The application of corpus-based approach in the Bulgarian new-word lexicography[1]

Sia Kolkovska, Diana Blagoeva & Atanaska Atanasova

Keywords: *new-word lexicography*, *corpus-based approach*, *Bulgarian lexicography*.

Abstract

The paper focuses on the specific directions of application of the corpus-based approach in the new-word lexicography. The research deals with the main application of corpus-based techniques in the elaboration of the latest academic neological Bulgarian dictionary. The usefulness of applying corpus-based techniques at the various stages of compiling this neological dictionary is described in more details. The usages of these techniques make compiling of the Dictionary easier at the following stages: working out the list of new units, included as head words in the Dictionary; determining of the degree of establishment of the new units in Bulgarian; determining of the representative variant among some graphic, phonetic or morphological variants of a new word; determining of the most typical collocations of a given head word.

1. Introduction

Our aim in the paper is to illustrate the usefulness of corpus-based approach for the elaboration of neological dictionaries. The usefulness of corpus-based approach for a dictionary making process is a well-known fact. Here we focus our attention on the specific directions of application of the corpus approach in the new-word lexicography.

Automatic (or semi-automatic) identification and information retrieval of lexical neologisms is very important for the registration of changes in the language occurring in a certain period of time (see Janicivic, Walker 1997, Rundell, Kilgarriff 2011). Well-known is the idea of "monitor corpus", which represents the dynamics of the language and allows its objective description. Specific procedures for information retrieval of words, not found in an earlier version of a corpus, but included in a newer version, are described for example in O'Donovan, O'Neil (2008).

The corpus-based approach is applied in the compilation of the latest academic neological Bulgarian dictionary: E. Pernishka, D. Blagoeva, S. Kolkovska. Dictionary of New Words in Bulgarian (at the End of the 20th and the First Decade of the 21st c.). Sofia: Naouka i Izkoustvo. 517 p. (referred to *the Dictionary* below). It is compiled at the Department of Bulgarian Lexicology and Lexicography at the Institute for Bulgarian Language and it is the most representative and complete dictionary of neologisms in Bulgarian. The aim of *the Dictionary* is to present as fully as possible the extremely high number of lexical, phraseological and semantic innovations in Bulgarian during the last two decades. The dictionary contains about 4300 newly-coined or newly-borrowed lexical units in Bulgarian, 700 neosemantisms, more than 600 new compound terms and almost 150 new phraseological units.

The "Dictionary of New Words in Bulgarian" is elaborated on the base of the corpus data extracted from the Lexicographic Electronic Corpus, which is developed for lexicographic purposes at the Department of Bulgarian Lexicology and Lexicography at the Institute for Bulgarian Language. The corpus contains more than 300 million words and includes more than 7600 electronic documents (digitalized versions of books or periodicals – newspapers, magazines, year-books, etc.). Subsequently, the corpus becomes a part of Bulgarian National Corpus (http://www.ibl.bas.bg/BGNC_bg.htm). The Lexicographic

Electronic Corpus includes four sub-corpora which have been set apart on the grounds of chronological characteristics. As empirical basis for lexicographic description of the dynamics of the new Bulgarian lexis we use a sub-corpus consisting of texts published after 1990. That neological sub-corpus is a reliable source for the extraction of lexicographically relevant information as it includes more than 6600 electronic documents which consist of more than 240 million words. The sub-corpus is representative for the new Bulgarian lexis and is relatively well-balanced as it presents a wide range of styles and genres. It comprises texts of both the informative and fictional type and includes 5289 documents (with 132 million words) selected from magazines and newspapers and 1374 documents (with 110 million words) selected from the fiction. The informative texts belong to various fields (scientific, popular science, documentary, memoir literature, political journalism, etc.) and the fictional texts present various genres of fiction, poetry, drama, etc.

The compilation of *the Dictionary* is crucially optimized as the corpus data are extracted from the neological sub-corpus by means of the search tool (developed at the Department of Computational Linguistics at the Institute for Bulgarian Language) and the lexical profiling system Sketch Engine (see Kilgarriff, Rundell 2002).

