I. Introduction

In this paper, I would like to describe the predictability experiment. The experiment represents the second step for the Russian Sentence Corpus, which was studied at [Laurinavichute et al., 2016].

Participants started with a blank screen and were asked to type any word. The participant had to guess the second word in such a way that the resulting phrase was a possible word combination in Russian. The script then would replace the word typed by the participant by the first actual word from one of the 144 sentences.

II. Data

- 144 modified sentences were used in a predictability norming study
- 750 native speakers participated in the predictability norming study
- We included data from every participant that made more than 20 guessing attempts out of the total number of 1362 words in the corpus
- 65% of words had low predictability, 9% average, and 23% of words had high predictability
- the whole Russian sentence corpus contains 468 nouns (34%), 282 verbs (21%), 126 adjectives (9%), 52 adverbs (4%), and 434 (32%) pronouns and function words.
- the experiment is at http://tayrinn.github.io/

III. Hypotheses

In the current study, I would like to examine, what determines the predictability of the word. There are some hypotheses, which I text with R and visualization:

1. In the beginning of the sentence, the participants' chance of guessing the actual word was close to **zero**, but it should be improved as they approached **the end** of the sentence although it remained low -> ***the dependence from word position***
2. Function words (e.g. conjunctions) are easy to guess, thus they should have **high predictability** and verbs do not. Because they are the main part of sentence and a sentence is built around the verbs -> ***the dependence from POS***
3. If the word has a high frequency, then it will have a high predictability -> ***the dependence from word frequency***

IV. Parameters

From our table, we take the following parameters for studying:

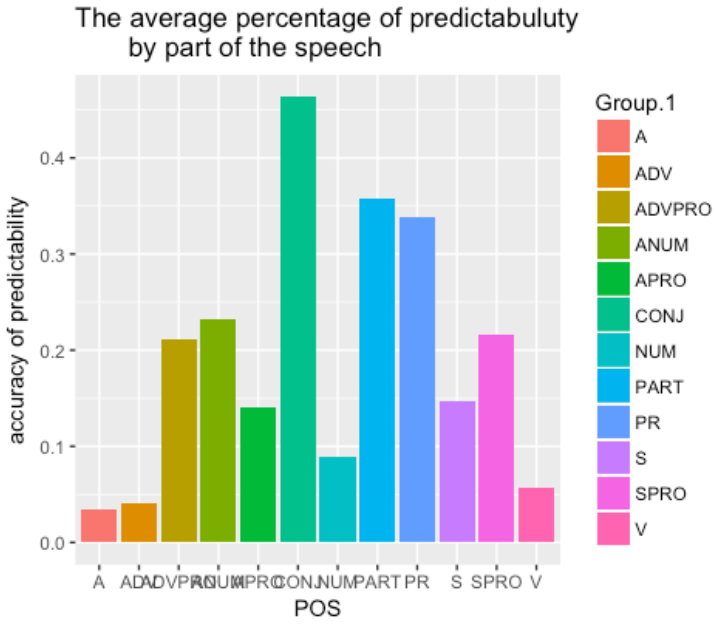
1. **word.id** – word
2. **word.serial.no** – word position
3. **POS** – part of speech
4. **fl –** if the word is the first in the sentence -> 1, else – 0
5. **X0** – number of times, when the word form) was not guessed
6. **X1** – number of times, when the word form) was guessed
7. **average accuracy** – percentage of predictability
8. **rsc_f_v_imp** – frequency of word from the subcorpus with the removed ambiguity [Lyashevskaya, Sharov, 2009]
9. **base.form** – if the word is in its base form -> 1, else – 0
10. **ambig** – if the word has morphological ambiguity -> 1, else – 0

Table 1. Data example

word.id	word.serial.no	POS	fl	X0	X1	average.accuracy	rsc_f_v_ipm
аварии	2	S	0	66	0	0.000000000	3.40
австралию	3	S	0	17	0	0.000000000	3.74
аквариуме	8	S	0	22	63	0.741176471	2.72
актуальные	6	A	0	21	0	0.000000000	1.87
американским	9	A	0	23	0	0.000000000	6.29
аптечку	7	S	0	41	22	0.349206349	0.34
атмосфера	4	S	0	16	0	0.000000000	11.23
атрибутику	5	S	0	14	0	0.000000000	0.17
багажнике	2	S	0	62	0	0.000000000	1.87
бампера	8	S	0	16	0	0.000000000	0.34
банке	8	S	0	4	20	0.833333333	16.50
банку	4	S	0	14	2	0.125000000	18.54

V. Results

a) The dependence of predictability from POS



Plot 1

From the plot 1, we can conclude that ***Function words*** such as conjunctions and participles have high level of predictability. But this table does not show us the dependence of POS from accuracy of predictability. Thus, we need to use some statistical models.

