Capitalization patterns on Russian Social Media

Nikolaeva Anna

21 June 2019

This research is aimed to examine how capitalization patterns on Russian social media vary with respect to three explanatory variables: the gender of the user, the age and the city where the user lives. I was inspired by another paper on capitalization for English, called “Social and Emotional Correlates of capitalization on Twitter”.

1. Hypotheses

I plan to test several different hypotheses:

- whether the fraction of capitalized words in a comment on Russian social media depend on the gender of the user;
- whether the fraction of capitalized words in a comment on Russian social media depend on the city where the user lives;
- whether the fraction of capitalized words in a comment on Russian social media depend on the age of a user.

Moreover, depending on the results of these tests I would like to explore whether it is possible to build a regression model that can predict the fraction of capitalized words in a comment, searching for possible interaction effects as well.

It would be interesting to compare some of these conclusions with the results of the previous research on the English capitalization patterns (the link above). Especially, S. Chan and A. Fyshe found out that capitalization on Twitter made by English-speaking Internet users does vary with respect to gender: female use words with uppercase characters more than male. There were more strengthened effects when the researchers considered only meaningfully capitalized tokens (which are not usually capitalized). To perform that, the authors also calculated each specific token’s probability of being capitalized.

Speaking about other expected results, I tend to think that young people use a uppercase on Runet more than elderly people.

2. Data Collection Method

Inspired by the mentioned study, I will conduct the experiments on two datasets: on the bigger one, where a word is regarded as a capitalized token even if it is an acronym, and the smaller one, which is the part of the bigger one, where only meaningfully capitalized tokens are taken into consideration.

There are no Russian corpora suitable for this study containing social metadata about users, so I crawled a small corpus myself from the online Russian social media Vkontakte using Python3 and VK API. You may see the way I collected and preprocessed the data on GitHub link1, link2, link3.

There are only two sources of the data: the https://vk.com/vdud wall page and the https://vk.com/echomsk wall page.

The main steps of collecting and preprocessing data from these two pages were as follows:

- Get all post ids from the page
- Get all the comments to these posts with caps lock words (if possible) including nested comments
- Get the user id of every comment and collect the user data (sex, birthday date, city)
- Save only those entries where “sex” is defined (minimal condition, other fields may be NA)

- Save only those entries where the user id (“from_id”) is unique The original datasets after these steps are: Vdud’ dataset, Echo Msk Dataset.

6) Concatenate the two datasets into one
7) Clean the text of a comment from hashtags and links on other users’ and clubs’ ids
8) Add columns which contain a number of tokens in a comment, a number of capitalized Russian words (if word length > 2 characters), a column with capitalized words themselves separated by space
9) Remove entries if they contain no caps lock words This “bigger” dataset can be found at link.
10) In order to create a smaller dataset with meaningfully capitalized tokens, I have computed for every capitalized word a ratio “a number of a word being capitalized in the word list divided by sum of a number of this word being capitalized and not being capitalized in the list”:

```
Count(upper)/(Count(upper) + Count(nonupper))
```

This metrics was proposed in a study mentioned before. If this ratio was smaller than 0.1, it meant that a certain word was capitalized only in less than 10% of cases in a corpus, and the authors considered such tokens as meaningfully capitalized, capitalized on purpose. I wanted to do the same in my research, but, unfortunately, my corpus is appeared to be rather small, containing only around 186 000 tokens, so I had to take a much higher threshold = 0.85 in order to remove acronyms from the capitalized words list and had to physically look through the words with a small frequency in my corpus in order to decide if a token is capitalized on purpose or not.
11) Using the meaningfully capitalized tokens list, I filtered only those dataframe entries which had these words and then added same kinds of columns as in 8. This smaller version of the dataset can be found at link.

The lists of usually capitalized words and words capitalized on purpose can be found at meaningful capitalization, usual capitalization.

