Numeral predicate agreement in Russian language typical for bilingual speakers

Elizaveta Ezhergina

2020

Introduction

This project aims at researching numeral predicate agreement tendencies in Russian made by bilingual speakers whose dominant language is either English or Finnish, Russian being either their heritage language or a learned one with current fluent level. Suggested hypotheses are concerning possible correlation between choice of singular or plural number in case of various quantitative-nominal constructions and informant's linguistic background (is their Russian heritage or studied, is their dominant language Finnish or English).

Hypotheses

In this project, a set of related hypotheses to be proved or denied is the following:

1) H0: dominant language does not affect speakers' choice of number in case of quantitative-nominal constructions in Russian.

   H1: speakers with dominant Finnish tend to choose plural number (pl. later on) in case of quantitative-nominal constructions in Russian more than their counterparts with dominant English.

2) H0: heritage or studied Russian does not affect speakers' choice of number in case of quantitative-nominal constructions in Russian.

   H1: speakers with heritage Russian tend to choose pl. in case of quantitative-nominal constructions more than their counterparts with studied Russian.

3) H0: dominant language does not affect speakers' choice of construction type.

   H1: dominant language affects speakers' choice of construction type.

4) H0: it is not possible to predict number by the dominant language, type of Russian or combination of these.

   H1: it is possible to predict number by the dominant language, type of Russian or combination of these.

The suggestions about certain choices affected by dominant languages stem from linguistic features of English and Finnish: English tends to use plural forms in such constructions, especially in colloquial speech, while for Finnish singular forms are default in these cases unless there is a communicative marking that denotes otherwise.

Data & Research design

To test the provided hypotheses, a dataset based on data from Russian Learners Corpus (RLC) was used. RLC contains examples of uttered and written speech of people who fall into two categories: those who have studied Russian as a foreign language and those who have begun studying it in childhood as a first language but for some reason (mostly emigration) began using some other language as a primary one (so-called heritage Russian).

The dataset (in aggregated state of plots) was obtained from Anastasia Dolgova's thesis presentation and disaggregated manually into a table with 5 columns. Example sequence can be observed in Table 1.

Table 1. Example data

index	dominant_lang	russian_type	construction_type	selected_number
1	finnish	heritage	percent	sg
2	english	studied	some	pl

Columns

index: entry number; entry is a case of using a certain construction that requires choosing a specific numeral predicate agreement.

dominant_lang: a language that is dominant to the speaker (either English or Finnish).

russian_type: whether Russian was studied by the speaker or whether it is heritage.

construction_type: a type of construction that requires choosing a specific numeral predicate agreement. Possible options: numeric_construction (i.e. четыре собаки прошли мимо (pl.), там погибло пятнадцать человек (sg.)), majority (большинство, множество), percent, some (несколько, столько, сколько), avg_many_little (приблизительно, много, мало).

selected_number: number used in the entry.

To test the aforementioned hypotheses #1 and #2, Fischer's Exact Test and Pearson's Chi-squared Test will be used. Hypotheses #3 and #4 will be tested with logistic regression.

Collected data & Exploratory analysis

```
data <- read.csv('https://raw.githubusercontent.com/iftwigs/LingData_project/master/numeric_constr.csv')
```

```
head(data, n=20)
```

```
##    index dominant_lang russian_type    construction_type selected_number
## 1      1       finnish     heritage numeral_construction              sg
## 2      2       finnish     heritage numeral_construction              sg
## 3      3       finnish     heritage numeral_construction              sg
## 4      4       finnish     heritage numeral_construction              sg
## 5      5       finnish     heritage numeral_construction              sg
## 6      6       finnish     heritage numeral_construction              sg
## 7      7       finnish     heritage numeral_construction              sg
## 8      8       finnish     heritage numeral_construction              sg
## 9      9       finnish     heritage numeral_construction              sg
## 10    10       finnish     heritage numeral_construction              sg
## 11    11       finnish     heritage numeral_construction              sg
## 12    12       finnish     heritage numeral_construction              sg
## 13    13       finnish     heritage numeral_construction              sg
## 14    14       finnish     heritage numeral_construction              sg
## 15    15       finnish     heritage numeral_construction              pl
## 16    16       finnish     heritage numeral_construction              pl
## 17    17       finnish     heritage numeral_construction              pl
## 18    18       finnish     heritage numeral_construction              pl
## 19    19       finnish     heritage numeral_construction              pl
## 20    20       finnish     heritage numeral_construction              pl
```

```
library(tidyverse)
library(ggplot2)
summary(data)
```

