Transport constructions in Russian

Olga Semenova

Problem Statement

The goal of my research is to explore the behavior of rival forms of construction describing transportation in Russian. There are two possible ways to specify the type of transport used for movement:

• construction with a noun in Instrumental case (e.g. доехать поездом, доставлять самолетами, путешествовать автобусом etc.)

• construction with a preposition and a noun in Prepositional case (e.g. добираться на поезде, лететь на самолете, уезжать на автобусе etc.)

Thus, the study aims to find out whether the choice of the construction depends on certain specific properties of the context, which it used in, or not.

Hypothesis

I formulate research hypotheses as follows:

The null hypothesis of the study: There is no association between the choice of the construction and certain contextual variables (such as properties of the verb in the construction, number of the noun, length of sentence and year of creation).

The alternative hypothesis H1: There is an association between the choice of the construction and certain contextual variables.

Regarding numerical variables, I have more precise assumptions that allow me to formulate more explicit hypotheses for them.

H2: The later the year of creation, the greater the probability that the sentence contains a prepositional construction.

H3: The longer the sentence, the more likely it is to contain a prepositional construction.

As for categorical variables, I do not have specific assumptions as to what value can be associated with which construction.

Data

For the purpose of the study I collected examples from Russian National Corpus containing constructions mentioned above. I annotated examples in order to identify contextual variables, which supposedly can influence construction choice. The dataset contents following columns:

• Verb – verbal part of the construction

• Transport – noun part of the construction

• Construction_type – one of two types of the construction (levels: INS (construction with a noun in Instrumental case) and PREP (construction with a preposition and a noun in Prepositional case))

• Normal_form - normal form of the verb

• Prefix – whether or not the verb has prefix (levels: yes for verbs with prefix and no for verbs without prefix)

• Tense – tense of the verb (levels: past (past tense), pres (present tense) and futr (future tense))

• Aspect – aspect of the verb (levels: impf (imperfective aspect) and perf (perfective aspect))

• Number – number of the transport noun (levels: plur(plural number) and sing (singular number))

• Full_context – complete sentence with the construction

• Sent_length – the length of the sentence in words

• Author – name of the author

• Header – name of the source of sentence

• Created – year of creation of the source.

```
library(ggplot2)
library(tidyverse)
library(caret)
library(party)
library(effects)
```

Data visualization

Let's take a look at the data. We can observe some of the descriptive statistics for both categorical and numerical variables.

```
df <- read.csv("https://raw.githubusercontent.com/olgasem10/Transport_constructi
ons/master/Transport_constructions.csv", encoding = "UTF-8")
summary(df[ , -which(names(df) %in% c("Full_context", "Author", "Header"))])
```