3. Data Decription

The result datasets (the bigger and the smaller ones) contain the following columns: * `from_id` — the unique user id * `post_id` — the id of the post on the wall * `comment_text` — the text of the comment under that post * `comment_id` — was needed to get a nested comment (the answer to some comment) * `first_name` — first name of the user * `last_name` — last name of the user * `sex` * `bdate` — birthday date of the user * `city` — the city where the user lives * `n_tokens` — number of tokens in a comment * `n_caps` — number of words capitalized * `caps` — capitalized words in a comment separated by space

4.Research Design

Variables of interest:

- 1 dependant numeric variable — the fraction of capitalized words in a comment.
- 3 explanatory factor variables — gender (2 levels), age (5 levels, will be created later), city (2 levels, will be created later).

The fraction of capitalized words is counted as n_caps / n_tokens.

The null hypothesis is that there is no connection between the fraction of capitalized words and other variables (one by one). I have chosen non-parametric criteria for testing this hypothesis - Wilcoxon rank sum test with continuity correction for pair tests (gender and city) + bootstrap as groups are unbalanced and Kruskal-Wallis rank sum test for checking the dependency of the dependant variable on the age of the user.

Formally in case of Wicoxon test we check if the differences in distributions of capitalization fractions are caused by a shift. If there is no shift, we cannot reject the null hypothesis of equality of the two distributions. We assume here, that the distribution of fractions of capitalized words is not Gaussian (we will see that later) and that the variance of the two samples is equal and unknown.

In case of Kruskal-Wallis rank sum test we formally check the null hypothesis that the mean ranks of all the groups are equal. Again we assume that the distribution of the samples is not normal (that is why we do not apply the ANOVA method). The advantage of Kuskal-Wallis criteria is that it can be used with unbalanced group sizes.

I have decided not to use ANOVA, as the researches of the mentioned paper did, because they took an absolute number of capitalized tokens in a tweet, but in my case the number of max words in a comment is not defined and it seems more reasonable to consider only relative numbers of words capitalized. Also there are many outlier observations for all the variables (for example, 1902 year of birth; a fraction of capitalized tokens can be equal to 1 but cannot be 0, as we do not include observations with no capslock words), and applying parametric criterion is very risky.

In case of regression we assume that we can predict the value of the dependant variable by explanatory variables.

Here you can find my code for the original dataset (the bigger one), which contains acronyms, after that I will analyze the smaller dataset.

5. Wilcoxon tests

5.0 Data loading

```
Sys.setlocale(category="LC_ALL", locale = "ru_RU")
```

```
## [1] "ru_RU/ru_RU/ru_RU/C/ru_RU/C"
```

```
if (!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(
  tidyverse,
  plyr,
  ggplot2,
  ggthemes,
  lme4,
  sjstats,
  stringr
)
theme_set(theme_bw())

cap <- read.csv("https://raw.githubusercontent.com/annnyway/capitalization/master/capitalization.csv")
```

5.1 Data description

```
cap <- cap[!duplicated(cap$from_id), ] # remove entries with same user_ids
# create a column with fraction of tokens capitalized
cap <- cap %>%
  mutate(caps_frac = n_caps/n_tokens)
psych::describe(cap$caps_frac) # descriptive statistics of this column
```

```
##    vars    n mean   sd median trimmed  mad min max range skew kurtosis se
## X1    1 4857 0.19 0.28   0.07    0.12 0.08   0   1     1 1.99     2.77  0
```

We can see that we have 4857 independent observations, the fraction mean is 0.19, the standart deviation is 0.28.

5.2 Preparations for the first Wilcoxon test (fraction of cap. words ~ gender of the user).

```
cap$sex <- factor(cap$sex) # convert numerical variable "sex" into factor
cap$sex <- revalue(cap$sex, c("1"="female", "2"="male")) # rename factor levels
psych::describeBy(cap$caps_frac,cap$sex)
```