Let's use Generalized Linear Model to test the dependence of predictability from POS

```
Call:
glm(formula = average.accuracy ~ POS, data = p)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.46319  -0.14692  -0.05721   0.03603   0.89950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.033836   0.021927   1.543   0.1231
POSADV      0.007519   0.042907   0.175   0.8609
POSADVPRO   0.176987   0.081716   2.166   0.0305 *
POSANUM     0.197717   0.138097   1.432   0.1525
POSAPRO     0.106249   0.038662   2.748   0.0061 **
POSCONJ     0.429359   0.038427  11.173  < 2e-16 ***
POSNUM      0.055967   0.107865   0.519   0.6040
POSPART     0.324646   0.050461   6.434 1.89e-10 ***
POSPR       0.304911   0.030877   9.875  < 2e-16 ***
POSS        0.113087   0.025521   4.431 1.04e-05 ***
POSSPRO     0.182893   0.036589   4.999 6.76e-07 ***
POSV        0.023378   0.026705   0.875   0.3815
---
```

As we can see there are no real dependence from POS.

b) *In the next step, I have normalized our Frequency and Predictability values*

```
p$normAcc <- lapply(p$average.accuracy, function(x) x/sum(p$average.accuracy))
p$normFreq <- lapply(p$rsc_f_v_ipm, function(x) x/sum(p$rsc_f_v_ipm))
```

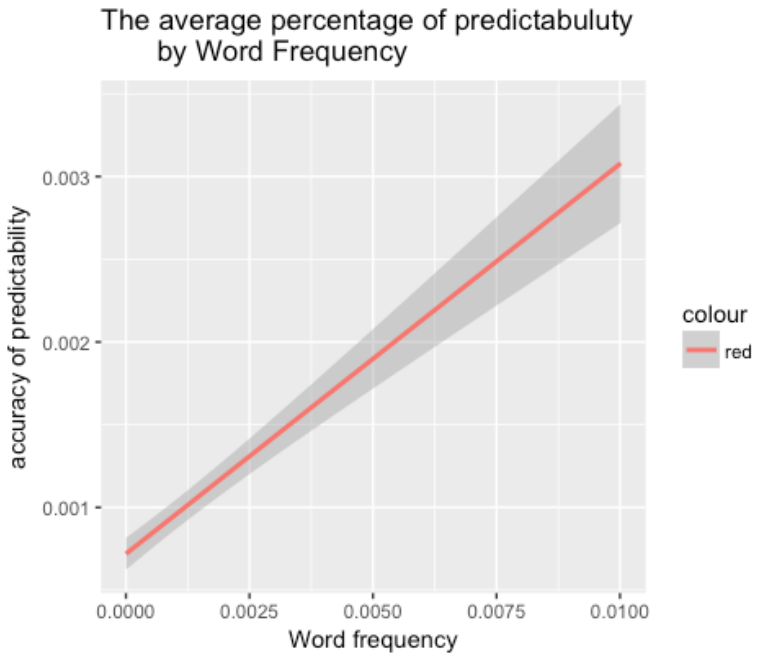
I take the values of each word frequency and predictability and divide to by the total sum of frequency and predictability

From the Pearson's correlation we can see, that the correlation is significant.

```
        Pearson's product-moment correlation

data:  as.numeric(p$normFreq) and as.numeric(p$normAcc)
t = 12.043, df = 1060, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2928846 0.3987613
sample estimates:
      cor
0.3469277
```