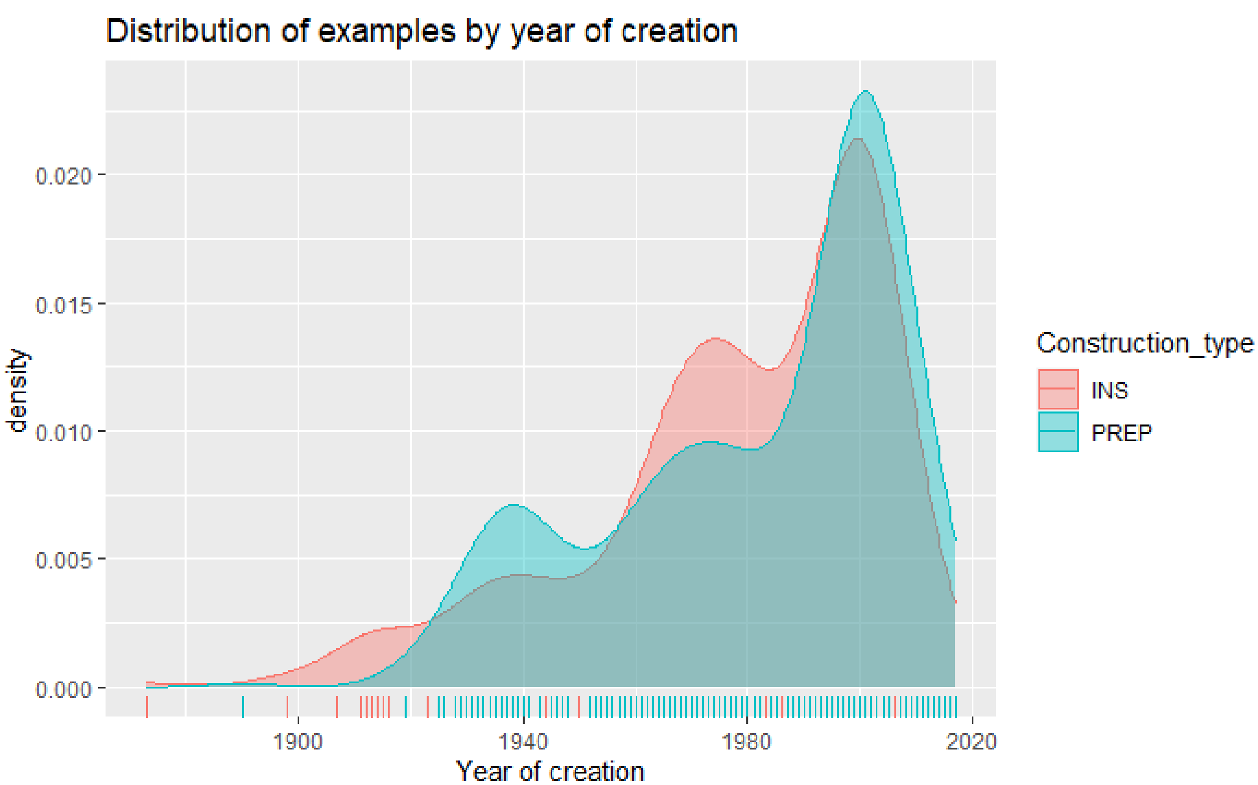
```
##       Verb           Transport    Construction_type   Normal_form  Prefix
##  ехали   : 44   поездом     :173   INS :325          ехать    :116   no :328
##  ехал    : 29   на самолете:126   PREP:401          летать   : 62   yes:398
##  летал   : 25   на поезде  : 96                     лететь   : 57
##  поехали: 20   на автобусе: 95                     поехать  : 57
##  выехали: 13   самолетом  : 78                     ездить   : 32
##  летел   : 13   автобусом  : 42                     приехать: 31
##  (Other):582   (Other)    :116                     (Other) :371
```

```
##    Tense       Aspect      Number      Sent_length       Created
##   futr: 65   impf:383   plur:102   Min.   : 3.00   Min.   :1873
##   past:524   perf:343   sing:624   1st Qu.: 9.00   1st Qu.:1964
##   pres:137                         Median :15.00   Median :1987
##                                    Mean   :17.74   Mean   :1980
##                                    3rd Qu.:22.00   3rd Qu.:2001
##                                    Max.   :67.00   Max.   :2017
##
```

To begin with, let's look at several graphs that give a more complete picture of the contents of the dataset and the distribution of data in it.

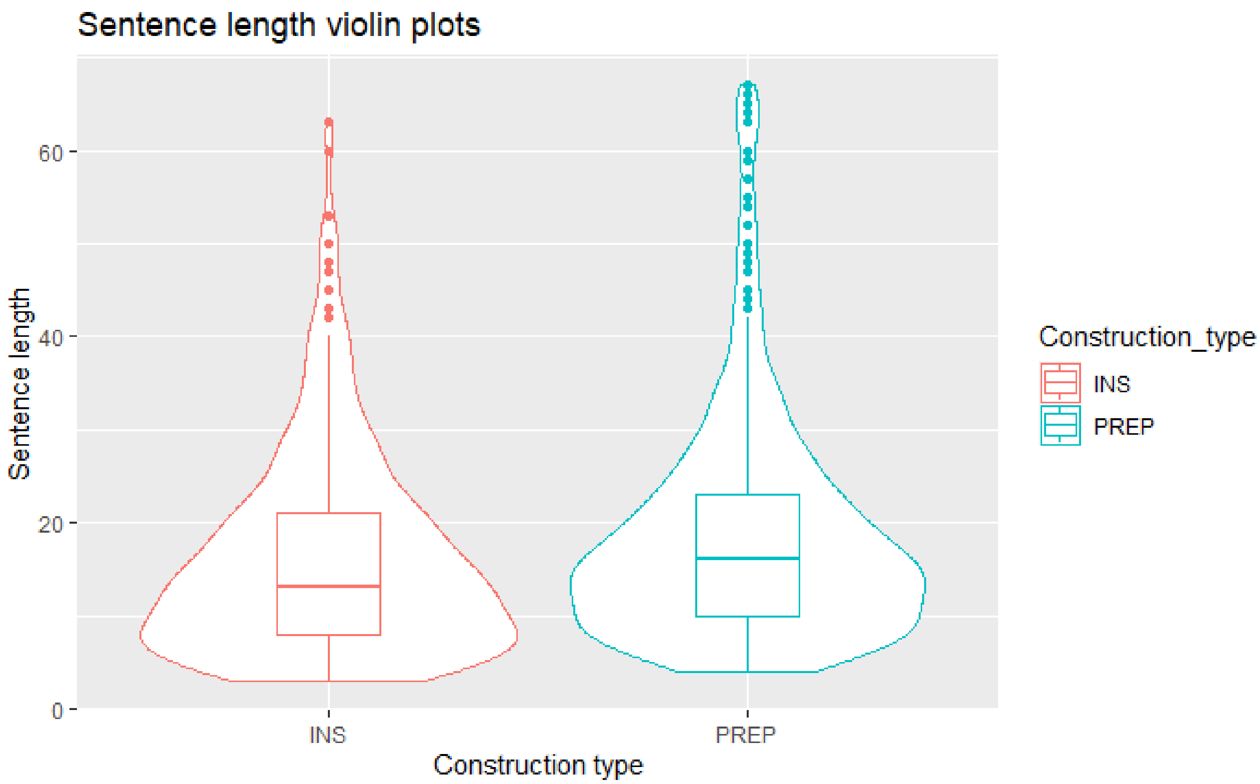
As one can see, most examples belong to the second half of the 20th century and the beginning of the 21th.

```
df %>%
  ggplot(aes(Created, fill = Construction_type, color = Construction_type))+
  geom_density(alpha = 0.4)+
  geom_rug()+
  labs(title = "Distribution of examples by year of creation",
       x = "Year of creation")
```



