Preposition drop in Daghestanian Russian

Final project on linguistic data analysis (version 2, 14.06.20)
Anastasia Panova, *anastasia.b.panova@gmail.com*

0. Preliminary remarks

This project is a statistical part of the bigger project "Preposition drop in Russian spoken in Daghestan" carried out at Linguistic Convergence Laboratory (HSE) by me and my colleague Tatiana Philippova. Although we annotated data together since November 2019, the statistical part was done by me during the course on linguistic data analysis (mainly in April 2020), so I hope I am allowed to submit this work as my research project now. Some parts of the present document are taken from our recently submitted paper.

1. Research objectives and hypothesis to be tested

As Daniel et al. (2009) note, "[a] very frequent, and indeed probably one of the most salient linguistic features of the local variety of Russian [in Daghestan] is dropping of the prepositions", cf. (1).

(1) tam naprimer Curibe jest' vrač
there for_example Tsurib.LOC COP.PRS.SG doctor
'For instance, [in] Tsurib there is a doctor.' (Daniel et al. 2010: 74)

In the previous studies (Daniel & Dobrushina 2009, 2013; Daniel et al. 2010), the phenomenon of preposition drop had been described primarily in qualitative terms. The purpose of the present project was a detailed quantitative study of this phenomenon across a large number of speakers of different L1s. In particular, we wanted to understand what factors condition the phenomenon of preposition drop in locative, directional and temporal phrases.

Based on existing literature on preposition drop in different variaties of different languages, we decided to check whether the probability of preposition drop in Daghestanian Russian depends on preposition type, phonetic environment, semantic type of prepositional complement and sociolinguistic characteristics of the speakers.

2. Description of input data: features and values, descriptive statistics, data visualisation

For this research we used data from the Corpus of Russian spoken in Daghestan (DagRus). Specifically, my input data was a dataset consisting of 2350 prepositional phrases, coming from sociolinguistic interviews with 47 speakers.

Each prepositional phrase (with or without preposition drop) was annotated with a number of parameters:

- speaker's ID;
- sex;
- year of birth;
- native language;
- education level;
- prepositional head;
- initial phoneme of the prepositional complement (consonant/vowel);
- complement type (toponym, temporal location, institution, other).

A csv file with annotated data can be found on Github.

In addition, for each speaker we annotated the degree of nonstandardness of his/her speech. The nonstandardness was calculated as a ratio of the total number of discrepancies from Standard Russian (excluding preposition drop) to the total number of words produced by a speaker.

The data on nonstandardness can be found on Github as well.

Now let me load and prepare these datasets for further analysis (you can surely skip section 2.1, if your are not intersted in these rather technical things).

2.1. Preparation of the data

The first dataset:

```
dat <- read.csv("Prep_drop_final_data.csv")
dat %>%
  select(3:5, 7:8, 14:16, 18) -> mydat
names(mydat)[9] <- "preposition"
# summary(mydat)
```

The second dataset (the nonstandardness is multiplied by one hundred to obtain the average number of discrepancies from Standard Russian per 100 words):

```
dat_lR <- read.csv("Prep_drop_DagRus - level of Russian.csv")
dat_lR %>%
  mutate (nonstandardness = non.standardness*100) %>%
  select(1, 16) -> dat_lR
names(dat_lR)[1] <- "respondent"
# summary(dat_lR)
```

Merging two datasets into one:

```
full_join(mydat, dat_lR, by = "respondent") -> mydat
```

```
## Warning: Column ‘respondent‘ joining factors with different levels, coercing to
## character vector
```

The variables `education` and`language group` have some values which are represented by a too small number of datapoints, so I unite five levels of education into just two (`higher` and `lower`) and unite language groups into langugae families (`Daghestanian`, `Indo-European`, `Turkic`).

```
mydat %>%
  mutate(ed_levels = ifelse(mydat$education == "higher education",
                            "higher", "non-higher")) -> mydat
mydat %>%
  mutate(ed_levels = as.factor(ed_levels)) -> mydat
mydat %>%
  mutate(lang_family = ifelse(mydat$language.group == "Turkic", "Turkic",
                              ifelse(mydat$language.group == "Indo-European",
                                     "Indo-European", "Daghestanian"))) -> mydat
mydat %>%
  mutate(lang_family = as.factor(lang_family)) -> mydat
```

I exclude speakers with Russian as L1 because they do not omit prepositions at all and this ruins the logistic regression in the end.

```
mydat %>%
  filter(lang_family != "Indo-European") -> mydat
```

Below I put into order the values of the parameter `omitted`. I do not include these lines in pdf because they contain cyrillic charachters which for some reason ruin the process of knitting to pdf.

2.2. Descriptive statisctics and data visualization

In this section I look at each of my parameters separately and try to visualize their possible correlations with preposition drop.