Let us visualize its correlation

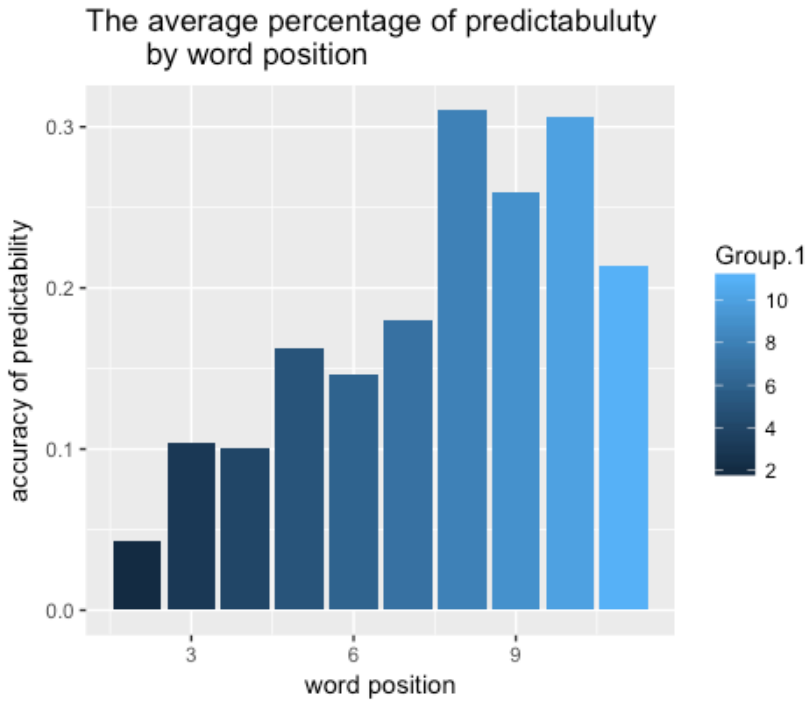


Plot 2

To sum up, the plot 2 shows us that the higher word frequency is, the higher the predictability

c) The dependence from word position

The plot 3 shows that the closer the word is to the end, the higher its predictability



Plot 3

From the Perason's correlation and Generalized Linear Model we can provide a proof that the correlation it statistically significant.

```
Call:
glm(formula = average.accuracy ~ word.serial.no + POS, data = |

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.47478  -0.12257  -0.05898   0.04632   0.91714

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.108617   0.027558  -3.941 8.64e-05 ***
word.serial.no  0.024961   0.003069   8.133 1.17e-15 ***
POSADV        0.012379   0.041639   0.297  0.76631
POSADVPRO     0.186311   0.079301   2.349  0.01899 *
POSANUM       0.132158   0.134245   0.984  0.32512
POSAPRO       0.110733   0.037520   2.951  0.00323 **
POSCONJ       0.408670   0.037375  10.934  < 2e-16 ***
POSNUM        0.058635   0.104667   0.560  0.57546
POSPART       0.312708   0.048987   6.383 2.59e-10 ***
POSPR         0.292306   0.030002   9.743  < 2e-16 ***
POSS          0.131342   0.024866   5.282 1.55e-07 ***
POSSPRO       0.189786   0.035515   5.344 1.12e-07 ***
POSV          0.042792   0.026023   1.644  0.10039
---
```

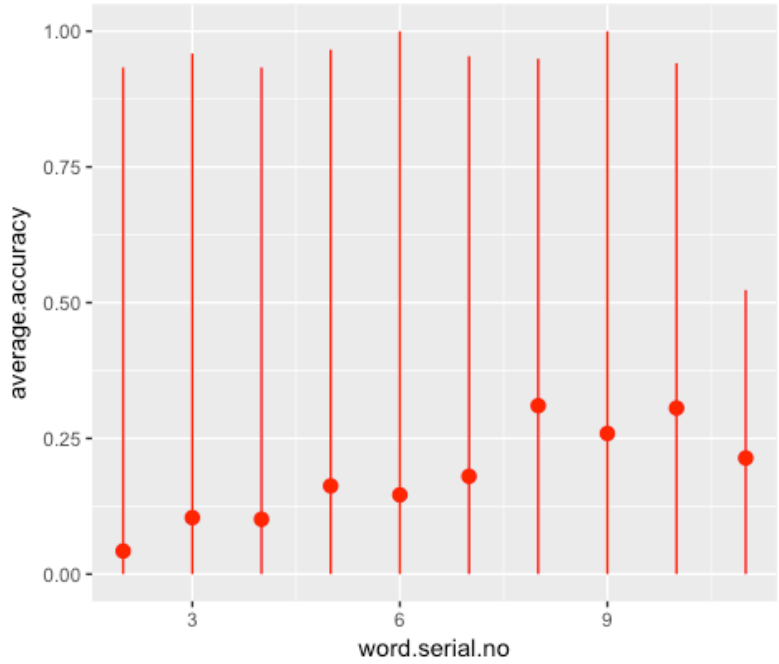
```
        Pearson's product-moment correlation

data:  p$average.accuracy and p$word.serial.no
t = 9.7171, df = 1060, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2297890 0.3402924
sample estimates:
      cor
0.2859912
```

I built the plot 4 with the confidence intervals.