Violin plots showing the distribution of examples depending on the value of the sentence length allow to see that the distribution differs slightly for sentences containing different types of structures. There is a tendency that sentences containing constructions with a preposition are longer on average.

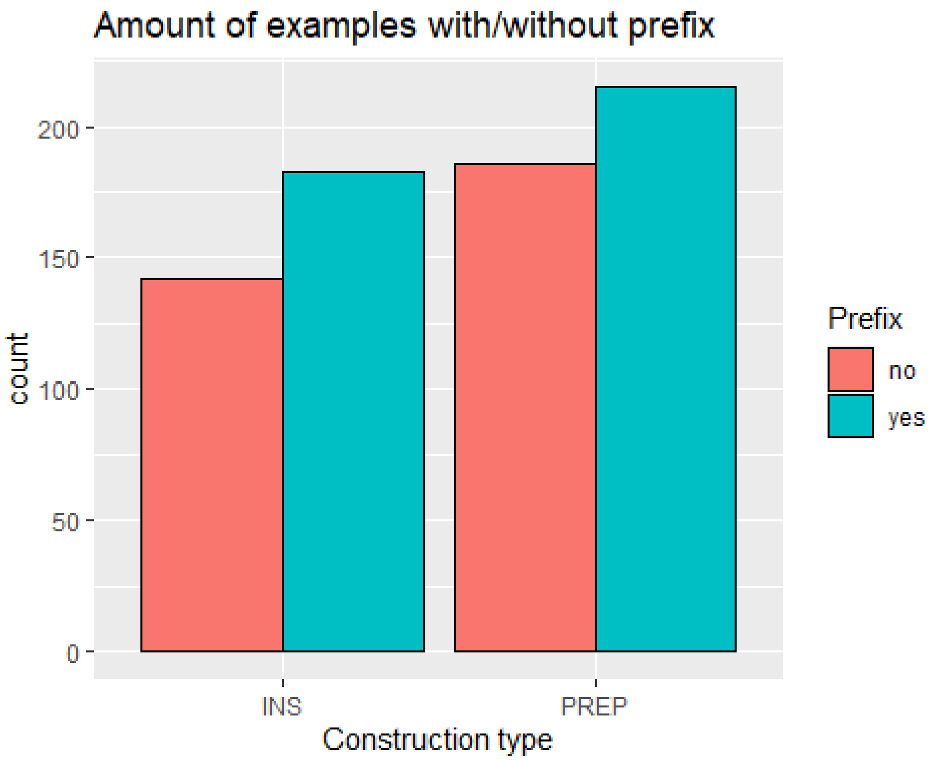
```r
ggplot(df, aes(x = Construction_type, y = Sent_length))+
  geom_violin(aes(color = Construction_type), position = position_dodge(1) )+
  geom_boxplot(aes(color = Construction_type), width = 0.25, position = position
_dodge(1))+
  labs(title = "Sentence length violin plots", x = "Construction type", y = "Sen
tence length")
```



The following histograms show the number of examples contained in the dataset with different values of categorical variables, taking into account the type of construction under study.

