Petr Filchenkov

Kirill Konovalov

Hypothesis

The aim of this research is to analyze the representation of such Russian comparative function words as “будто”, “словно”, “а-ля”, “аки”, “как”, “подобно” in the Russian National Corpus.

The hypothesis is that each of these words tends to be used in texts of specific register and genre. Such conjunctions as “словно” and “подобно” are supposed to occur mostly in written belles-lettres texts, while “как” and “будто” seem to be stylistically neutral and should have even distribution in texts of all genres and registers.

Research design

In terms of statistics, we are going to test two null hypotheses in our research separately. Hypothesis one is that the usage of words under consideration should depend only on genre. Hypothesis two is about their dependance on text register.

There are some other variables which can be taken into account such as text date and context. According to the null hypotheses, they should not possess any significant influence on speaker’s choice of comparative conjunctions or prepositions.

As for statistical tools, we will use the logistic regression model and chi-squared test to check our predictions.

Data collection method

The data for this research is based on the Russian National Corpus which includes several smaller corpora.

First of all, we are going to examine the words of the study in the written corpus and how they are presented according to their genres, context and date of creation. In this set of experiments we are going to compare the usage of the word "как" (as the most universal) with the other words so there will be five datasets with about 500 words (250 occurrences for each word) gained from written corpus search queries.

Secondly, we intend to combine the written and spoken corpora in one dataset excluding their genres since they are different. It is worth noting that we are going to add the words proportionally according to their distribution. We will not only look at their register but also at their context and date of creation as well. Comparing the usage of the word "как" with other words, we will test null hypothesis two. The maximum size of each dataset will be 500 words (250 occurrences for each word) but some words do not have enough occurrences and their dataset will be reduced accordingly.