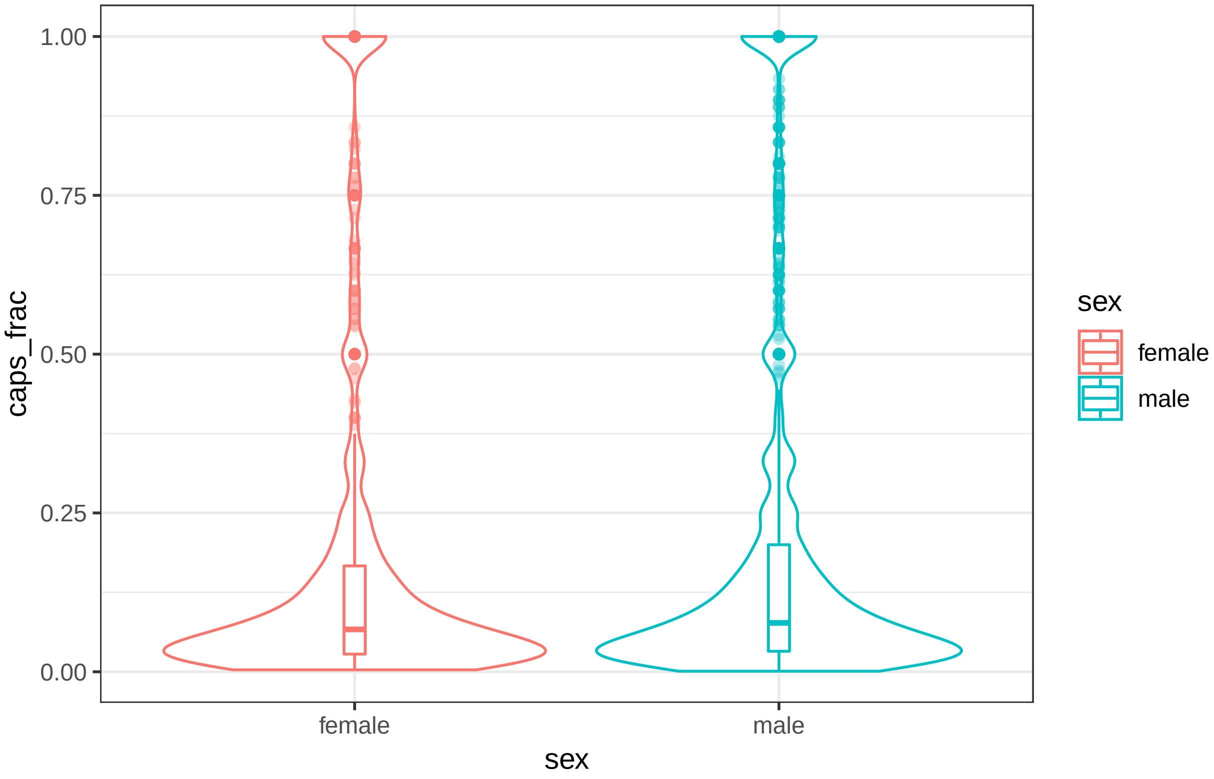
```
##
##  Descriptive statistics by group
## group: female
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 923 0.18 0.28   0.07    0.11 0.07   0   1     1 2.06     3.02 0.01
## --------------------------------------------------------
## group: male
##    vars    n mean   sd median trimmed  mad min max range skew kurtosis se
## X1    1 3934  0.2 0.28   0.08    0.13 0.08   0   1     1 1.98     2.71  0
```

```
bartlett.test(cap$caps_frac ~ cap$sex) # check homogeneity of variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  cap$caps_frac by cap$sex
## Bartlett's K-squared = 0.40455, df = 1, p-value = 0.5247
```

We can see that the female and male groups are unbalanced - 923 vs 3934 observations. We see that we have a slightly different means for capitalization fraction - 0.18 for female vs 0.2 for male. However the standart deviation and medians are almost close. The Bartlett's test of homogeneity of variances shows that the samples' variances are equal. Let us visualize the distribution:

```
ggplot(cap,aes(x = sex, y = caps_frac, color = sex)) +
  geom_violin() +
  geom_boxplot(width=0.05, alpha = .3) +
  labs(
    caption = paste0('n female = ',length(cap$sex[cap$sex == 'female']),',
                     n male = ',length(cap$sex[cap$sex == 'male']))
  )
```



n female = 923,
n male = 3934

We can see, that the distrubution of capitalization fraction is not normal, there are a lot of outliers including caps_frac = 0. Now let us check if there is a significant diffrence between male and female in capitalization patterns.