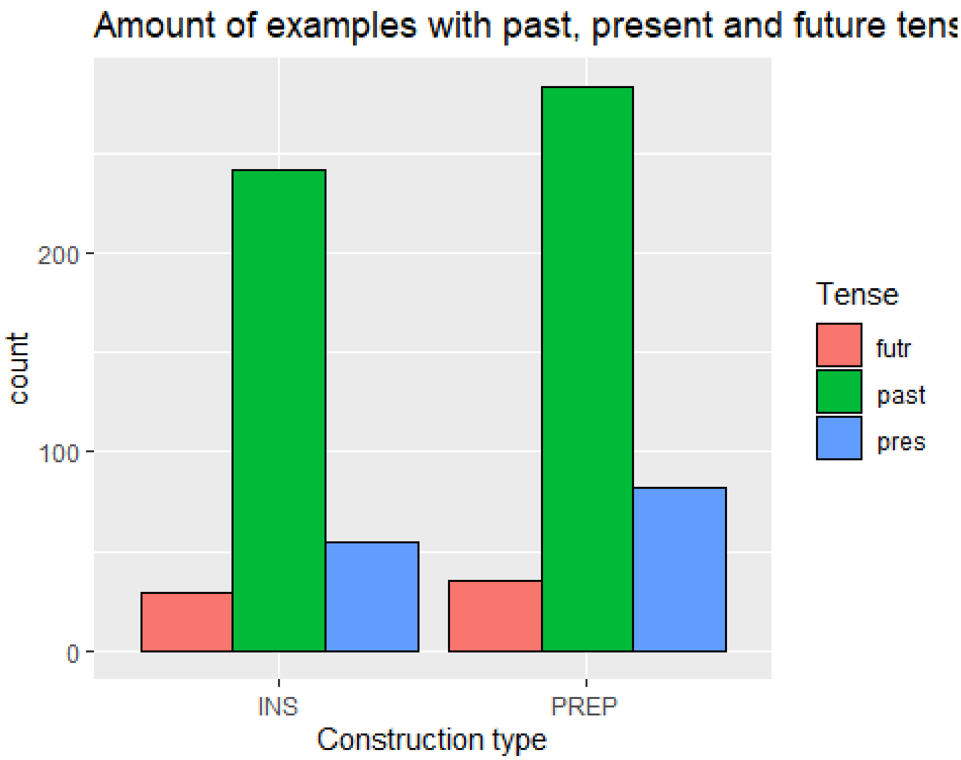
Most of the verbs in the examples have prefixes.

```r
ggplot(df, aes(x = Construction_type, fill = Prefix))+
  geom_bar(colour = "black", position = "dodge")+
  labs(title = "Amount of examples with/without prefix", x = "Construction type"
)
```



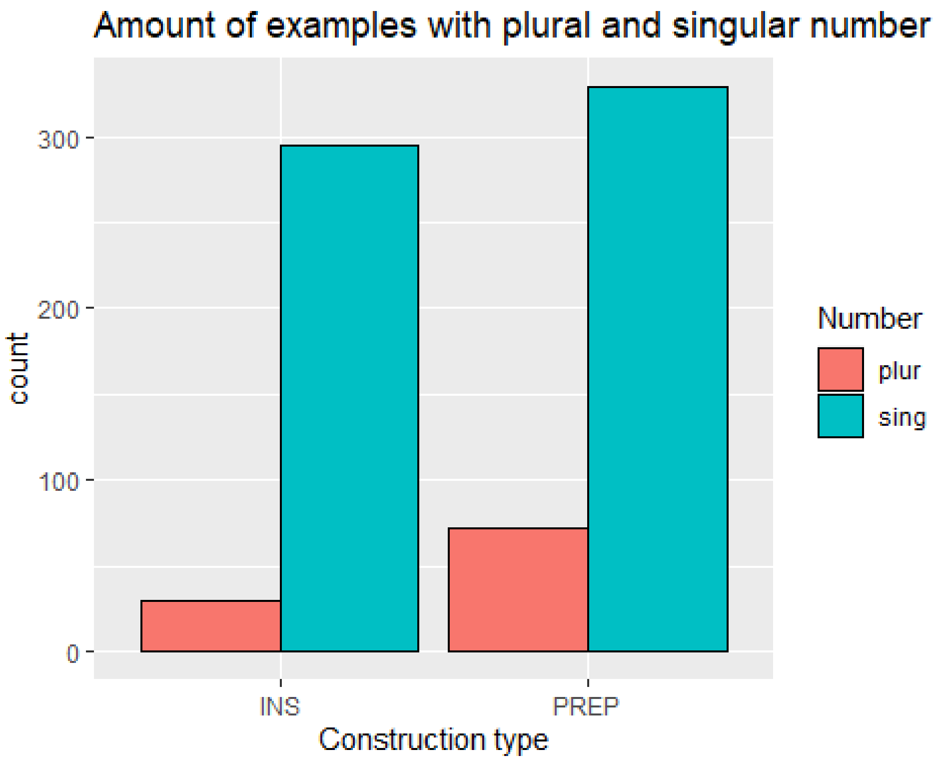
Vast majority of verbs are in past tense.

```
ggplot(df, aes(x = Construction_type, fill = Tense))+
  geom_bar(colour = "black", position = "dodge")+
  labs(title = "Amount of examples with past, present and future tense", x = "Co
nstruction type")
```



Vast majority of the transport nouns are in singular form.

```
ggplot(df, aes(x = Construction_type, fill = Number))+
  geom_bar(colour = "black", position = "dodge")+
  labs(title = "Amount of examples with plural and singular number", x = "Constr
uction type")
```



In constructions with a noun in Instrumental case the number of perfect and imperfect verbs is almost equal, while in the constructions with a noun in Prepositional case, the imperfective verbs prevail.