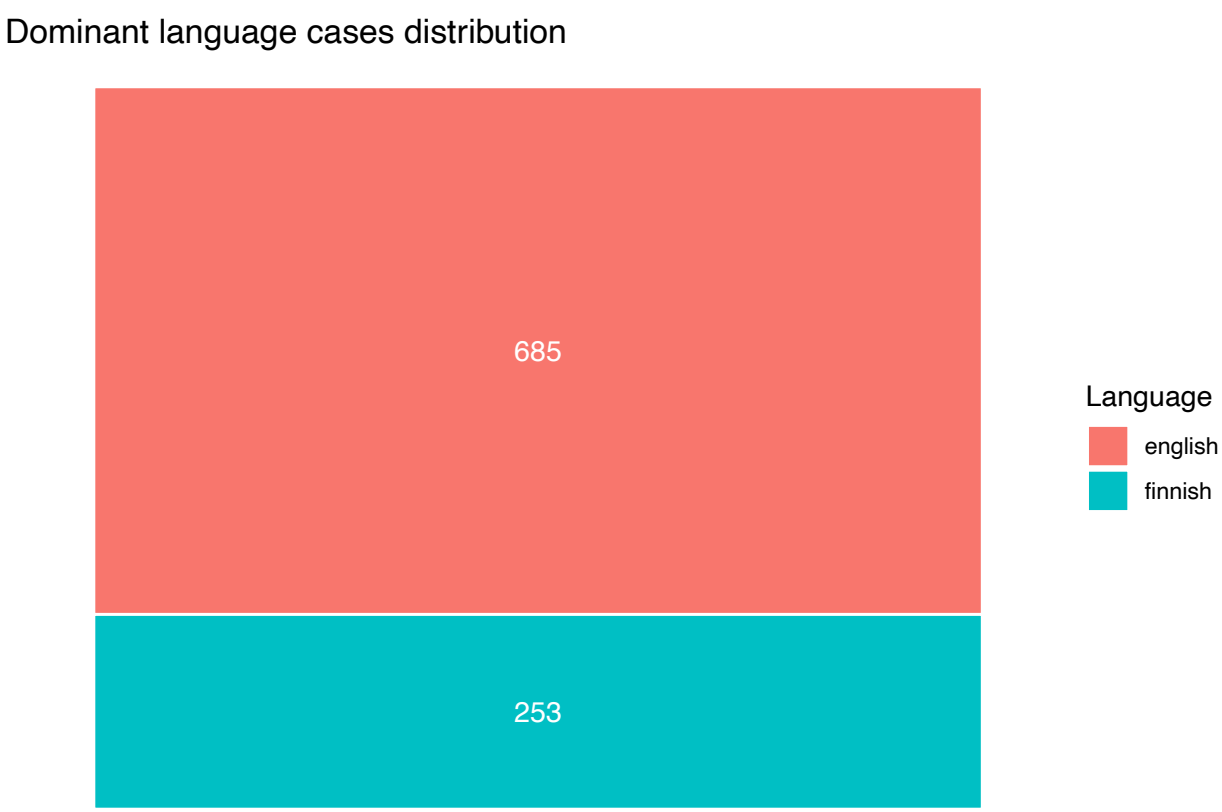
Checking the distribution of languages, we can see that English is almost 3 times more frequent in the dataset than Finnish, and this tendency is preserved after splitting speakers by their Russian background. Moreover, it can be noted that Russian as a foreign language is more common in the dataset. The most frequen construction type for Finnish is numeral construction and for English these are the words like приблизительно, много, мало.

