Basic Russian Dictionary: a Corpora-Based Learner's Dictionary of Russian

Maria Shvedova, Dmitri Sitchinava
Kyiv National Linguistic University,
Institute of Russian language, Russian academy of Sciences
e-mail: masha.shvedova@gmail.com, mitrius@gmail.com

Abstract

The paper presents a project of a learner's dictionary of Russian based on the most frequent Russian lexicon, and the most common words are also used for definitions. Both frequencies and examples are defined using the Russian National Corpus. The corpora methods make possible to include into the dictionary some meanings, constructions absent from the existing dictionaries of Russian let alone learner's dictionaries. Some elements of thesaurus are used in the dictionary. The structure of metalanguage, the study of collocations, and the use of corpora are discussed and exemplified in the paper.
Keywords: learner's dictionary; frequency lists; Russian National Corpus; metalanguage

1 Introduction

The Basic Russian Dictionary (henceforth: BRD), currently compiled by M. Shvedova, is a description of the most frequent Russian lexical units within their respective semantic and grammatical relationships. The dictionary is intended for foreign students of Russian. Words are given in the alphabetical order accompanied by definitions. In the definitions only the simplest grammatical constructions and frequent words are used. Some elements of thesaurus are used as illustrations. The dictionary is corpus-based, and uses the collocations and frequencies provided by the Russian National Corpus (henceforth: RNC).

The learner's dictionaries of Russian, intended for foreign students, often using pictures or elementary translations into main foreign languages, exist since 1960s. In latest years, corpora-based dictionaries are also appearing. The Active Dictionary of Russian edited by Yu. D. Apresjan (Проспект… 2010) is not a learner's dictionary but can serve as a basis for future ones: it is designed to enable the production of texts and is corpora-based. It is not, however, frequency-oriented, and features rather rare idioms.

The dictionary is intended to have an online version integrated with the Russian National Corpus and enabling a quick search that provides more examples.

2 Selection of Words

The selection of words for a learner's dictionary should be narrower as compared with a general dictionary; existing Russian learner's dictionaries include typically 3000-5000 words, selected by different criteria including frequency, stylistic neutrality, relevance for a student and semantic productivity. We chose as the basic criterion the frequency of a word in different genres of modern

texts, according to the Frequency Dictionary of Modern Russian by Olga Lyashevskaya and Serge Sharoff (Ляшевская, Шаров, 2009) (henceforth FD). The FD is based on the RNC as it stood in 2009. We use the top 5000 words of the FD as the basis for the selection for the BRD. The frequency top of the FD is built on written texts and includes mainly abstract vocabulary, whereas concrete lexemes like *grusha* 'pear' or *rascheska* 'comb' are not chosen. Such words (with no polysemy or idioms and very complicated formal definitions) are better explained to a foreign student by a bilingual dictionary or a picture, whereas abstract lexicon with difficult usage rules, combinations and idioms, that describes basic ideas, relations and concepts, is to be presented in the explanatory dictionary.
The cumbersome definitions of concrete objects are often accompanied by illustrations in learner's dictionaries. For example, the word *kist'* that can signify 'paintbrush', 'truss', 'racemation', 'tassel' and 'hand' (several meanings or even homonyms) is defined by an illustration (Figure 1).

Figure 1: Definition of *kist'*.

We may note that there are still no comprehensive illustrated linguistic dictionaries of Russian. Only the dictionaries for foreigners, including the so-called "pictorial dictionaries", are published with illustrations, and so are some technical dictionaries and encyclopaedias. Nevertheless, this idea seems to be effective and is widely used in the West.

3 Defining Vocabulary

Some existing learner's dictionaries simplify the definitions or even substitute pictures or examples for them. We propose building definitions in a simplified metalanguage not exceeding the selection of the dictionary entries itself, and even narrower (only top 4000 frequent words according to the FD). The syntax of the definitions is also simplified and does not use Russian participles or gerunds which are not familiar to the beginner students. A similar approach is used in Longman's dictionary

(Longman 2012), where a 2-thousand-word defining vocabulary (listed in the dictionary in full) with simplified syntax is used for definitions (however, it is not based on frequency but rather on the basic thesaurus). For example, *infusion* is defined in Longman's as "if you infuse tea or HERBS in hot water, you leave them in very hot water while their taste passes into the water". The word HERBS is printed in all caps as it is not included into the defining vocabulary and is to be additionally consulted in the dictionary. A similar task is fulfilled in the well-known French basic dictionary (*Dictionnaire fondamental*) by Gougenheim with a restricted defining vocabulary (Gougenheim 1958); the idea of teaching Basic French (*français fondamental*) was popular in the 1960s but rejected during the following years. This idea was never used in compiling a Russian dictionary.

