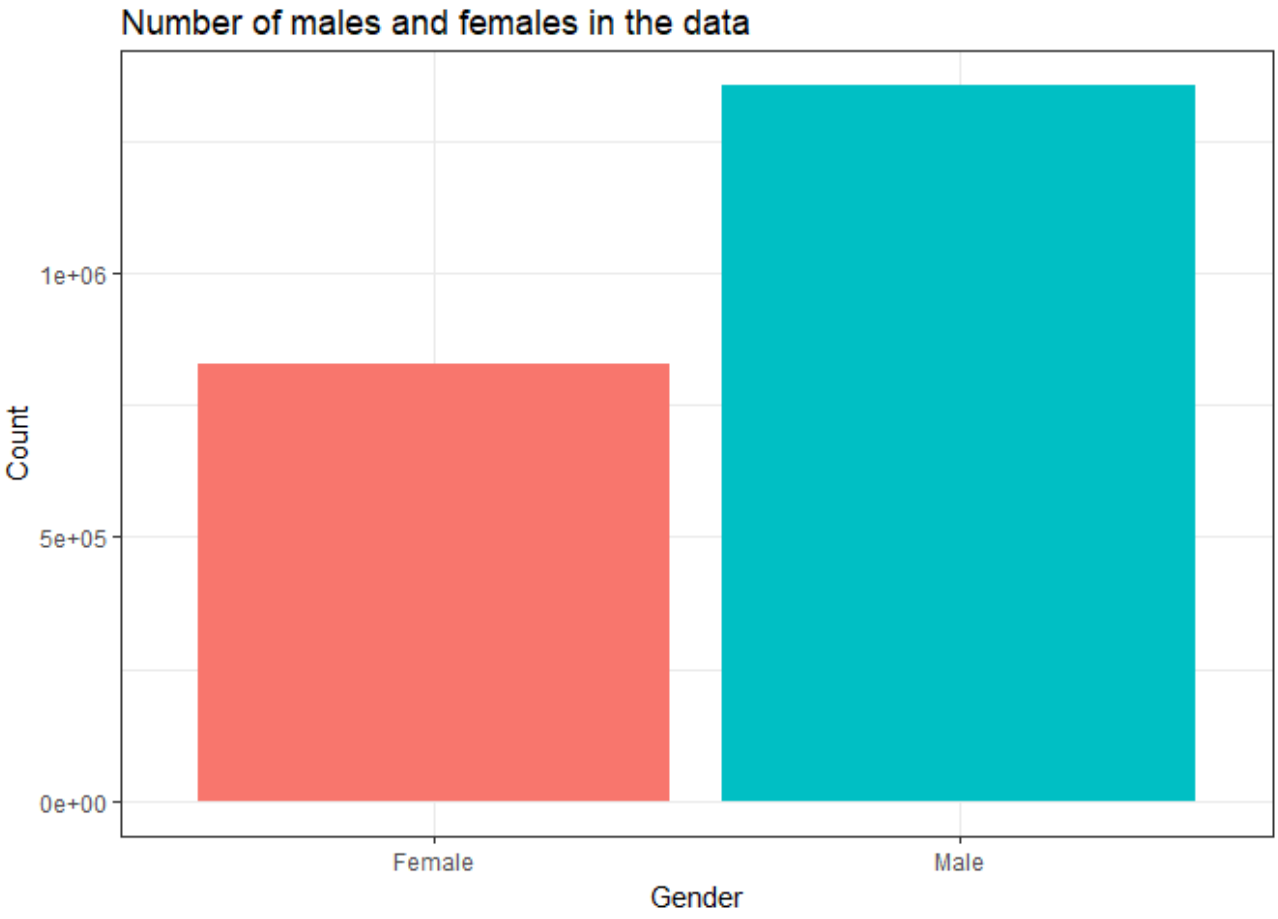
data are more than 30 Mbs).

Loading the data:

```
gender <- read.csv("https://goo.gl/938nJg")
gender$sex_factor <- factor(gender$sex,
                            levels = c(1, 2),
                            labels = c("Female", "Male"))
```