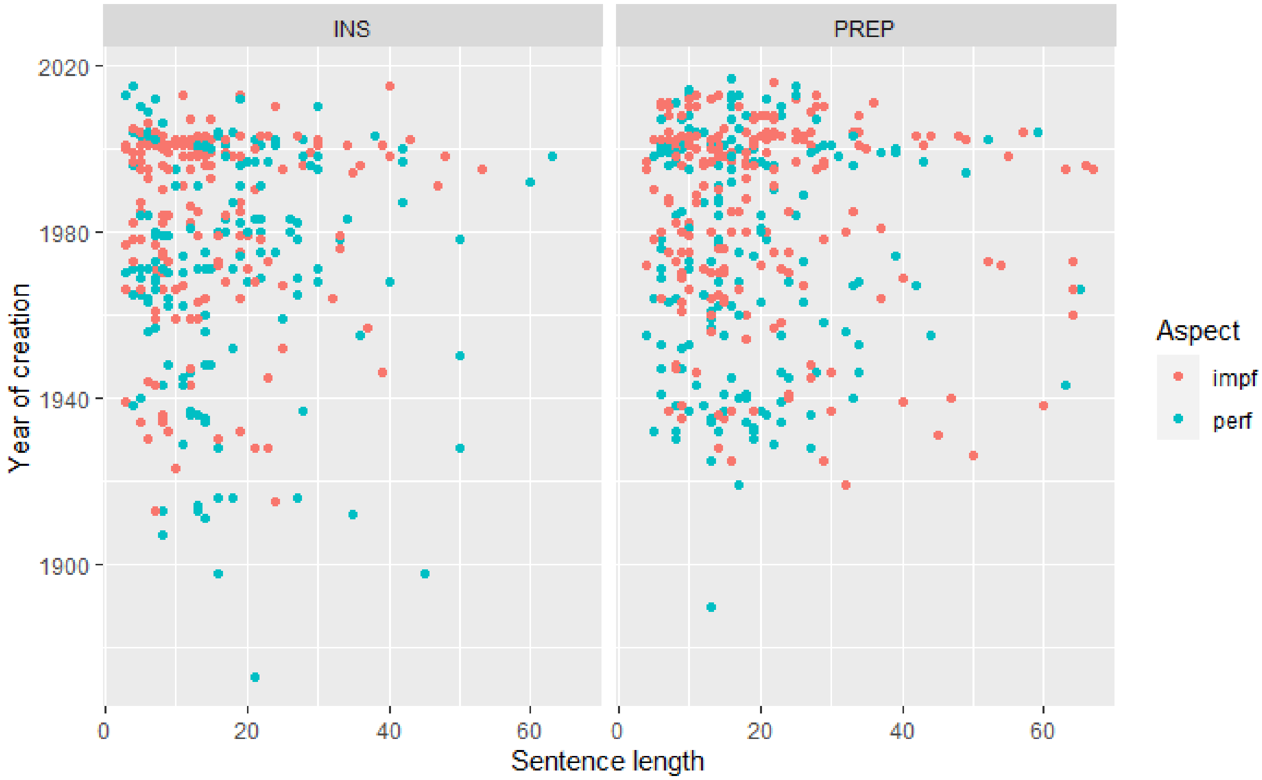
```
ggplot(df, aes(x = Construction_type, fill = Aspect))+
  geom_bar(colour = "black", position = "dodge")+
  labs(title = "Amount of examples with imperfective and perfective aspect", x = 
"Construction type")
```

The following scatterplots show the distribution of examples along the length of the sentence and year of creation, taking into account the type of construction and the value of categorical variables.

```
ggplot(df, aes(x = Sent_length, y = Created, color = Prefix))+
  geom_point()+
  facet_wrap(~Construction_type)+
  labs(x = "Sentence length", y = "Year of creation")
```


```r
ggplot(df, aes(x = Sent_length, y = Created, color = Tense))+
  geom_point()+
  facet_wrap(~Construction_type)+
  labs(x = "Sentence length", y = "Year of creation")
```

```r
ggplot(df, aes(x = Sent_length, y = Created, color = Aspect))+
  geom_point()+
  facet_wrap(~Construction_type)+
  labs(x = "Sentence length", y = "Year of creation")
```




```r
ggplot(df, aes(x = Sent_length, y = Created, color = Number))+
  geom_point()+
  facet_wrap(~Construction_type)+
  labs(x = "Sentence length", y = "Year of creation")
```

Statistical analysis

First, I examine if there is an association between the target variable and the predictors for each variable separately.