First, I look at different prepositions. The table shows that only two prepositions *v* 'in(to)' and *na* 'on(to)' are omitted frequently.

```
mydat %>%
  count(preposition, omitted) %>%
  spread(omitted, n, fill = 0) %>%
  mutate(total_n = no+yes) %>%
  mutate(yes_percent = (yes/total_n)*100) %>%
  arrange(desc(total_n))
```

```
## # A tibble: 33 x 5
##    preposition               no   yes total_n yes_percent
##    <fct>                  <dbl> <dbl>   <dbl>       <dbl>
##  1 v 'in(to)'               479   351     830       42.3
##  2 u 'at'                   400     0     400        0
##  3 na 'on(to)'              289    45     334       13.5
##  4 s 'with/from/off'        193    14     207        6.76
##  5 iz 'from, of'             88     4      92        4.35
##  6 do 'up to, until'         69     0      69        0
##  7 za 'behind; for'          64     3      67        4.48
##  8 k 'to'                    61     3      64        4.69
##  9 po 'along/about/up to'    64     0      64        0
## 10 posle 'after'             44     0      44        0
## # ... with 23 more rows
```

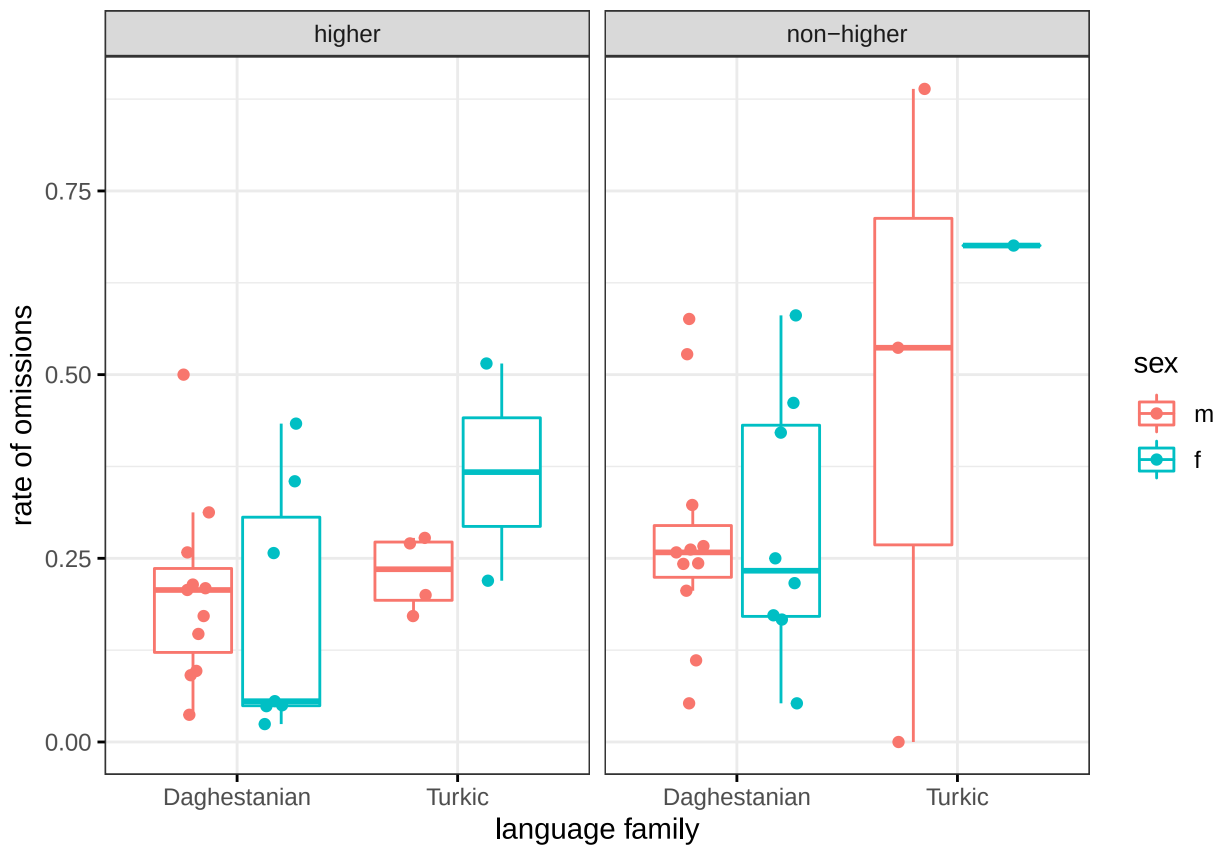
Second, I look at the first phoneme of the prepositional complement. The difference is very small.

```
mydat %>%
  count(initial.phoneme, omitted) %>%
  spread(omitted, n, fill = 0) %>%
  mutate(total_n = no+yes) %>%
  mutate(yes_percent = (yes/total_n)*100)
```

```
## # A tibble: 2 x 5
##   initial.phoneme    no   yes total_n yes_percent
##   <fct>           <dbl> <dbl>   <dbl>       <dbl>
## 1 consonant        1601   354    1955        18.1
## 2 vowel             328    67     395        17.0
```

Then I turn to sociolinguistic parameters. Level of education, sex and language family are visulalized together with the library `ggpubr`. In the folowing figures I consider only prepositional phrases that are headed by seven prepositions that are in principle omittable: this way I partially solve the problem of an uneven distribution of omittable and non-omittable prepositions across speakers.