5.3 Wilcoxon test + Bootstrap from female and male samples

The bootstrap procedure is as follows: we choose 100 elements from male and female samples 1000 times, run Wilcoxon test 1000 times and collect p-values. Then we count the percent of p-values < 0.05.

```r
# doing bootstrap:

pvals <- c()
for(i in 1:1000){
  male_sample <- sample(cap$caps_frac[cap$sex == "male"], 100)
  female_sample <- sample(cap$caps_frac[cap$sex == "female"], 100)
  pvals <- c(wilcox.test(male_sample, female_sample)$p.value,pvals)
}

cat(length(pvals[pvals < 0.05])/1000*100,"%")
```

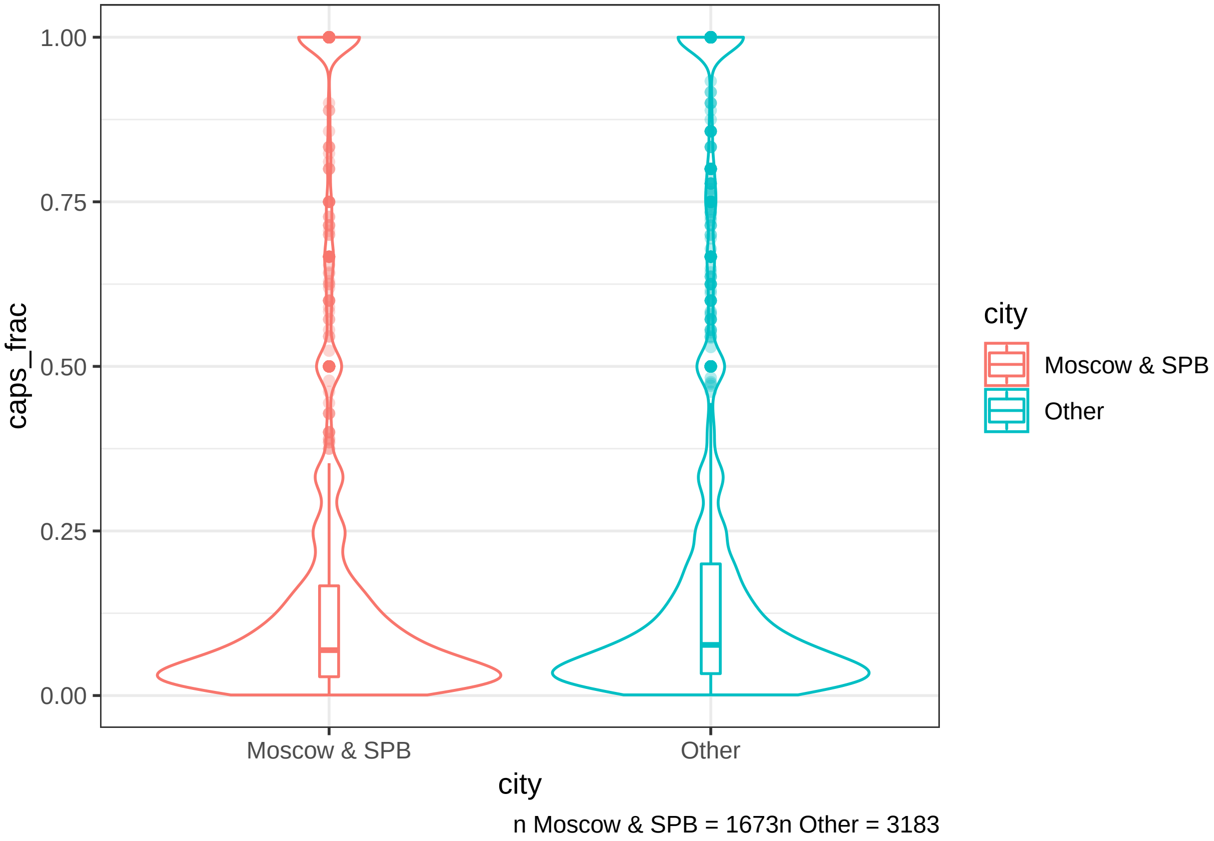
```
## 9.7 %
```

The percent of p-values < 0.05 is far less than 50%, so we cannot regect the null hypothesis about independence of gender and the proportion of capitalized words in a comment.