Numerical variables

To test hypothesis H2 I apply logistic regression model with one predictor (variable Created).

```
fit1 <- glm(Construction_type~Created, data = df, family = "binomial")
summary(fit1)

##
## Call:
## glm(formula = Construction_type ~ Created, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.336  -1.267   1.039   1.080   1.257
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.504044   5.679048  -1.497    0.134
## Created      0.004402   0.002869   1.534    0.125
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 998.48  on 725  degrees of freedom
## Residual deviance: 996.12  on 724  degrees of freedom
## AIC: 1000.1
##
## Number of Fisher Scoring iterations: 4
```

As one can see, p-value for this variable is over 5%, so I can not declare the variable as significant and reject null hypothesis.

Next, to test hypothesis H3 I apply logistic regression model with one predictor (variable Sent_length).

```
fit2 <- glm(Construction_type~Sent_length, data = df, family = "binomial")
summary(fit2)

##
## Call:
## glm(formula = Construction_type ~ Sent_length, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6807  -1.2194   0.9565   1.1106   1.2047
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.145243   0.136795  -1.062  0.28835
## Sent_length  0.020292   0.006639   3.057  0.00224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 998.48  on 725  degrees of freedom
## Residual deviance: 988.65  on 724  degrees of freedom
## AIC: 992.65
##
## Number of Fisher Scoring iterations: 4
```

P-value for ariable Sent_length is 0.00224, which means that I now can claim that variable Sent_length (the length of the sentence) is significant.

Categorical variables

To examine if there is a relationship between categorical variables and the target variable, I apply chi-squared tests for each variable.

```
chisq.test(df$Construction_type, df$Prefix)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
```

```
## data:  df$Construction_type and df$Prefix
## X-squared = 0.42208, df = 1, p-value = 0.5159

chisq.test(df$Construction_type, df$Tense)

##
##  Pearson's Chi-squared test
##
## data:  df$Construction_type and df$Tense
## X-squared = 1.502, df = 2, p-value = 0.4719

chisq.test(df$Construction_type, df$Aspect)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$Construction_type and df$Aspect
## X-squared = 1.4137, df = 1, p-value = 0.2344

chisq.test(df$Construction_type, df$Number)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$Construction_type and df$Number
## X-squared = 10.604, df = 1, p-value = 0.001129
```

Only the latter chi-squared test showed the p-value less then 5%, so the only variable which has the association with target variable is Number.

Model with all variables

Now I try to combine all the variables into one logistic regression.

```
fit3 <- glm(Construction_type~Prefix+Tense+Aspect+Number+Created+Sent_length, da
ta = df, family = "binomial")
summary(fit3)

##
## Call:
## glm(formula = Construction_type ~ Prefix + Tense + Aspect + Number +
##     Created + Sent_length, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8382  -1.1970   0.8238   1.1257   1.3408
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.107499   5.917867  -1.201  0.22974
## Prefixyes    0.008869   0.233067   0.038  0.96965
## Tensepast   -0.171196   0.279939  -0.612  0.54084
## Tensepres   -0.099936   0.358910  -0.278  0.78067
## Aspectperf  -0.091288   0.257146  -0.355  0.72259
```

```
## Numbersing  -0.677448   0.238734  -2.838  0.00454 **
## Created      0.003917   0.002967   1.320  0.18681
## Sent_length  0.019032   0.006707   2.838  0.00454 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 998.48  on 725  degrees of freedom
## Residual deviance: 975.83  on 718  degrees of freedom
## AIC: 991.83
##
## Number of Fisher Scoring iterations: 4
```

I used all the variables and try different interactions of them, but most of the variables remain insignificant.

Therefore, I remove most of the variables and the final version of the model contains only Number and Sent_length. As one can see, AIC (Akaike Information Criterion) is also smaller for this model.

```
fit4 <- glm(Construction_type~Number + Sent_length, data = df, family = "binomia
l")
summary(fit4)
```

```
##
## Call:
## glm(formula = Construction_type ~ Number + Sent_length, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7932  -1.1907   0.8364   1.1327   1.2360
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.506159   0.252924   2.001  0.04537 *
## Numbersing  -0.717722   0.233283  -3.077  0.00209 **
## Sent_length  0.018682   0.006677   2.798  0.00514 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 998.48  on 725  degrees of freedom
## Residual deviance: 978.60  on 723  degrees of freedom
## AIC: 984.6
##
## Number of Fisher Scoring iterations: 4
```