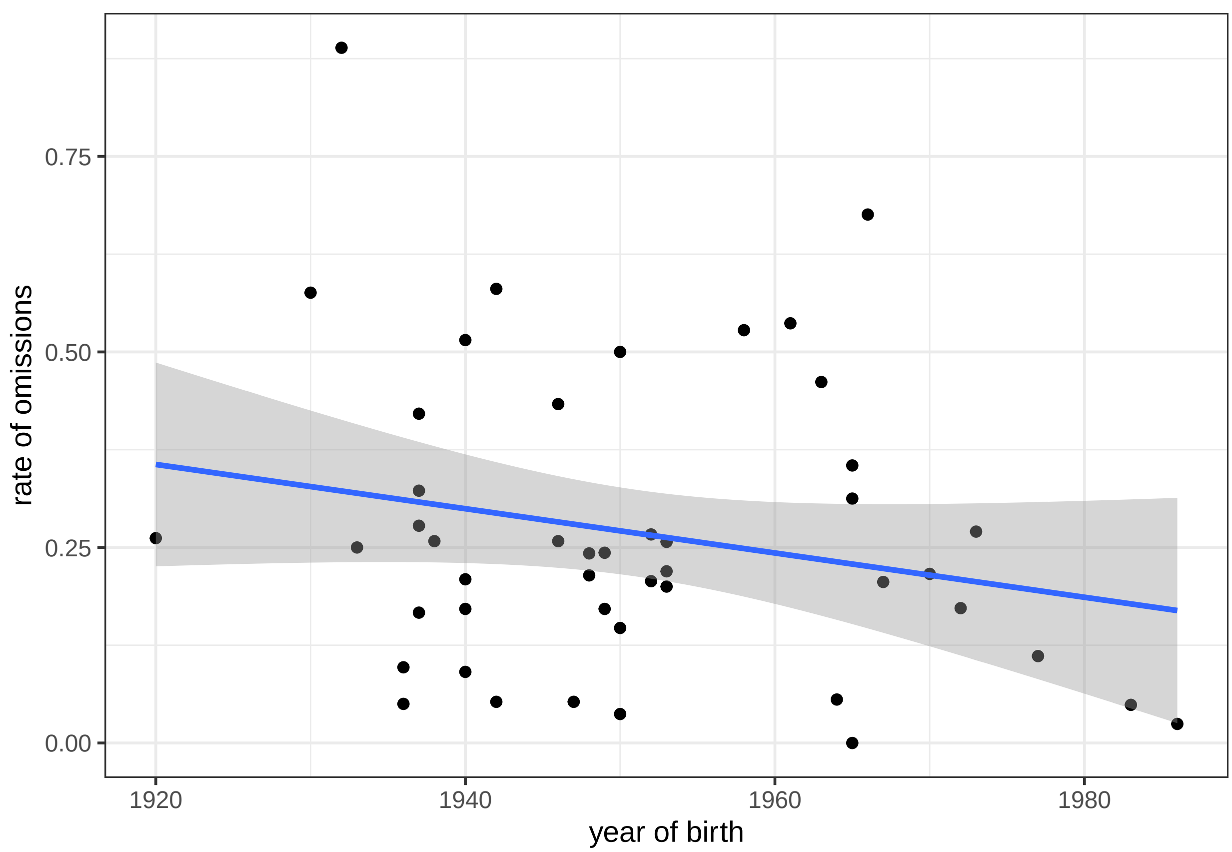
```
mydat %>%
  filter(preposition == "v 'in(to)'" | preposition == "na 'on(to)'" |
           preposition == "s 'with/from/off'" | preposition == "iz 'from, of'" |
           preposition == "za 'behind; for'" | preposition == "k 'to'" |
           preposition == "pro 'about") %>%
  count(respondent, ed_levels, lang_family, sex, omitted) %>%
  spread(omitted, n, fill = 0) %>%
  mutate(n = yes+no)%>%
  mutate(ratio = yes/n)%>%
  ggboxplot(x = "lang_family", y = "ratio",
            color = "sex",
            add = "jitter",
            outlier.shape = NA,
            ggtheme = theme_bw(),
            add.params = list(jitter = 0.3),
            ylab = "rate of omissions",
            xlab = "language family",) -> p
facet(p, facet.by = "ed_levels")
```