5.4 Preparations for the second Wilcoxon test (fraction of cap. words ~ city where the user lives).

```
cap_city <- filter(cap, city != '')
# rename factor levels in order to get two groups to compare

cap_city$city <- ifelse(cap_city$city == "Moscow" | cap_city$city == "Saint Petersburg",
                   "Moscow & SPB",
                   "Other")
ggplot(cap_city,aes(x = city, y = caps_frac, color=city)) +
  geom_violin() +
  geom_boxplot(width=0.05, alpha = .3) +
  labs(caption = paste0('n Moscow & SPB = ',
                        length(cap_city$city[cap_city$city == 'Moscow & SPB']),
                        'n Other = ',
                        length(cap_city$city[cap_city$city == 'Other'])))
```



```
psych::describeBy(cap_city$caps_frac, cap_city$city)
```

```
##
##  Descriptive statistics by group
## group: Moscow & SPB
##    vars    n mean   sd median trimmed  mad min max range skew kurtosis
## X1    1 1673 0.18 0.27   0.07    0.11 0.07   0   1     1 2.16     3.55
##      se
## X1 0.01
## --------------------------------------------------------
## group: Other
##    vars    n mean   sd median trimmed  mad min max range skew kurtosis
## X1    1 3183  0.2 0.29   0.08    0.13 0.08   0   1     1 1.92     2.41
##      se
## X1 0.01
```

We see that the distribution is almost identical.

5.5 Wilcoxon tests + Bootstrap from different "city" groups

```
pvals <- c()
for(i in 1:1000){
  moscow_sample <- sample(cap_city$caps_frac[cap_city$city == "Moscow & SPB"], 100)
  other_sample <- sample(cap_city$caps_frac[cap_city$city == "Other"], 100)
  pvals <- c(wilcox.test(moscow_sample, other_sample)$p.value, pvals)
}
cat(length(pvals[pvals < 0.05])/1000*100, "%")
```

```
## 9.3 %
```

Again we get a very low percent of low p-values – there is no significant differences in city groups with respect to the percent of capitalized words in a comment.

6. Kruskal-Wallis rank sum test for age groups

6.1 Data Preprocessing

I will preprocess the data in a column "bdate" in order to use these values as factors and divide observations into groups by age.

```
# the code below converts "03.09.1993" into "1993"
cap_age <- filter(cap, bdate != '')
cap_age$bdate <- as.character(cap_age$bdate)
cap_age %>%
  mutate(n_dots = str_count(cap_age$bdate, fixed("."))) %>%
  filter(n_dots == 2) -> cap_age
cap_age$bdate <- strsplit(cap_age$bdate, "[.]")
new_df <- as.data.frame(do.call(rbind, cap_age$bdate))
names(new_df)[3] <- paste("year")
cap_age$bdate <- as.integer(as.character(new_df$year))

# remove bdate < 1960 because of outliers
# anyway we save more than 92% of observations that have information about age
cap_age_filtered <- filter(cap_age, bdate > 1960)

# add rounded age groups
cap_age_filtered$age_rounded <- round(as.integer(as.character(cap_age_filtered$bdate)),-1)

nrow(cap_age_filtered)
```