The majority of speakers of both languages used numeral constructions, majority-type words and words with the meaning of "some" with Pl., and tended to use words like приблизительно, много, мало with Sg.

The least represented construction type for both languages is the one that uses the word "percent".

```
languages <- data.frame(table(distinct(data[,1:2])$dominant_lang))

ggplot(languages, aes(x="", y=Freq, fill = Var1)) +
  geom_bar(width = 1, stat = "identity", color = "white")+
  geom_text(aes(label = Freq), position = position_stack(vjust = 0.5), color = "white") +
  theme_void() +
  labs(fill = "Language", title = "Dominant language cases distribution")
```

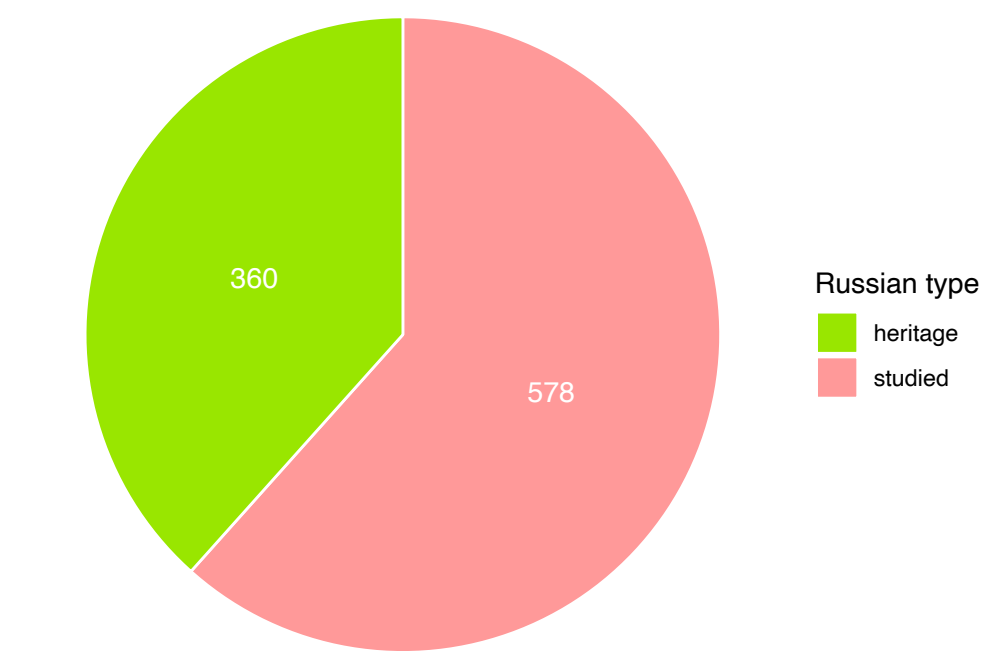



```
russian <- data.frame(table(data$russian_type))
russian
```

```
##       Var1 Freq
## 1 heritage  360
## 2  studied  578
```

