Church Slavonic words distribution in historical secular texts

Salnikov Egor

Initial hypothesis

The initial hypothesis of the research is the assumption that some lexemes of Church Slavonic language were somehow grouped in secular texts and became a part of larger semantic chunks than just an isolated word.

The main goal of the project is to test whether the presence of Church Slavonic words in the sentence could affect the probability of another Church Slavonic word (target word) in the same sentence. In Modern Russian such lexemes are often a part of idiomatic phrases and it should be interesting to trace idiomatisation process.

To achieve this goal two additional intuitive hypothesis should also be tested:

1. The distribution and frequency of Church Slavonic words reduces over time. The time-period of the text should also be a factor affecting the probability of the presence of Church Slavonic words in the text.

2. Thematic domain of the text is another factor affecting the probability of Church Slavonic words presence.

Method

To test this hypothesizes I will be using text collection extracted from RNC Middle Russian Corpus.

The probit model:

P - probability of target word presence in the sentence
N - the number of another Church Slavonic words in the sentence
T - thematic domain of the text
C - century of the text

P $= \mathrm{b}_0 + b_1 N + b_2 T + b_3 C$

Formal hypothesis

H_0: variables are insignificant
$b_1 = 0$
$b_2 = 0$
$b_3 = 0$

H_A: some variable(s) is/are significant

$b_1 \neq 0$

or/and

$b_2 \neq 0$

or/and

$b_3 \neq 0$

Statistical methods: LRtest, Macfadden pseudo-R^2

Data

To make the experiment more explicit I decided to concentrate only on synonymous pairs with one word from Church Slavonic and other being more universal (like нощь/ночь, дщерь/дочь, очи/глаза, etc). Such conditions reduce the chance of unexpected external factor which could possibly affect the results.

Also I plan to use the list of most frequent Church Slavonic words, extracted from Church Slavonic thematic domain from RNC Middle Russian Corpus.

The dataset will contain the collection of target words and their universal synonyms. The dataframe created from such samples will contain columns: century, oth.words, th.domain, and targ.word.

century column consists of century-id (15, 16, 17) of the text the sentence was taken from

oth.words shows the number of non-target Church Slavonic words in the sentence

th.domain – thematic domain of secular text

targ.word – the target-word from synonymous pair