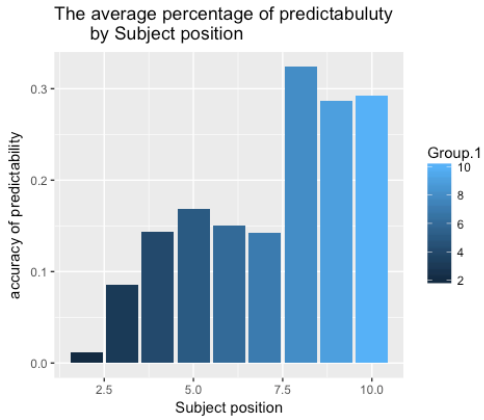
The plot 4 shows that:

1) First, disjoint confidence intervals confirm our conclusion about statistically significant differences.
2) Secondly, for the attentive observer, this graph also provides additional information about the details of our data: it is easy to see that the confidence is wider for all positions.

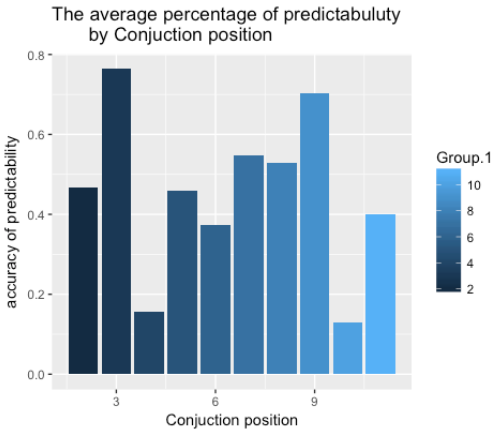


Plot 4

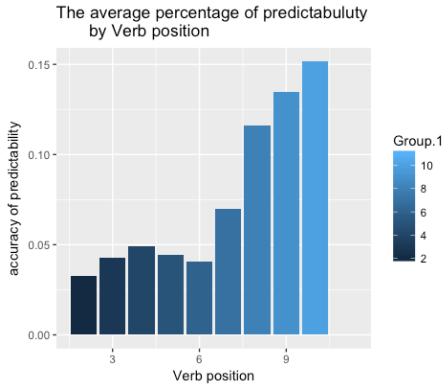
d) ***In the next steps I examine the correlation between predictability and several Part of Speech***



Plot 5



Plot 6



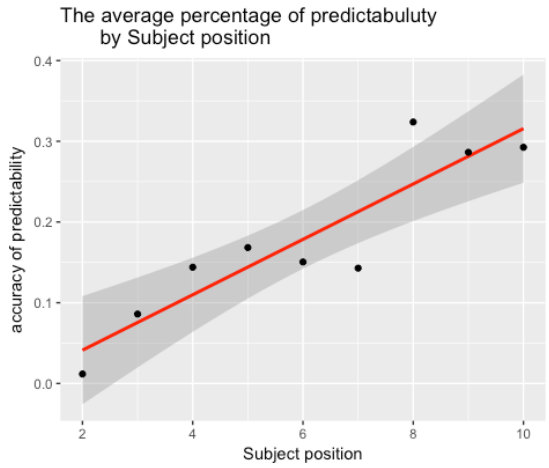
Plot 7

As we can see Conjunction position (plot 6) has no dependence of predictability, but Subjects' position have (plot 5).

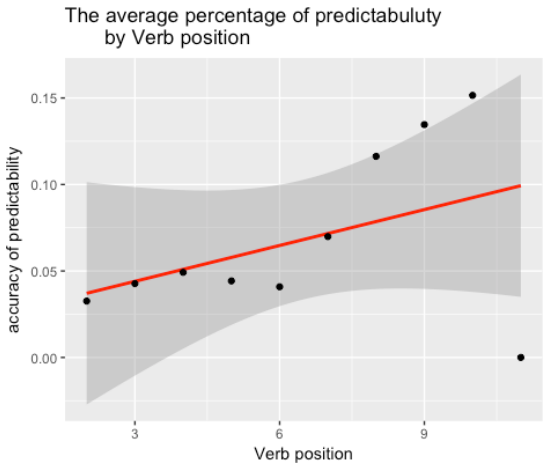
There are several explanation:

1) we have unregulated sample
2) ***Conjunctions have a high frequency.*** Thus, there is no matter where they are in the sentence