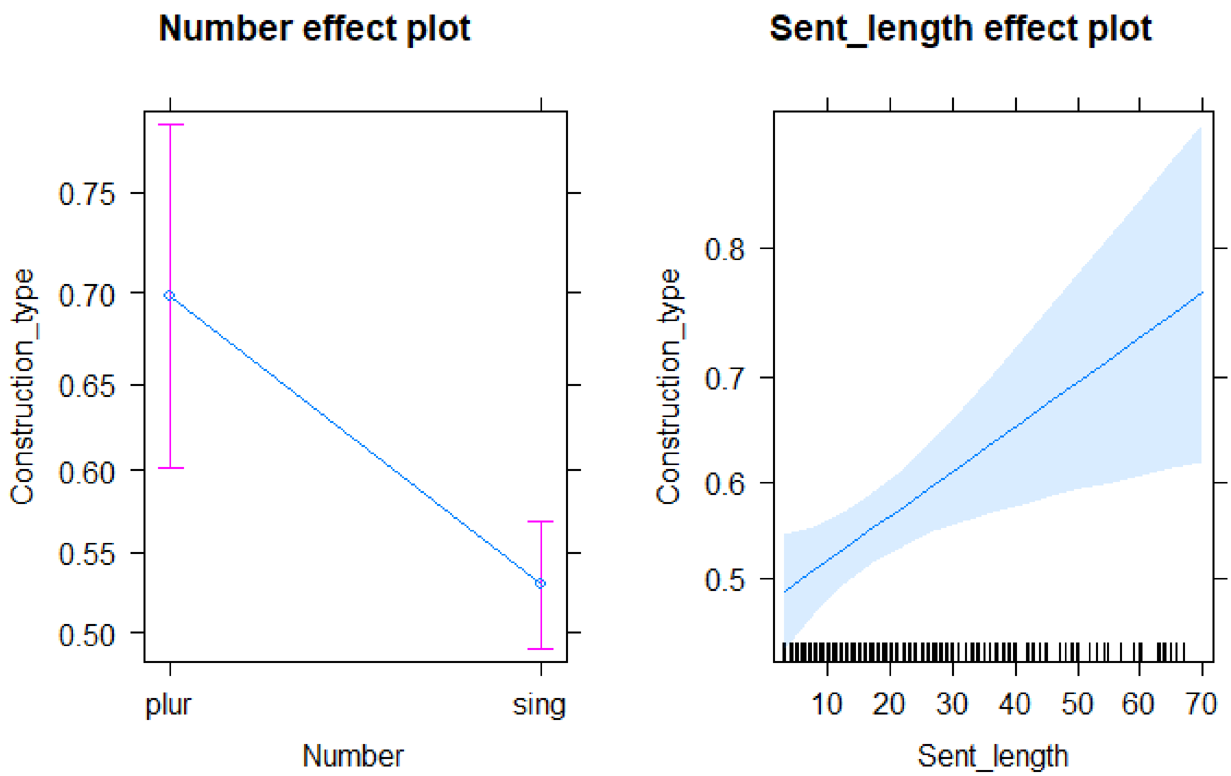
I also try to include the interaction of variables in the model, but it only worsens the result.

```
fit5 <- glm(Construction_type~Number + Sent_length + Number:Sent_length, data =
df, family = "binomial")
summary(fit5)
```

```
##
## Call:
## glm(formula = Construction_type ~ Number + Sent_length + Number:Sent_length,
##     family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8238  -1.1913   0.8372   1.1327   1.2340
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.459428   0.414869   1.107    0.268
## Numbersing             -0.664740   0.439737  -1.512    0.131
## Sent_length             0.021141   0.018635   1.134    0.257
## Numbersing:Sent_length -0.002827   0.019962  -0.142    0.887
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 998.48  on 725  degrees of freedom
## Residual deviance: 978.58  on 722  degrees of freedom
## AIC: 986.58
##
## Number of Fisher Scoring iterations: 4
```