The next figure shows how the ratio of omissions to the number of produced omittable prepositions depends on the year a speaker was born in. Each point corresponds to one speaker. We see that there is no significant correlation.

```
mydat %>%
  filter(preposition == "v 'in(to)'" | preposition == "na 'on(to)'" |
           preposition == "s 'with/from/off'" | preposition == "iz 'from, of'" |
           preposition == "za 'behind; for'" | preposition == "k 'to'" |
           preposition == "pro 'about") %>%
  count(respondent, year.of.birth, omitted) %>%
  spread(omitted, n, fill = 0) %>%
  mutate(n = yes+no)%>%
  mutate(ratio = yes/n)%>%
  ggplot(aes(year.of.birth, ratio))+
  geom_point()+
  geom_smooth(method=lm, se=TRUE) +
  labs(x = "year of birth",
       y = "rate of omissions")+
  theme_bw()
```

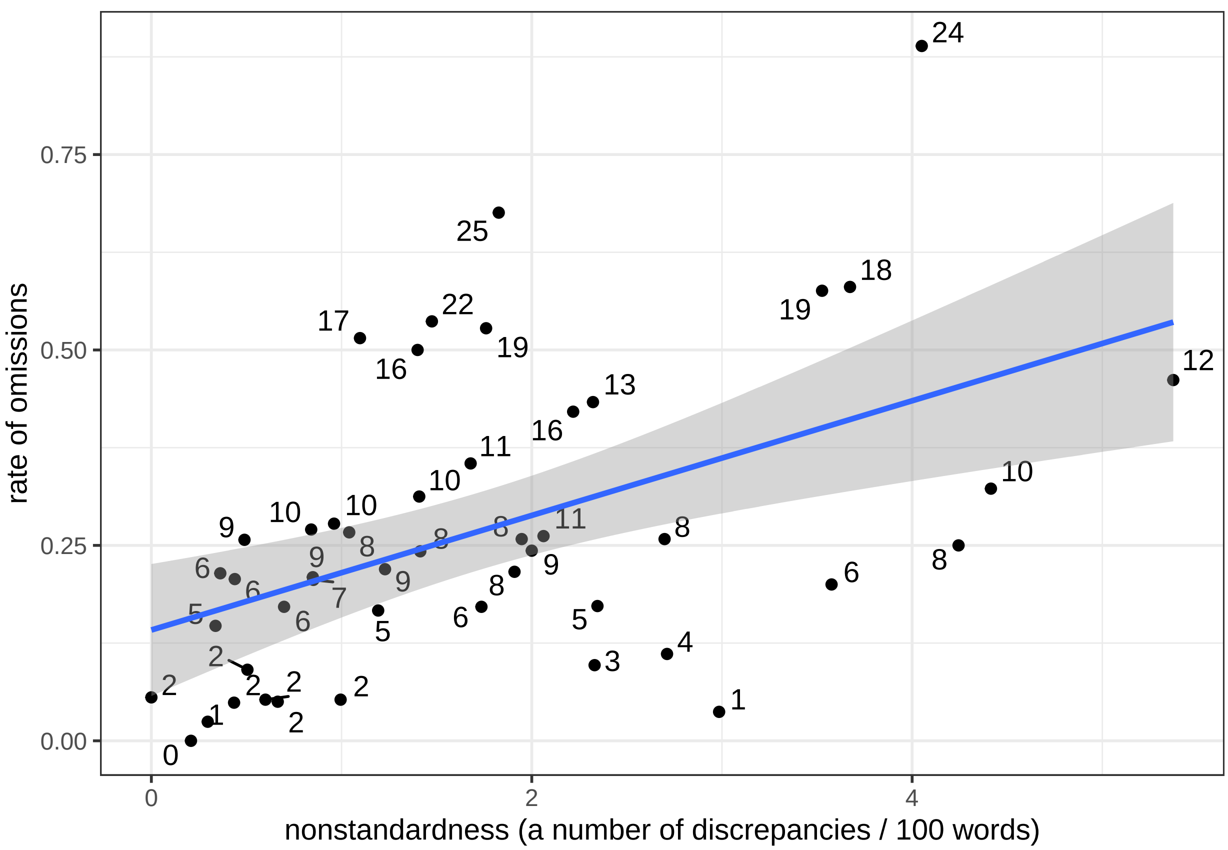
```
## `geom_smooth()` using formula 'y ~ x'
```