A further comparison may be drawn to different projects of simplified vocabularies for beginners like the 850-words Basic English by C. K. Ogden, the 2000-words Simple English that has its own Wikipedia project (simple.wikipedia.org) and to some more artificial projects of defining vocabulary, including *Lingua mentalis* by Anna Wierzbicka.

4 Definitions and Meanings

More frequent meanings and uses are given first in the BRD. Many existing dictionaries give different meanings as equal regardless of their relative frequency and relevance for a modern reader. Some dictionaries list a historically original meaning first (e.g. the first meaning for Russian *mrachnyj* would be 'dark', 'unlit' as the word is formed from *mrak* 'darkness', whereas 'gloomy, sinister' is by far more frequent in the texts). While compiling the dictionary we often find that within Russian aspectual pairs a semantic asymmetry is attested, that is, a meaning or use is used more frequently only in perfective or only in imperfective aspects, whereas the other item of the same aspectual pair prefers another meaning. Usually in existing dictionaries only one article is provided for the whole aspectual pair. For example, the perfective verb *osushchestvitsja* is used in the sense 'to come true' (*ego mechta osushchestvilas'* 'his dream came true'), but its imperfective counterpart *osushchestvljatsja*, judging by the RNC, is used mainly in the sense 'to be organized, undertaken' (*meroprijatija osushchestvljajutsja* 'the events are organized'). This asymmetry is consistently shown in the BRD, and we seldom provide a single article for both aspects.

In the BRD we use definitions that are further specified in the comments accompanying the examples. In the article CENTR 'centre' defined as 'middle; the main part of something, the most important place' after the example *zhivet v samom centre* '(he) lives in the very centre' a comment is added in brackets: (*goroda*) 'of the city'; furthermore, after the example *nauchnyj centr* 'scientific centre' the BRD gives a comment (*vazhnoe uchrezhdenie, krupnoe predprijatie*) 'an important establishment, a major enterprise'.

5 Exploring Collocations in Corpora

Modern corpora technologies spare lexicographer's time, giving full access to vast amounts of texts, and at the same time offer new opportunities and challenges. Users may explore usage of words, word combinations and grammar constructions in different genres of texts. It is possible to extract from the RNC collocations (consisting of 2, 3, 4, or 5 tokens) and frequency lists. The RNC allows searching the collocations by their grammatical features (e.g. all the adjectives combined with the given word).

Such a search yields a lot of idioms, stable collocations and clichés that are important for language

students but are not always reflected in dictionaries. The "Idioms" section is the most innovative part of the BRD, as this level of language is unstable and constantly changing. Whereas the basic lexicon is included in virtually all the dictionaries, some idioms that are actively used in Modern Russian are often absent there, like *ne vernutsja* lit. 'not to return' = 'to lose one's life somewhere'; *ne znat', s kakogo konca vzjatzja* lit. 'not to know from which end one should take smth' = 'not to know how to address some problem', *nachat' i konchit'* lit. 'to begin and to finish' = 'the job is not even started, there is yet the whole job to be done' etc. At the same time, many dictionaries include obsolete idioms that are not frequent any more judging by the RNC. It is of particular importance for students to take into consideration the contemporary situation with idioms.

As a rather objective criterion, frequency allows to collect material for a dictionary, assess the relevance of examples and meanings, given in existing dictionaries, for modern language. For example, it can help adding most frequent collocations as examples of free combinations. The examples chosen for the BRD, as well as the definitions, are intended to include likewise only words from the top 4000 frequent lexemes according to the FD. Only fixed idioms and literary quotations can feature words not belonging to the list, but not more than 1-2 per example. A student can easily use the articles and at the same time s/he gets acquainted with a lot of different models of usage.

The adjective *dlinnyj* 'long' is illustrated by the frequent two-term collocations (bigrams): *dlinnye volosy* 'long hair' (458 examples), *dlinnyj rjad* 'a whole set, a lot of' lit. 'long range' (344), *dlinnye nogi* 'long legs' (287), *dlinnyj stol* 'oblong table' (220), *dlinnye ruki* 'long arms' (185), *dlinnoe pis'mo* 'long letter' (177) etc. The most frequent collocations with *idti* 'go' are *rech' idet* 'the question is, we are talking about…' lit. 'the speech goes' (6040), *delo idet* 'the job is successful / under way'; 'the question is…' lit. 'the business goes' (1431), *vremja idet* 'time is passing' (475), *zhizn' idet* 'life is passing' (247), *doroga idet* 'the road goes (to…)' (239), *dela idut* 'we are doing what we should do; things are OK' lit. 'businesses go' (221), *ljudi idut* 'people go' (189).