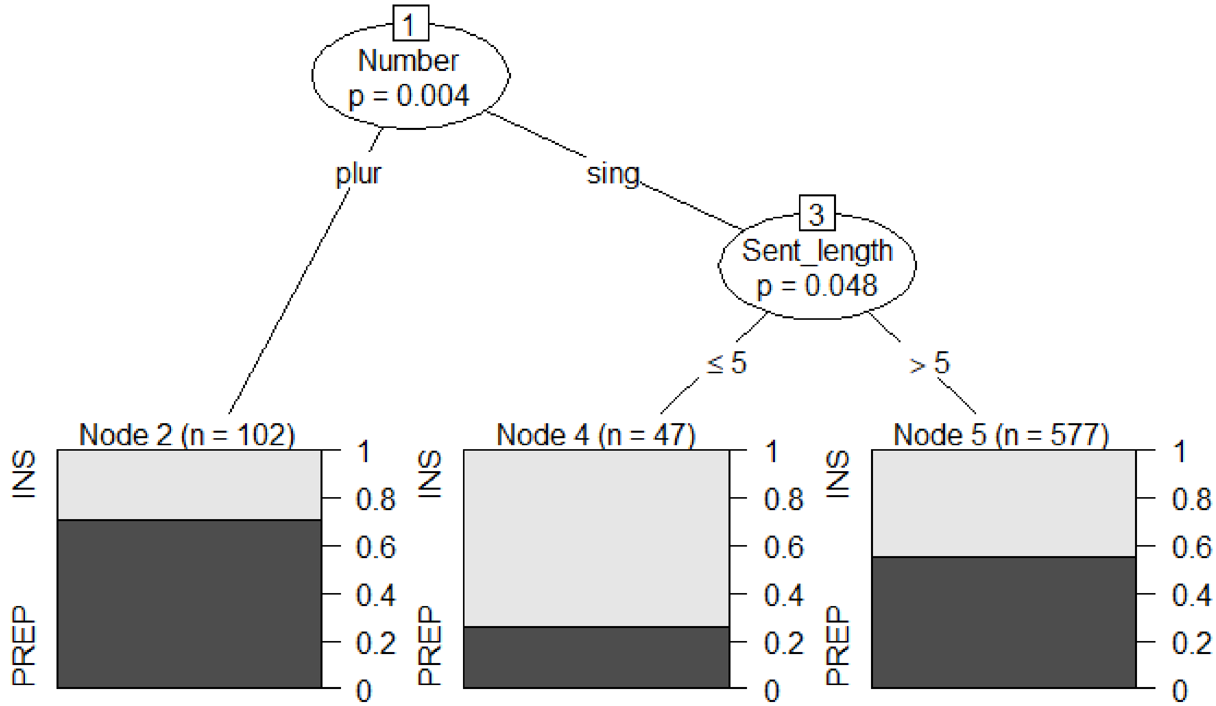
Following graps are effect plots for variables Number and Sent_length.They show how predicted probabilities value changes with the value of the variables involved in the model.

```
plot(allEffects(fit4))
```



I also use the decision tree model to further check the significance of variables. The division into classes took into account the same variables.

```
tree <- ctree(Construction_type~Number+Sent_length+Aspect+Tense+Prefix, data = d
f)
plot(tree)
```



Prediction

To test the predictive capabilities of the model, I divide the data into test and train and predict the values for the test sample

```
data1 <- df[order(runif(nrow(df))),]
split <- createDataPartition(y = data1$Construction_type, p = 0.9, list = FALSE)
train_set <- data1[split, ]
test_set <- data1[-split, ]

fit6 <- glm(Construction_type~ Number + Sent_length, data = train_set, family =
"binomial")

response <- predict(fit6, newdata=test_set, type="response")
scores <- data.frame(response=response,
                         construction_obs=test_set$Construction_type,
                         stringsAsFactors=FALSE)
v <- rep(NA, nrow(scores))
v <- ifelse(scores$response >= .5, "PREP", "INS")
scores$construction_pred <- as.factor(v)

confusionMatrix(data = scores$construction_pred, reference = scores$construction
_obs, positive="INS")

## Confusion Matrix and Statistics
##
##           Reference
## Prediction INS PREP
```

```
##       INS   13   11
##       PREP  19   29
##
##                Accuracy : 0.5833
##                  95% CI : (0.4611, 0.6985)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 0.3626
##
##                   Kappa : 0.1346
##
##  Mcnemar's Test P-Value : 0.2012
##
##             Sensitivity : 0.4062
##             Specificity : 0.7250
##          Pos Pred Value : 0.5417
##          Neg Pred Value : 0.6042
##              Prevalence : 0.4444
##          Detection Rate : 0.1806
##    Detection Prevalence : 0.3333
##       Balanced Accuracy : 0.5656
##
##        'Positive' Class : INS
##
```

By dividing into test and train several times, I got values ranging from 56.7 to 65%. Given that the class ratio is 45% and 55%, we can say that the model, although it shows low quality and tends to predict PREP in most cases, still outperforms the model just predicting a larger class.

Linguistic interpretation

Having examined and compared the results of the models, I can draw some linguistic conclusions.

The hypothesis about the time of creation (H2) was not confirmed, which means that on the basis of our dataset we cannot make a claim that over time, one construction began to be used less often and the other more often.

The hypothesis about the length of the sentence (H3) was confirmed, which may mean that shorter sentences also contain a shorter form of the construction under study, that is, a form without an preposition.

As for the study of the properties of the verb, such as its tense, aspect and presence or absence of prefix, none of these properties showed a statistical correlation with the choice of construction, which means that the construction is chosen regardless of these properties of the verb.

Having examined the behavior of a variable containing a number of the transport noun, I can claim that the plural form tends to occur more often in constructions with prepositions than in constructions without prepositions. That is, constructions such as ездить на поездах in the Russian are more common than ездить поездами.

Conclusion

After analyzing and summarizing the results of the work of different models, I can conclude that when considering the choice of the construction describing transportation in Russian, the length of the sentence and the number of the transport noun show a significant correlation. But at the same time the overall quality demonstrated by the models based on these variables nevertheless remains low, which leaves room for possible further research and search for other properties of the context that influence the choice of one of these rival forms.