The next figure shows the relation between the rate of omissions and the speaker's command of Standard Russian. The linear trend reveals that speakers with a better command of Standard Russian (displaying fewer non-standard features) tend to omit prepositions less frequently.

```r
mydat %>%
  filter(preposition == "v 'in(to)'" | preposition == "na 'on(to)'" |
           preposition == "s 'with/from/off'" | preposition == "iz 'from, of'" |
           preposition == "za 'behind; for'" |
           preposition == "k 'to'" | preposition == "pro 'about") %>%
  count(respondent, nonstandardness, omitted) %>%
  spread(omitted, n, fill = 0) %>%
  mutate(n = yes+no)%>%
  mutate(ratio = yes/n)%>%
  ggplot(aes(nonstandardness, ratio, label = paste0(yes)))+
  geom_point()+
  ggrepel::geom_text_repel()+
  geom_smooth(method=lm, se=TRUE) +
  labs(x = "nonstandardness (a number of discrepancies / 100 words)",
       y = "rate of omissions")+
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



3. Discussion of the methods of analysis and their application

3.1. What kind of prepositional phrases allow preposition drop

I decided to use a logistic regression to assess the significance of the factors discussed above.

Before running a regression, I had to reduce a number of levels in the parameter `preposition`. I grouped prepositions in two types based on linguistic grounds: the prepositions v 'in(to)' and na 'on(to)' in Standard Russian are precisely those used in general locative and directional phrases, not necessarily specifying the relation between the locatum and the location. Therefore, they are grouped together and contrasted to all other prepositions.

```
mydat %>%
  mutate(prep_type = ifelse(mydat$preposition == "v 'in(to)'", "prep_v/na",
                                             ifelse(mydat$preposition == "na 'on(to)'",
                                                    "prep_v/na", "prep_other"))) -> mydat
mydat %>%
  mutate(prep_type = as.factor(prep_type)) -> mydat
```

Then I change values of the parameter `year.of.birth` in order to make it more centered. I save all values as factors.

```
mydat %>%
  mutate(year.of.birth = year.of.birth-1900) -> log_dat
log_dat %>%
  mutate(year.of.birth = as.integer(year.of.birth)) -> log_dat
log_dat %>%
  mutate(omitted = as.factor(omitted)) -> log_dat
```

I use library `lme4` to run a mixed-effects model. I have one random effect which is a speker.

```
library("lme4")
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
glmer.res <- glmer (omitted ~ sex + year.of.birth + ed_levels + lang_family + nonstandardness + initial
summary(glmer.res)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## omitted ~ sex + year.of.birth + ed_levels + lang_family + nonstandardness +
##     initial.phoneme + prep_type + (1 | respondent)
##    Data: log_dat
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   1588.2   1640.0   -785.1   1570.2     2341
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1187 -0.4091 -0.1320 -0.0710 17.4458
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  respondent (Intercept) 0.5852   0.765
## Number of obs: 2350, groups:  respondent, 47
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -4.658645   0.616555  -7.556 4.16e-14 ***
## sexf                  0.028581   0.277834   0.103  0.91807
## year.of.birth        -0.009553   0.009703  -0.985  0.32485
## ed_levelsnon-higher   0.296403   0.290527   1.020  0.30762
## lang_familyTurkic     0.626626   0.317312   1.975  0.04829 *
## nonstandardness       0.366063   0.120291   3.043  0.00234 **
## initial.phonemevowel -0.426730   0.174045  -2.452  0.01421 *
## prep_typeprep_v/na    3.503915   0.223124  15.704  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) sexf   yr.f.b ed_lv- lng_fT nnstnd intl.p
## sexf         0.034
## year.f.brth -0.824 -0.231
## ed_lvlsnn-h -0.081  0.012 -0.034
## lng_fmlyTrk -0.090  0.087 -0.082  0.102
## nnstndrdnss -0.461 -0.097  0.257 -0.414 -0.006
## intl.phnmvw -0.020  0.025 -0.005 -0.008 -0.008 -0.005
## prp_typpr_/ -0.346 -0.004 -0.008  0.007  0.009  0.070 -0.070
```

After that I am trying to choose the best model based on AIC.