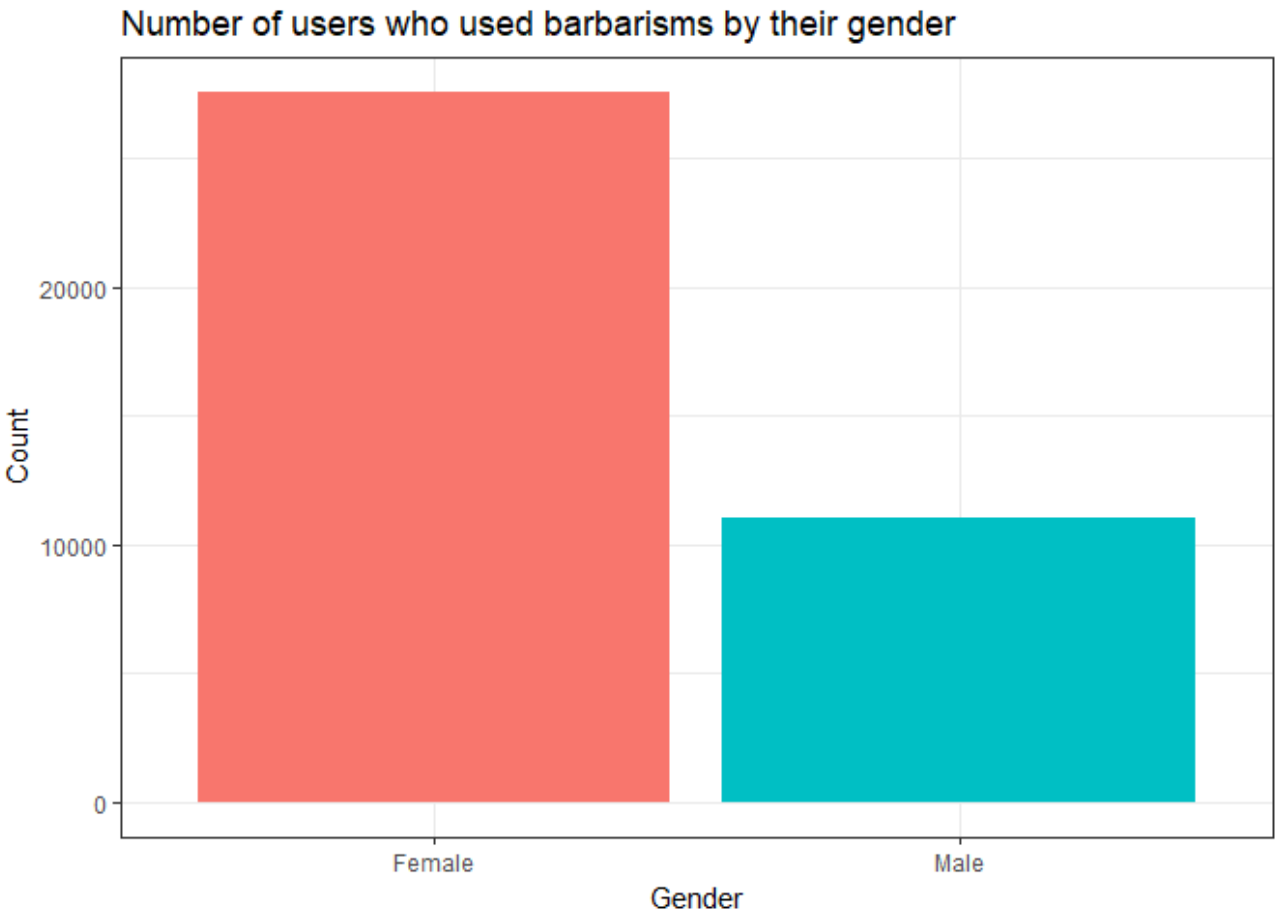
Let's look at how the genders are distributed:

```
gender %>%
  ggplot(aes(sex_factor, fill = sex_factor)) +
  geom_bar() +
  labs(x = "Gender", y = "Count", title = "Number of males and females
  theme_bw() +
  theme(legend.position="none")
```



Now, let's see how many of them used barbarisms:

```
subset(gender, hasBar == 1) %>%
  ggplot(aes(sex_factor, fill = sex_factor)) +
  geom_bar() +
  labs(x = "Gender", y = "Count", title = "Number of users who used ba
  theme_bw() +
  theme(legend.position="none")
```



As we can see, more than twice as many female users used barbarisms in their comments. However, we still have to test it for significance. This time, we use the Chi-squared test. But we have to construct the matrix first:

```
female_bar <- nrow(subset(gender, hasBar == 1 & sex == 1))
male_bar <- nrow(subset(gender, hasBar == 1 & sex == 2))
female_noBar <- nrow(subset(gender, hasBar == 0 & sex == 1))
male_noBar <- nrow(subset(gender, hasBar == 0 & sex == 2))
gender_matrix <- matrix(c(female_bar, male_bar, female_noBar, male_noB
                          nrow = 2, byrow = TRUE)
rownames(gender_matrix) <- c("hasBar", "noBar")
colnames(gender_matrix) <- c("female", "male")
gender_matrix
```

```
##        female    male
## hasBar  27564   11025
## noBar  799313 1345046
```

The Chi-squared test:

```
chisq.test(gender_matrix)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender_matrix
## X-squared = 18791, df = 1, p-value < 2.2e-16
```

As we can see, p-value is less than 0.05, so we can make the conclusion that there is a significant difference between the two groups.

Conclusion

Thus we are left with two groups that are significantly different: Small-Big (pop. less than 250,000) and Large-Biggest (pop. between 250,000 and 1,000,000). Cities with more than million population kind of got out of our way. We suppose that the explanation to this may lie in a fact that many people from smaller cities move to the bigger ones, thus bringing their linguistic habits with them and mixing our samples.

However, even that the difference is present, it is very small (less than 1%). We should probably look at other sociolinguistic factors like age, gender or education to get some bigger difference in the usage of Anglophone Barbarisms in Russian web-discourse.

Moreover, we analysed one more variable which is gender and found out that female users tend to use more barbarisms than male ones.