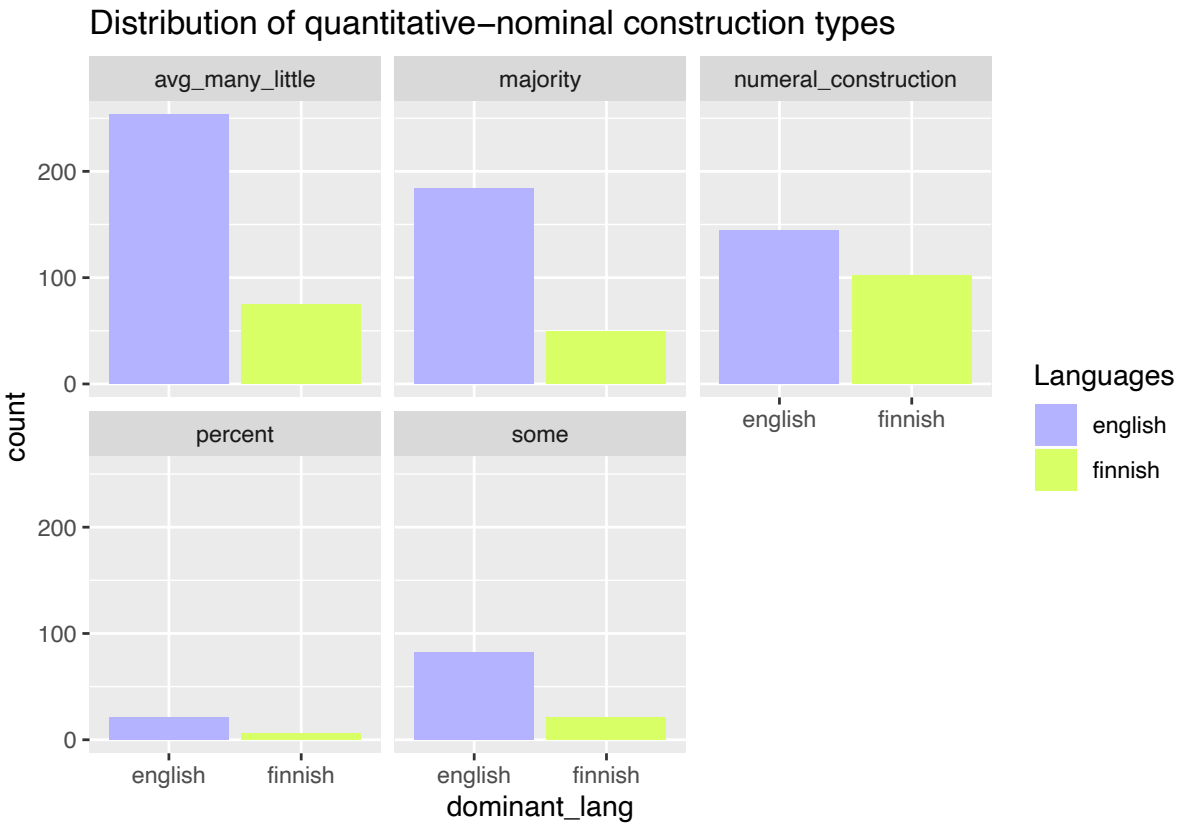
```
colors <- c("#99e600", "#ff9999")

ggplot(russian, aes(x="", y=Freq, fill = Var1)) +
  geom_bar(width = 1, stat = "identity", color = "white")+
  coord_polar("y", start=0) +
  geom_text(aes(label = Freq), position = position_stack(vjust = 0.5), color = "white") +
  theme_void() +
  labs(fill = "Russian type", title = "Distribution of types of Russian") +
  scale_fill_manual(values=colors)
```