```
drop1(glmer.res, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## omitted ~ sex + year.of.birth + ed_levels + lang_family + nonstandardness +
##     initial.phoneme + prep_type + (1 | respondent)
##                 npar    AIC    LRT Pr(Chi)
## <none>               1588.2
## sex                1 1586.2   0.01 0.91855
## year.of.birth      1 1587.2   0.96 0.32629
## ed_levels          1 1587.2   1.02 0.31311
## lang_family        1 1589.9   3.67 0.05530 .
## nonstandardness    1 1594.8   8.58 0.00340 **
## initial.phoneme    1 1592.3   6.15 0.01312 *
## prep_type          1 2086.6 500.42 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The line which has the smallest AIC is `sex`, so I remove `sex`.

```r
glmer.res2 <- glmer (omitted ~ year.of.birth + ed_levels + lang_family + nonstandardness + initial.phon
summary(glmer.res2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: omitted ~ year.of.birth + ed_levels + lang_family + nonstandardness +
##     initial.phoneme + prep_type + (1 | respondent)
##    Data: log_dat
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   1586.2   1632.3   -785.1   1570.2     2342
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1205 -0.4076 -0.1321 -0.0707 17.5168
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  respondent (Intercept) 0.5857   0.7653
## Number of obs: 2350, groups:  respondent, 47
##
## Fixed effects:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.660779   0.616411  -7.561    4e-14 ***
## year.of.birth         -0.009324   0.009445  -0.987  0.32354
## ed_levelsnon-higher    0.296038   0.290584   1.019  0.30831
## lang_familyTurkic      0.623800   0.316189   1.973  0.04851 *
## nonstandardness        0.367283   0.119757   3.067  0.00216 **
## initial.phonemevowel -0.427187   0.173991  -2.455  0.01408 *
## prep_typeprep_v/na     3.504032   0.223122  15.705  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) yr.f.b ed_lv- lng_fT nnstnd intl.p
## year.f.brth -0.839
## ed_lvlsnn-h -0.082 -0.032
## lng_fmlyTrk -0.093 -0.064  0.101
## nnstndrdnss -0.460  0.243 -0.415  0.003
## intl.phnmvw -0.021  0.000 -0.008 -0.010 -0.003
## prp_typpr_/ -0.346 -0.009  0.007  0.009  0.070 -0.070
```

```r
drop1(glmer.res2, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## omitted ~ year.of.birth + ed_levels + lang_family + nonstandardness +
##     initial.phoneme + prep_type + (1 | respondent)
##                 npar    AIC    LRT   Pr(Chi)
```

```
## <none>                1586.2
## year.of.birth       1 1585.2   0.97  0.324629
## ed_levels           1 1585.2   1.01  0.313735
## lang_family         1 1587.9   3.67  0.055497 .
## nonstandardness     1 1592.9   8.70  0.003188 **
## initial.phoneme     1 1590.4   6.17  0.012995 *
## prep_type           1 2084.8 500.55 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The lines which have the smallest AIC is `year.of.birth` and `ed_levels`, so I remove `year.of.birth`.

```r
glmer.res3 <- glmer (omitted ~ ed_levels + lang_family + nonstandardness + initial.phoneme + prep_type +
summary(glmer.res3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## omitted ~ ed_levels + lang_family + nonstandardness + initial.phoneme +
##     prep_type + (1 | respondent)
##    Data: log_dat
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   1585.2   1625.5   -785.6   1571.2     2343
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0955 -0.4012 -0.1346 -0.0696 17.8065
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  respondent (Intercept) 0.5969   0.7726
## Number of obs: 2350, groups:  respondent, 47
##
## Fixed effects:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -5.1775     0.3362 -15.402  < 2e-16 ***
## ed_levelsnon-higher    0.2864     0.2925   0.979 0.327524
## lang_familyTurkic      0.6038     0.3178   1.900 0.057387 .
## nonstandardness        0.3969     0.1169   3.395 0.000687 ***
## initial.phonemevowel  -0.4279     0.1739  -2.460 0.013889 *
## prep_typeprep_v/na     3.5047     0.2230  15.717  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) ed_lv- lng_fT nnstnd intl.p
## ed_lvlsnn-h -0.200
## lng_fmlyTrk -0.271  0.099
## nnstndrdnss -0.487 -0.421  0.019
## intl.phnmvw -0.038 -0.008 -0.010 -0.003
## prp_typpr_/ -0.649  0.007  0.008  0.074 -0.070
```

```r
drop1(glmer.res3, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## omitted ~ ed_levels + lang_family + nonstandardness + initial.phoneme +
##     prep_type + (1 | respondent)
##                 npar    AIC    LRT   Pr(Chi)
## <none>               1585.2
## ed_levels          1 1584.1   0.94  0.332832
## lang_family        1 1586.6   3.41  0.064783 .
## nonstandardness    1 1593.7  10.52  0.001182 **
## initial.phoneme    1 1589.4   6.19  0.012825 *
## prep_type          1 2084.2 501.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The line which has the smallest AIC is `ed_levels`, so I remove `ed_levels`.

```r
glmer.res4 <- glmer (omitted ~ lang_family + nonstandardness + initial.phoneme + prep_type + (1|responde
summary(glmer.res4)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## omitted ~ lang_family + nonstandardness + initial.phoneme + prep_type +
##     (1 | respondent)
##    Data: log_dat
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   1584.1   1618.7   -786.1   1572.1     2344
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0888 -0.3847 -0.1340 -0.0667 18.5750
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  respondent (Intercept) 0.6195   0.7871
## Number of obs: 2350, groups:  respondent, 47
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.1178     0.3317 -15.430  < 2e-16 ***
## lang_familyTurkic       0.5725     0.3210   1.783   0.0745 .
## nonstandardness         0.4461     0.1075   4.148 3.35e-05 ***
## initial.phonemevowel   -0.4273     0.1740  -2.456   0.0141 *
## prep_typeprep_v/na      3.5060     0.2229  15.726  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation of Fixed Effects:
##              (Intr) lng_fT nnstnd intl.p
## lng_fmlyTrk -0.257
## nnstndrdnss -0.645  0.065
## intl.phnmvw -0.040 -0.009 -0.007
## prp_typpr_/ -0.656  0.007  0.084 -0.071
```