Longer collocations are useful for finding popular literary quotations or proverbs where the word in question is used, for example, a 5-word collocation query for *zhdat'* 'wait' yields K. Simonov's famous line *Zhdi menja, i ja vernus'* 'Wait for me and I will come' on the fourth frequency place. Sometimes longer collocations help find words which frequently accompany one another syntactically but are not found obligatory in direct contact, as *kazatsja* 'to seem' and *inogda* 'sometimes' (they can be separated by a pronoun or particle).

6 Exploring Semantics in Corpora

Corpus helps clarify the meaning of a word. E.g. the word *povtorjatsja* lit. 'to repeat oneself' in new texts may refer not to a previous action of the subject ('to repeat one's words') but to a piece of previously well known information, typically in a context like *ne budu povtorjatsja* lit. 'I will not repeat myself' (='I am not going to tell what you already know'). The lexicographer may also use the parallel corpora within the RNC containing, for example, English texts translated to Russian. Here the peculiarities of some Russian words can be explored as compared with their English counterparts. For example, the synonyms *bol'shoj* and *krupnyj*, both translated in the dictionaries largely in the same way (as 'big', 'large', 'great…') have the following counterparts in the English-Russian parallel corpora (list compiled by foreign students, Table 1):

bol'shoj	*krupnyj*
large (10 times)	*large* (14 times)

great (5 times)	*major* (8 times)
Big	*large-scale*
Strong	*significant*
Wide	*big*
Deep	
full-length	
Grand	
plenty of	
a lot of	
much (work)	

Table 1: The English correspondences for Russian synonyms.

It follows from this analysis that the main translation equivalent for both words is *large* (and not *big* given in a bilingual dictionary: *big* attested once while *large* 10 and 14 respectively), the second equivalent is *great* for *bol'shoj* and *major* for *krupnyj*, and the whole "translation spectre" is overall different. Thus the parallel corpus helps to clarify the meaning of a word, and these data can be used also in a monolingual explanatory dictionary.

7 Other Information

In the text of the BRD some tables are included as illustrations, showing the thesaurus-like structure of lexicon (not unlike Longman's or other English dictionaries). For example, sets of synonyms are given, with their corresponding idioms and collocations.
A fragment of thesaurus may include shortened explications (Table 2) or just give a paradigmatic list of entries (Table 3).

DEPARTMENT
Otdelenie 'department'. A part of an establishment or its separate office. [eg *pochtovoe otdelenie* 'post office']
Otdel 'department'. A part within the structure of an establishment [eg *otdel prodazh* 'sales department / office']
Filial 'branch'. An autonomous department of an establishment or organization. [eg *filial banka* 'branch of a banking company']
Podrazdelenie 'subdivision'. A part included into a larger part

Table 2: Thesaurus for DEPARTMENT.

DAY & NIGHT
den' 'day'
noch' 'night'
utro 'morning'
vecher 'evening'
polden' 'noon'
polnoch' 'midnight'
rassvet 'dawn'
zakat 'sunset'
sumerki 'twilight'

Table 3: Thesaurus for DAY & NIGHT.

Grammar information is simplified in the BRD. Syntactically ambiguous indeclinable words (adverbs/prepositions, adverbs/particles/conjunctions…) are usually given in the most frequent function, sometimes following the part-of-speech tagging within the RNC (e.g. adverbs like *vesnoj* 'in the spring', that are analyzed in the corpus as instrumental case of the word *vesna* 'spring').
We see that corpora technologies are a useful tool for compiling an explanatory learner's dictionary. The articles of the dictionary that have been already compiled are used in classroom with foreign students who sometimes suggest clarifying some points and adding examples.

8 References

Апресян, Ю. Д. (ред.) (2010). *Проспект активного словаря русского языка*. М.: Языки славянских культур.

Ляшевская, О. Н., Шаров, С. А. (2009). *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)*. М.: Азбуковник.

Gougenheim, G. (1958). *Dictionnaire fondamental de la langue française*, Paris: Didier.

Longman dictionary of contemporary English (2012). Harlow: Pearson Education Ltd. (1 ed. 1978).