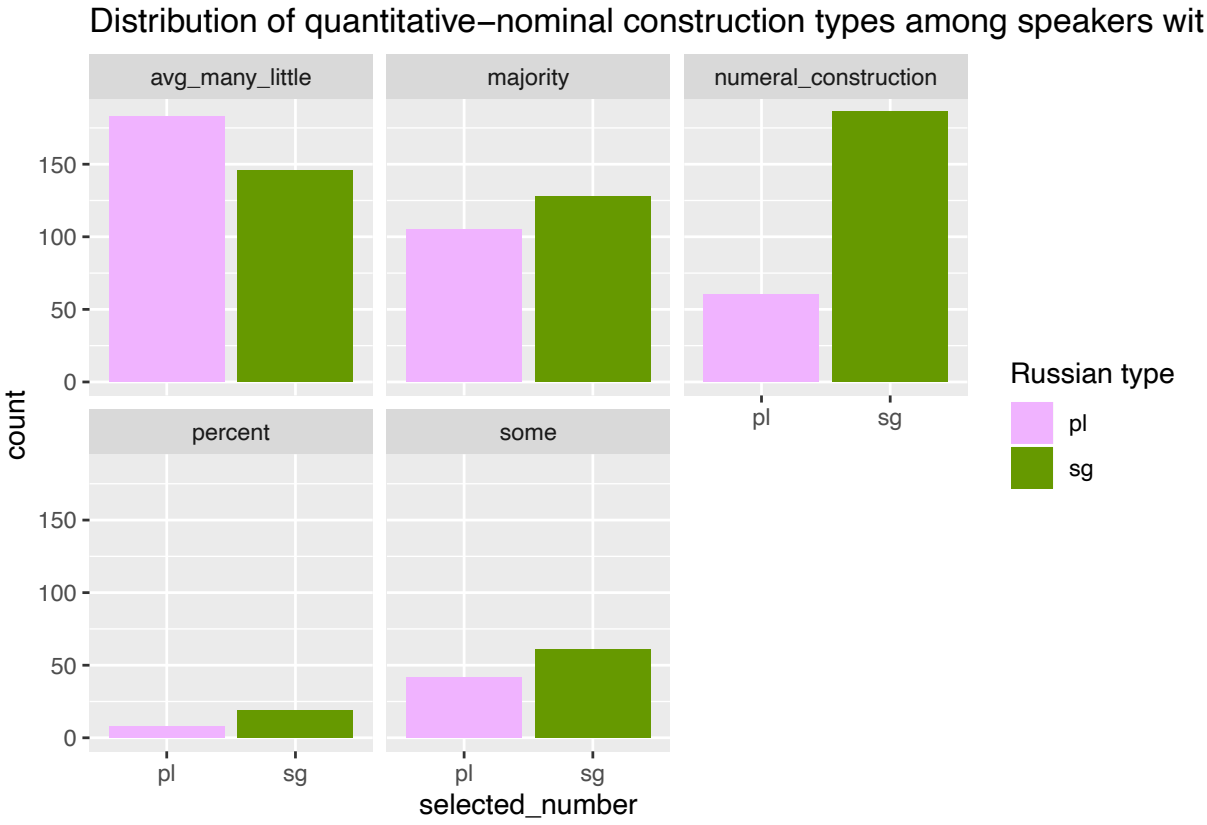

```
colors <- c('#b3b3ff', '#d9ff66')

data %>%
   ggplot(aes(dominant_lang, fill=dominant_lang)) +
   geom_bar() +
   facet_wrap(~construction_type) +
   labs(fill = "Languages", title = "Distribution of quantitative-nominal construction types") +
  scale_fill_manual(values=colors)
```




```
colors <- c('#f0b3ff', '#669900')

data %>%
  ggplot(aes(selected_number, fill=selected_number)) +
  geom_bar() +
  facet_wrap(~construction_type) +
  labs(fill = "Russian type", title = "Distribution of quantitative-nominal construction types among speakers with different
 scale_fill_manual(values=colors)
```




```
contingency <- table(data$dominant_lang, data$russian_type)
contingency
```

```
##
##          heritage studied
##   english     279     406
##   finnish      81     172
```

Statistical tests

The first hypothesis to check is "speakers with dominant Finnish tend to choose pl. in case of quantitative-nominal constructions in Russian more than their counterparts with dominant English". It can be tested with Pearson's Chi-squared Test because the two factors are independent.

```
test_1 <- chisq.test(data$dominant_lang, data$selected_number)
test_1
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data$dominant_lang and data$selected_number
## X-squared = 0.17993, df = 1, p-value = 0.6714
```

```
odds_ratio <- (test_1$observed[1,1] / test_1$observed[2,1] / test_1$observed[1,2] / test_1$observed[2,2])
odds_ratio
```

```
## [1] 4.852337e-05
```

The p-value > 0.005 so the zero hypothesis is not rejected. The choice of number is not affected by the

dominant language of the speaker.