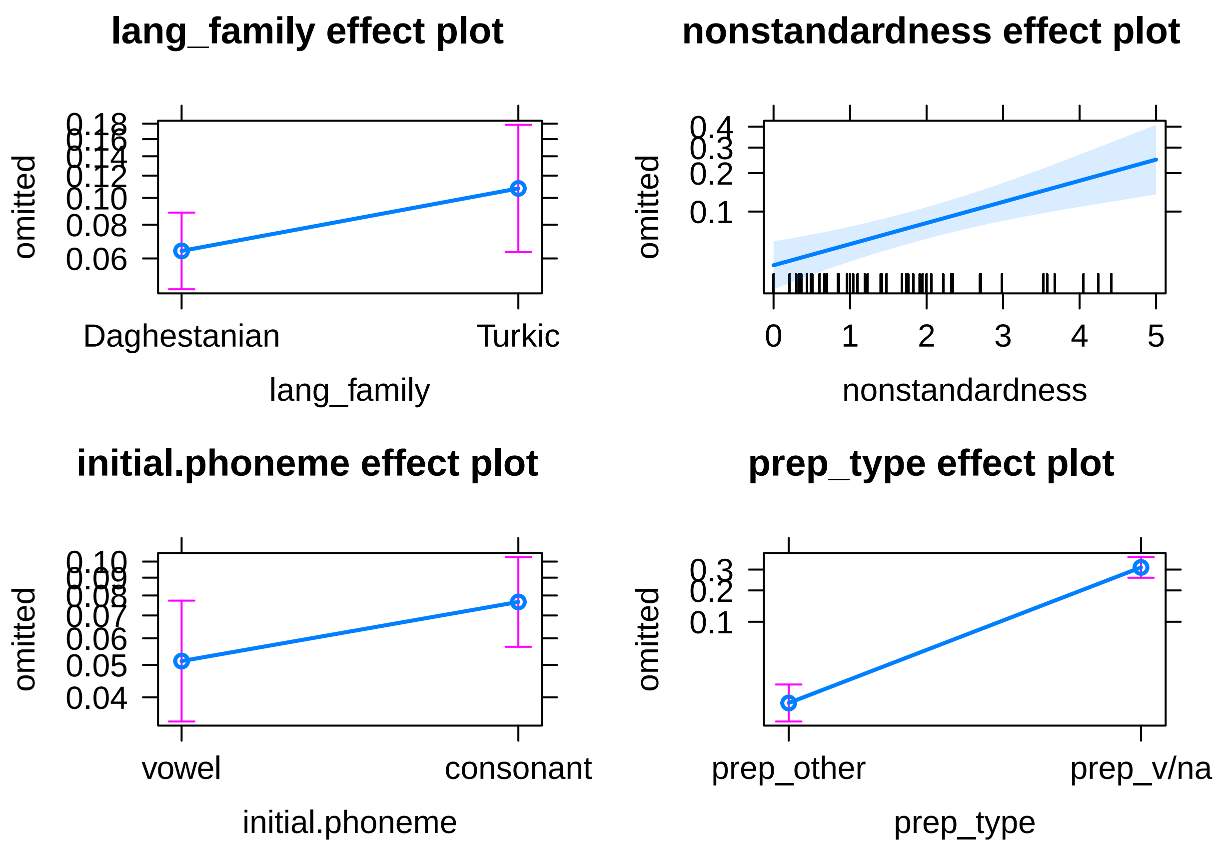
```r
drop1(glmer.res4, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## omitted ~ lang_family + nonstandardness + initial.phoneme + prep_type +
##     (1 | respondent)
##                 npar    AIC    LRT   Pr(Chi)
## <none>               1584.1
## lang_family        1 1585.1   3.02 0.0820835 .
## nonstandardness    1 1597.1  14.95 0.0001102 ***
## initial.phoneme    1 1588.3   6.17 0.0130045 *
## prep_type          1 2083.7 501.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The line which has the smallest AIC is `<none>`, so this is the best model.

I visualize the obtained model with `effects` package.

