A Case Study of High-Frequency Dictionary Collocations in a Spoken Corpus

Maria Khokhlova

St Petersburg State University

RASLAN 2021

Outline

- Introduction
- Spoken Russian Corpora
- Methods
- Results
- Conclusion
- Acknowledgement

Introduction

- collocations in written corpora vs collocations in spoken corpora;
- oral data remain mainly outside the scope of works on collocations;
- questions:
  1. do high-frequency collocations collected from dictionaries occur in spoken texts?
  2. do their frequencies differ from the ones in written corpora?

Spoken Russian Corpora

- Spoken Corpus of Russian (SCR), a part of the Russian National Corpus (RNC): 13.4 mln tokens;
- project "Night Dream Stories and Other Corpora of Oral Speech": 14,000 tokens;
- Corpus of Russian Oral Speech: 22,000 tokens.

Methods

- Gold Standard of Russian collocations;
- 6 different Russian dictionaries:
  - two explanatory dictionaries, i.e. the Dictionary of the Russian Language; the Large Explanatory Dictionary of the Russian Language;
  - three collocation dictionaries [Borisova 1995; Oubine 1987; Reginina, Tjurina, Shirokova 1980];
  - an online dictionary based on the Russian National corpus [Kustova 2008].
- dictionary indices: 5 dictionaries vs 2 dictionaries;
- adjacent vs distance collocations (e.g. *polnaya svoboda* 'complete freedom' and *polnaya i bezgranichnaya svoboda* 'complete and unlimited freedom');
- corpora: SCR and the disambiguated subcorpus of RNC.

Results

- dictionary index 5:
  - the most frequent collocate is *glubokiy* 'deep' (8 examples), while *zheleznyy* 'iron', *ostryy* 'sharp' and *polnyy* 'complete, full' show 2 examples;
  - no correlation between two distributions in corpora (0.36 according to the Spearman coefficient, p $>$0.05);
  - frequencies are small and do not differ in the corpora (V=80 according to the Wilcoxon test, p $>$0.05).
- distance n-grams:
  - collocations show low permeability;
  - the longest n-grams: *tverdaya, khotya i mgnovenno sozrevshaya uverennost'* 'firm, albeit instantly ripe, confidence'; *polnoy i ravnoy dlya vsekh svobody* 'full and equal freedom for all'.

Results

- dictionary index 2:
  - more than half of collocations from this group had no examples in corpora;
  - tend to occur rarer compared to dictionary index 5;
- distance n-grams:
  - tend to occur only in their adjacent forms (exceptions: *dlinnaya avtomatnaya ochered'* 'a long gun burst', *chrezmernoye issledovatel'skoye svimaniye* 'excessive research attitude', *bol'shoy vas poklonnik* 'a big fan of you' and *svezhaya nemetskaya gazeta* 'a fresh German newspaper');

Results: textual and syntactic characteristics

- selected collocations are more characteristic of journalistic texts (compared to fiction);
- their usage prevails in the position of the end of the clause;
- more used in plural form;
- typical for texts written by men.

Conclusion

- low occurrences in corpora;
- dictionary collocations are rare linguistic phenomena;
- spoken corpora are not sufficient enough.

Acknowledgement

The database was compiled with the support from the Russian Science Foundation (Project No. 19-78-00091). The work on collocation evaluation in spoken corpora was supported by St Petersburg State University, project No. 75254082 “Modeling of Russian megalopolis citizens’ communicative behavior in social, speech and pragmatic aspects using artificial intelligence methods”.