The second hypothesis suggests that speakers with heritage Russian tend to choose pl. in case of quantitative-nominal constructions more than their counterparts with studied Russian.

```
test_2 <- fisher.test(data$russian_type, data$selected_number)
test_2
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data$russian_type and data$selected_number
## p-value = 9.707e-07
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.479722 2.572847
## sample estimates:
## odds ratio
##   1.949987
```

The p-value > 0.005 so the zero hypothesis is not rejected. Heritage or studied Russian does not affect speakers' choice of number in case of quantitative-nominal constructions in Russian.

The third hypothesis suggests that the dominant language affects speakers' choice of construction type.

```
test_3 <- chisq.test(data$dominant_lang, data$construction_type)
test_3
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$dominant_lang and data$construction_type
## X-squared = 35.892, df = 4, p-value = 3.046e-07
```

The p-value > 0.005 so the zero hypothesis is not rejected. Dominant language does not affect speakers' choice of construction type.

The fourth hypothesis is in fact three hypotheses that require testing and suggest it is possible to predict number by the dominant language, type of Russian or combination of these. Linear regression model will be used to test them.

```
data_2 <- data
data_2$dominant_lang <- factor(data_2$dominant_lang)
data_2$russian_type <- factor(data_2$russian_type)
glm_1 <- glm(selected_number ~ dominant_lang, data_2, family='binomial')
summary(glm_1)
```

```
##
## Call:
## glm(formula = selected_number ~ dominant_lang, family = "binomial",
##     data = data_2)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.333  -1.301   1.029   1.059   1.059
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)           0.28513    0.07719   3.694 0.000221 ***
## dominant_langfinnish  0.07443    0.14928   0.499 0.618087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1278.8  on 937  degrees of freedom
## Residual deviance: 1278.5  on 936  degrees of freedom
## AIC: 1282.5
##
## Number of Fisher Scoring iterations: 4
```

We turn dominant language and russian type into factors to indicate that they are categorical variables here. The output signifies that the p-value is small enough to reject the null hypothesis. It is possible to predict number when the dominant language is known.

```
glm_2 <- glm(selected_number ~ russian_type, data_2, family='binomial')
summary(glm_2)
```

```
##
## Call:
## glm(formula = selected_number ~ russian_type, family = "binomial",
##     data = data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4264  -1.1352   0.9474   0.9474   1.2202
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.1001     0.1055  -0.948    0.343
## russian_typestudied   0.6685     0.1365   4.898  9.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1278.8  on 937  degrees of freedom
## Residual deviance: 1254.6  on 936  degrees of freedom
## AIC: 1258.6
##
## Number of Fisher Scoring iterations: 4
```

The output signifies that the p-value is not small enough to reject the null hypothesis. It is not possible to predict number when the russian type is known.

```
glm_3 <- glm(selected_number ~ russian_type + dominant_lang, data_2, family='binomial')
summary(glm_3)
```

```
##
## Call:
## glm(formula = selected_number ~ russian_type + dominant_lang,
##     family = "binomial", data = data_2)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.4316  -1.1406   0.9430   0.9493   1.2218
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.10380    0.11092  -0.936    0.349
## russian_typestudied   0.66736    0.13694   4.874  1.1e-06 ***
## dominant_langfinnish  0.01652    0.15162   0.109    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1278.8  on 937  degrees of freedom
## Residual deviance: 1254.6  on 935  degrees of freedom
## AIC: 1260.6
##
## Number of Fisher Scoring iterations: 4
```

The output signifies that the p-value is not small enough to reject the null hypothesis. It is not possible to predict number when the dominant language and russian type is known.

Summary

Upon having conducted the research it was found that all of the suggested hypotheses should be rejected due to the lack of statistical significance. Selected number is not affected neither by russian type or dominant language of the speakers; the only case where the zero hypothesis was rejected was the suggestion that it is possible to predict number when the dominant language is known.