```r
log_dat$initial.phoneme<- relevel(log_dat$initial.phoneme, ref = "vowel")
glmer.res4 <- glmer (omitted ~ lang_family + nonstandardness + initial.phoneme + prep_type + (1|responde
plot(allEffects(glmer.res4))
```



3.2. Does preposition drop pattern significantly correlate with nonstandardness?

As an additional observation, in the paper we note that context type and preposition type reveal the existence of three groups of speakers in the sample.

The groups are the following:
- speakers who only omit prepositions v 'in(to)', na 'on(to)' and only in core contexts*
- speakers who only omit prepositions v 'in(to)', na 'on(to)' in core and non-core contexts
- speakers who omit prepositions v 'in(to)', na 'on(to)' in core and non-core contexts and also omit other prepositions

*Core contexts are contexts where the prepositional complement is a toponym, an exact temporal location or an institution. Non-core contexts are all other contexts.

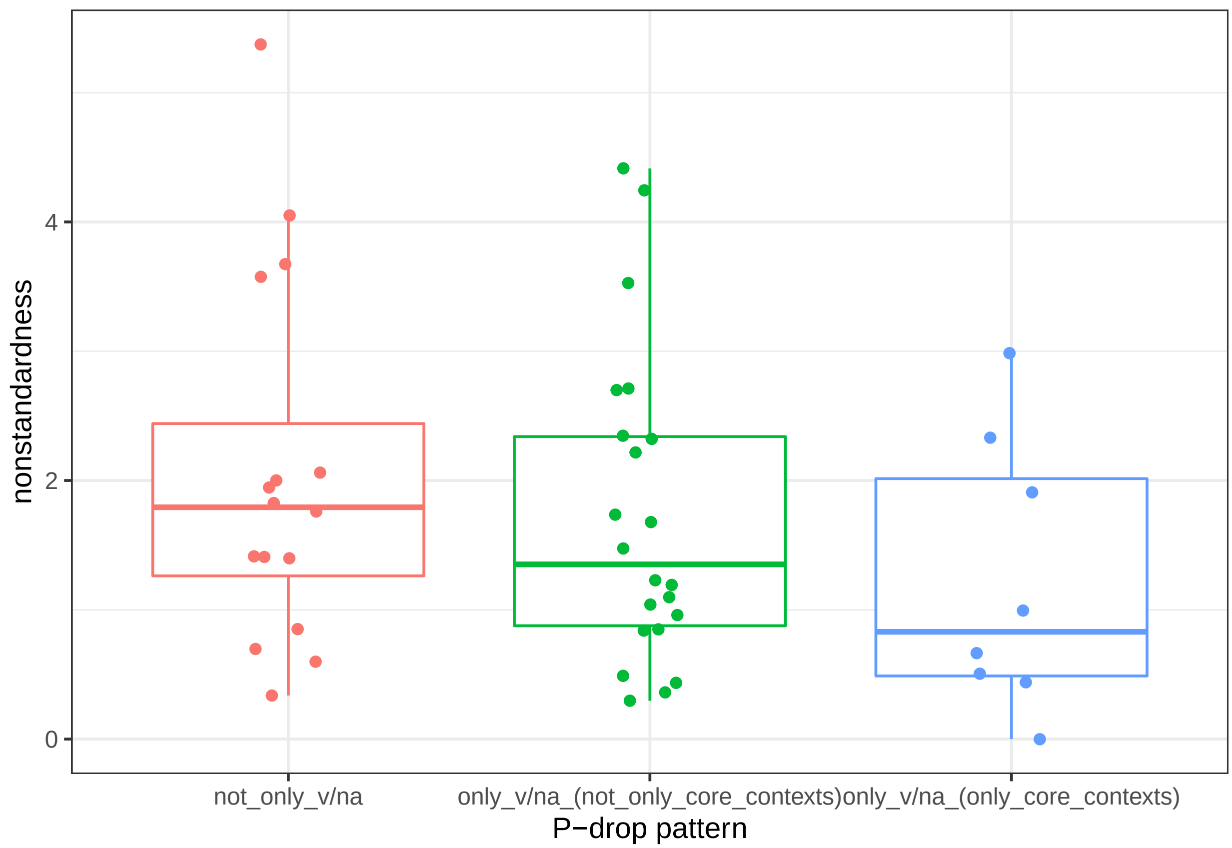
A natural question to ask at this point is whether the observed patterns correlate with the speakers' command of Standard Russian. Below I am trying to check this hypothesis.

For each speaker we annotated his/her pattern (membership in one of three groups), so I load the csv file with annotation of speakers one more time, and clean it a little (this chunk is not included because of the problem with cyrillic charachters).

We can see from the figure that the smaller the average number of non-standard features (the better the command of Standard Russian), the narrower the range of environments with preposition drop.

```
dat_speakers  %>%
  mutate(P_drop_pattern = factor(P_drop_pattern, levels = c("not_only_v/na",
                                                 "only_v/na_(not_only_core_contexts)",
                                                 "only_v/na_(only_core_contexts)"))) %>%
```

```r
ggplot(aes(P_drop_pattern, nonstandardness, color = P_drop_pattern))+
geom_boxplot(outlier.shape = NA)+
geom_jitter(width = 0.1)+
labs(x = "P-drop pattern",
     y = "nonstandardness")+
theme_bw()+
theme(legend.position = "none")
```



However, the difference between the groups does not reach statistical significance (p = 0.32, ANOVA test).

```r
aov_res <- aov(nonstandardness ~ P_drop_pattern, data = dat_speakers)
summary(aov_res)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## P_drop_pattern   2   3.72   1.862   1.179  0.317
## Residuals       43  67.90   1.579
```

This is the end of the statistical part of the research. In the paper we disuss the obtained results also in light of some cross-linguistic considerations.