Viktor P. Zakharov, Candidate of Science (Philology), Associate Professor, Department of Computational Linguistics, SPbGU

Russian Corpora and Corpus Research

Abstract

Corpus linguistics is part of the computational linguistics that develops the general principles of how to build and employ, by using computers in linguistics, corpora of texts. Nowadays, linguistic research and compiling dictionaries and grammars tend to incorporate linguistic corpora. Today, to develop the intellectual software to process the texts written in a natural language we must have experimental linguistic background. All these make corpus linguistics vital for Russian science.

The author is a Russia's leading expert in the corpus linguistics. Apart from his academic and research scope of responsibilities, he is an organizer of the only in Russia conference on the corpus linguistics regularly held at SPbU.

The author has 24 publications indexed in the Web of Science and Scopus, among them 21 published during the last 5 years (2017 including). For the competition, we have selected 6 publications that focus on how to build and use the Russian corpora and computational methods to identify lexical units and semantic relations in the Russian language.

As an independent branch of science, corpus linguistics had appeared by the early 1990s. Since then, the corpus methodology has become an indispensable part of the linguistics, and all linguists inevitably incorporate corpus data in their research. Although Russia was not among the first to explore the corpus linguistics, it is nevertheless on the rise. So is the author who makes his best endeavours to contribute to the Russian corpus linguistics (publications # 1, 3, and 4).

Importantly, the author has been actively engaged in building Russian-Chinese parallel corpus (publication #3) and contributed to developing a new technology to build web-based mega-corpora (publication #4). The publication #1 is a report on the conference "Text, Speech and Dialogue", Czech Republic (as an invited lecturer).

The second group of the publications focuses on how automatically identify, by using corpus data, lexical units (key words, set expressions, lexical and semantic fields). One of the traditional and well-known methods in linguistics is a statistical analysis and distribution, that is word frequency distribution in the texts. As early as when computational linguistics appeared, scientists tried to reveal, by the word frequency distribution in the texts of a definite size, the frequency of distribution of a pair or group of words that are habitually juxtaposed. In contemporary linguistics, it is known as methods to identify collocations and set group of words by associations (#2 and 5). Simultaneously, scientists expressed an idea that this method could be extrapolated to study paradigmatic aspect of a language. In pther words, they believed that the paradigmatic relations could be based on the syntagmatic relations. However, only big text corpora made it possible to prove in practice. This is the focus of the publication #6, so as of many others that are not mentioned in the references. I must say that multilingual parallel corpora seem highly promising area for research. This is the main concern of our publication #3.