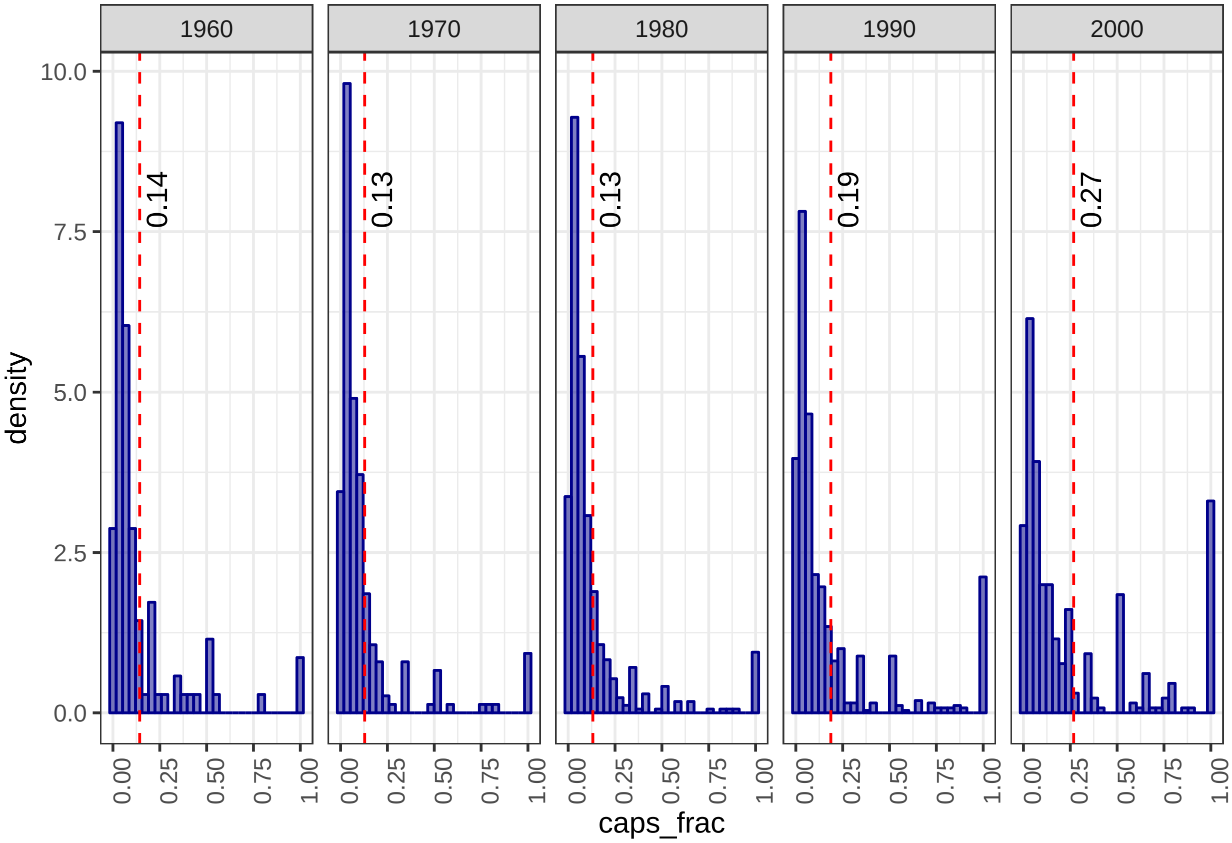
```
## [1] 1943
```

```
# we have 1943 observations.

# visualize rounded age groups
cap_age_filtered_avg <-
  cap_age_filtered %>%
  dplyr::group_by(age_rounded) %>%
  dplyr::summarize(avg_caps_frac = mean(caps_frac))
```

```r
ggplot(cap_age_filtered,aes(x=caps_frac)) +
  geom_histogram(aes(y=..density..),col = "darkblue", fill = "darkblue", alpha = .5) +
  geom_text(data = cap_age_filtered_avg,
            mapping = aes(x = avg_caps_frac,
                          y = 8,
                          label = round(avg_caps_frac,2)),
            angle = 90,
            vjust = 1.3) +
  facet_grid(~age_rounded) +
  geom_vline(cap_age_filtered_avg,
             mapping = aes(xintercept = avg_caps_frac),
             colour = 'red',
             linetype = 'dashed') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We see that distribution of capitalization patterns in different age groups changes. People born in 1960's tend to use capitalization less than those who was born in 1990's.

```r
# describe
psych::describeBy(cap_age_filtered$caps_frac, cap_age_filtered$age_rounded)
```

```
##
##  Descriptive statistics by group
## group: 1960
##    vars   n mean   sd median trimmed  mad  min max range skew kurtosis
## X1    1 101 0.14 0.21   0.07    0.09 0.06 0.01   1  0.99 2.66     7.15
##      se
## X1 0.02
## --------------------------------------------------------
## group: 1970
##    vars   n mean   sd median trimmed  mad  min max range skew kurtosis
## X1    1 219 0.13 0.21   0.06    0.08 0.05 0.01   1  0.99 3.07     9.26
##      se
## X1 0.01
## --------------------------------------------------------
## group: 1980
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 491 0.13 0.21   0.06    0.08 0.06   0   1     1 3.03     9.13 0.01
## --------------------------------------------------------
## group: 1990
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 754 0.19 0.28   0.07    0.12 0.08   0   1     1 2.09      3.2 0.01
## --------------------------------------------------------
## group: 2000
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 378 0.27 0.33   0.11    0.21 0.13   0   1     1 1.33     0.34 0.02
```

6.2 Kruskal-Wallis test

