Methods and tools for development of the Russian Reference Corpus

Serge Sharoff, Russian Research Institute for Artificial
Intelligence, P.O.Box 85, 125190, Moscow, Russia,
sharoff@aha.ru

From the viewpoint of corpus linguistics, Russian is one of few major world languages lacking a comprehensive corpus of modern language use, even though the need for constructing such a corpus is growing in the corpus linguistics community both in Russia and in the rest of the world.

1. The history of development of Russian corpora

The best-known attempt to develop a comprehensive Russian corpus has been made in Uppsala. The Uppsala Corpus consists of 1 million words of fiction and non-fiction texts, so it is too small and restricted in the genre coverage for modern standards. It also lacks morphosyntactic annotations. Another attempt has been made in the Soviet Union in the mid 1980s under the heading of the Machine Fund of Russian, though it did not produce the expected outcome. There are also multiple ad hoc collections of Russian texts, but they are not balanced and representative.

2. The objective

The objective of the project is to develop the Russian equivalent of the BNC, namely a corpus of 100 million words with proportional coverage of various functional registers, POS annotation and lemmatisation (the latter is required for Russian, which has dozens of word forms for a lemma). The annotation scheme (based on TEI) also allows to mark noun phrases and prepositional phrases, because they are important for the resolution of ambiguity.

3. Problems and solutions

First, there are problems in obtaining source texts. Some sources are readily available: fiction and news texts are widely accessible via the Internet and can be legally available for the corpus. Other types of the discourse, like business or private correspondence, are hard to obtain and make available in a corpus because of legal obstacles. Yet other types of sources, like samples of spontaneous speech, are rare for technical reasons. The decision is to increase the amount of ephemera whenever possible, because news and fiction texts will take the rest of the share. Personal and business letters are subjected to an anonymization procedure with respect to names of persons and companies. Another set of problems with sources concerns the choice of diachronic sampling, because the turbulent history of Russia in the 20th century radically affected the language. For instance, according to the frequency list (Zasorina, 1977) that was compiled on the basis of texts from 1930-1960, such words as *sovetskij* (Soviet) and *tovarishch* (comrade) belonged to the first hundred of Russian words on a par with function words, but this is no longer valid in modern texts. The decision on the chronological limits of the study is different for different functional registers, for instance, fiction texts are taken from 1970, scientific texts from 1980, while news texts from 1997. Second, there are problems with resolving the ambiguity of word forms. Many word forms correspond to several lemmas and POS classes, for instance, the *pole* is an instance of three different nouns *pol* (floor), *pole* (field) and *pola* (lap). Since they have distinct morphological properties (the case, number and gender), the ambiguity can be resolved on the basis of simple syntactic analysis, like the agreement in noun phrases or between the subject and the predicate in a sentence. Yet other frequent types of ambiguity can be resolved only on the basis of semantic and pragmatic constraints: *Xranite svoi denjgi v banke* (keep your money in a bank/in a jar). Such cases of genuine ambiguity are kept in the corpus using multiple <ana> tags. Third, there are problems with the query language for accessing the corpus. Typically corpus query languages (e.g. SARA or CQP) assume the fixed order of tokens, while in Russian the order of participants in a clause is not fixed, but depends on thematic development conditions. Special operators are introduced for expressing such conditions.

4. The current state of the project

Currently tools and techniques for working with the reference corpus are tested using a corpus of 40 million words. Its sub corpus of about 1 million words of fiction texts (The Russian Standard) has automatically assigned and manually inspected POS annotations (it is available from http://corpora.yandex.ru). It can be also used for correcting POS taggers used for processing the corpus.