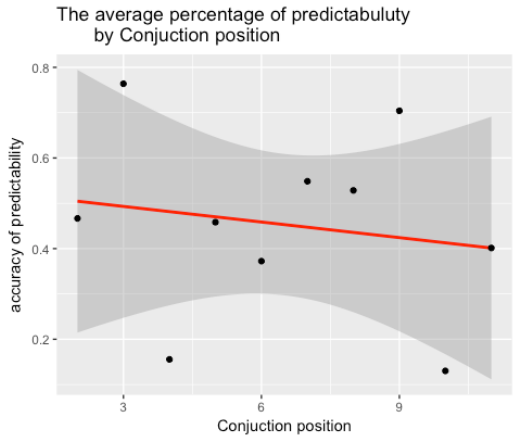
And there are also simple linear regression models, which shows, that subject position is significant *(plot 8),* verb position has no real significance *(plot 9)* and there no dependency from conjunction position *(plot 10)*



Plot 8

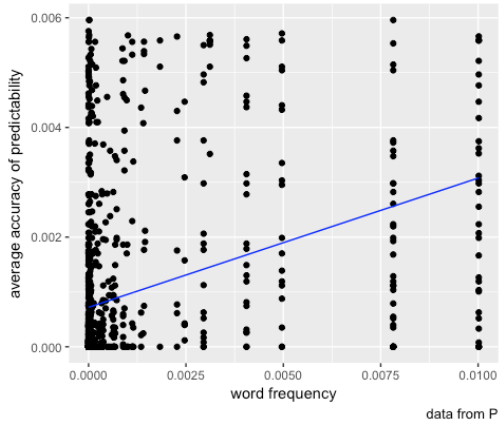


Plot 9

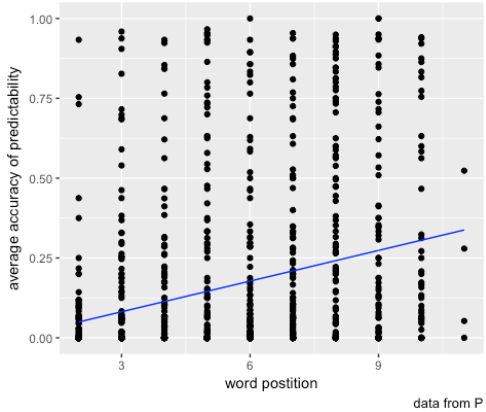


Plot 10

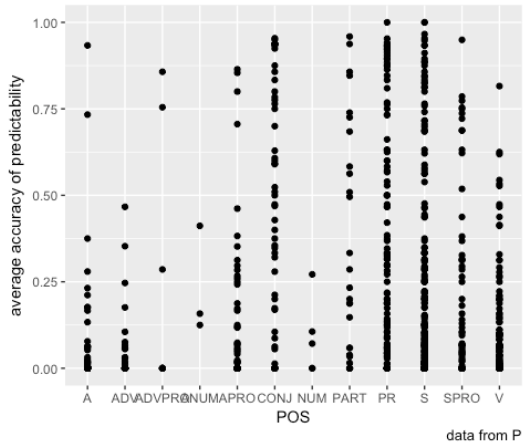
e) In this step I use mixed effects models



Plot 11



Plot 12



Plot 13

In these plots, we can see a strong correlation between predictability and word frequency *(plot 11)* and between word position *(plot 12)*. And one more time: there are no correlation between POS and predictability *(plot 13)*.

VI. Summary

In this paper I examined, what exactly determines the predictability of word.

1) From the data, that I have found, **there are no dependence from POS. Thus, the second hypothesis is not approved.**
2) On the other hand, there is a strong dependence from *word frequency and word position*. Thus, the first and the third hypotheses are proved.
    a. However, positions of some part of speech (e.g. Conjunction) has no correlation with predictability.
    b. This can be explained in the following way: conjunctions have the high frequency itself, so there is no need to examine its position.
3) Further we should study other parameters such as:
    a. ambiguity
    b. word inflection
    c. simultaneity of different parameters

Sources:

1. Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. European Journal of Cognitive Psychology, 16(1/2), 262-284.
2. Anna K. Laurinavichyute, (2016). Irina A. Sekerina, Svetlana Alexeeva, Kristina Bagdasaryan and Reinhold Kliegl. *Abstracts*. Russian Sentence Corpus. Benchmark measures of eye movements in reading in Cyrillic. The Seventh International Conference On Cognitive Science (COGSCI-2016), Svetlogorsk, Russia.
3. Russian National Corpus
4. Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) //URL: http://dict. ruslang. ru/freq. php. – 2009.