```r
kruskal.test(caps_frac ~ as.factor(age_rounded), data = cap_age_filtered)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  caps_frac by as.factor(age_rounded)
## Kruskal-Wallis chi-squared = 43.452, df = 4, p-value = 8.336e-09
```

At last we have found out a significant dependence of capitalization fraction on the age of the user.

7. Regression attempts.

I have tried to build several regression models and could not find a model that can explain different capitalization patterns. No interaction effects, such as of gender and age or of city and gender, were found. The perfomance of the best regression model is as follows:

```r
# preprocess
cap_regr <- cap_age_filtered # use data from previous section
cap_regr$age_rounded <- as.character(cap_regr$age_rounded)
cap_regr <- filter(cap_regr, city != '')
cap_regr$city <- ifelse(cap_regr$city == "Moscow" | cap_regr$city == "Saint Petersburg",
                        "Moscow & SPB",
                        "Other")
nrow(cap_regr)
```

```
## [1] 1943
```

```r
# regression
model <- lmer(caps_frac ~ (1 | age_rounded) +  sex + city, data=cap_regr)
summary(model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: caps_frac ~ (1 | age_rounded) + sex + city
##    Data: cap_regr
##
## REML criterion at convergence: 316.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.0300 -0.5341 -0.3544  0.0102  3.4185
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  age_rounded (Intercept) 0.003267 0.05716
##  Residual                0.067856 0.26049
## Number of obs: 1943, groups:  age_rounded, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.13968    0.03123   4.473
## sexmale      0.03079    0.01730   1.779
## cityOther    0.01039    0.01259   0.825
##
## Correlation of Fixed Effects:
##           (Intr) sexmal
## sexmale   -0.451
## cityOther -0.240 -0.061
```

```
coef(model)
```

```
## $age_rounded
##      (Intercept)    sexmale  cityOther
## 1960   0.1177988 0.03079047 0.01038537
## 1970   0.0989464 0.03079047 0.01038537
## 1980   0.0991260 0.03079047 0.01038537
## 1990   0.1527657 0.03079047 0.01038537
## 2000   0.2297653 0.03079047 0.01038537
##
## attr(,"class")
## [1] "coef.mer"
```

```
icc(model)
```

```
##
## Intraclass Correlation Coefficient for Linear mixed model
##
## Family : gaussian (identity)
## Formula: caps_frac ~ (1 | age_rounded) + sex + city
##
##   ICC (age_rounded): 0.0459
```

It is a mixed effects model with a random intercept. The reability of the model is still low (see ICC).

Simple linear regression models (e.g. $caps_{f}rac\ age_{r}ounded$) showed Rˆ2 around 0.03. Std.errors of coefficients were big, residuals were not normally distributed. However, in that model intercept and a “age_rounded2000” variable were significantly different from 0. So anyway we can conclude, that there is some linear significant connection between capitalization and age, but it is still very small.

In my opinion, to build a more reliable regression model, we should collect more data and consider other predictors such as social status, family status, political views, relion of the user, user interests etc.

8. Analyzing meaningully capitalized tokens.

I used the same code above to analyze the smaller dataset with meaninfully capitalized tokens. Although in Wilcoxon tests (caps_frac ~ sex/city) the percent of p-values < 0.05 was twice as bigger (around 19%), it still was not big anough to reject the null hypothesis. Kruskal-Wallis test also showed significant connection between age and capitalization:

```
Kruskal-Wallis chi-squared = 22.067, df = 4, p-value = 0.0001944)
```

The regression models did not give any significant results.

9. Conclusions

During my small research I have tested several hypotheses about possible connection between the percent of capitalized words in a comment and sex, gender and age of the user. The results showed that the age of the user unfluences on a fraction of capitalized words. The connection with gender was not found, also it was found in the previous research for English. I guess more data and more predictors could develop this analysis into aa more deeper and reliable one and reveal some significant differences of capitalization patterns on Russian social media.