The application of the corpus-based techniques facilitates essentially the process of compilation of *the Dictionary*, especially at the following stages of that process:

2. When working out the list of new units to include in *the Dictionary*

The corpus-based approach is very useful for the creation of neological dictionaries mainly because it enables lexicographers to objectify and justify the neological status of individual units. Instead of using only traditional techniques of manually elaboration of the list of words in *the Dictionary* mentioned above, we applied the procedure of semi-automatic extraction of neologisms from the neological sub-corpus. We automatically extracted an alphabetical frequency list of the word forms from that sub-corpus. The list was compared (by means of a special computer programme) with a reference list of the word forms belonging to the core Bulgarian lexis. As a result of the comparison, a list of units from the sub-corpus, which are not included in the reference list, was created. That was an expanded list of about 30 000 units which were considered as candidates for inclusion in *the Dictionary*. This extended list was manually processed by lexicographers. During processing, some cases were removed as irrelevant ones: proper names, words containing spelling mistakes and misprints, unrecognized forms of well-known lemmas, words containing foreign graphics, etc. The shortlist, a result of the processing (comprising true neologisms as well as various types of occasionalisms, potential words, etc.) became a basis for selection of items to be included in *the Dictionary* as entries. The restrictions of described semi-automatic method for extracting neologisms (in particular: its inapplicability to detect neosemantisms, new set phrases, terminological or other phrases) were offset by traditional methods (manual extraction).

3. When determining the degree of establishment of the new units in Bulgarian

The corpus-based approach enables us to objectify and justify the neological status of the units included in the expanded list mentioned above. Only these new units from the list, which are not occasional words, are included in *the Dictionary*. The mechanism applied by us for distinguishing the new words from the occasionalisms is based on the understanding that the new words are units of language and they are lexicalized. On the contrary, the occasionalisms

are created for the single act of communication and are bound with definite context. That difference between the new words and the occasionalisms determines the difference in their frequency. Namely, the frequency we used as an objective criterion for the automatic recognition of these two types of units in the neological corpus. We assumed that a certain new unit tends to be lexicalized if it has shown a minimum frequency of more than three occurrences in the corpus (excluding repetitions and cited reproductions of the same text in different documents). New words with occurrences below the minimum were considered as ocasionalisms and they were not included in *the Dictionary*. Except the frequency of the candidates for new words, we also took into account the distribution of the words in accordance with the understanding of P. Hanks that for the corpus lexicography the frequency is an important criterion but it is not enough as the distribution also should be taken into account (Hanks 2003: 58).

The usage of statistical methods for automatic or semi-automatic detection of neologisms is limited because the new words tend to have uncommon use. As M. Rundell and A. Kilgarriff point out: "For new-word finding we will want to include items in a candidate list even though they occur just once or twice. Statistical filtering can therefore only be used minimally." (Rundell, Kilgarriff 2011).

4. When determining the representative variant among some graphic, phonetic or morphological variants of a new word

The corpus-based approach helps the lexicographers to take a decision which one of several variants of a new word is the most representative and perspective to establish in the language system. The objective corpus-driven data about the frequency of variants of a new word is very helpful because a lot of neologisms (especially borrowings) are in the process of gradually adaptation to the language and have graphical and/or morphological variants. With respect to *the Dictionary* mentioned above, the decision in a lot of cases is taken on the base of the frequency of the different variants in the neological sub-corpus. For example, because of the prevailing frequency (397 occurrences) of the graphic variant of the borrowing from English *плейър* (Engl. *player*) in comparison with the variant *плейер* (6 occurrences) we preferred the first one as a head word in the *Dictionary*. The same is the case with the forms *мърчандайзинг* (Engl. *merchandising*) and *мърчандайзер* (Engl. *merchandiser*) which were included as head words instead of less frequent forms respectively *мърчъндайзинг and мърчъндайзер*. The borrowing noun *билборд* (Engl. *billboard*) is represented in *the Dictionary* with a plural form *билбордове*, because of the prevailing frequency (514 occurrences) in comparison with other morphological variant - *билборди* (8 occurrences) (see the Figure 1 below).

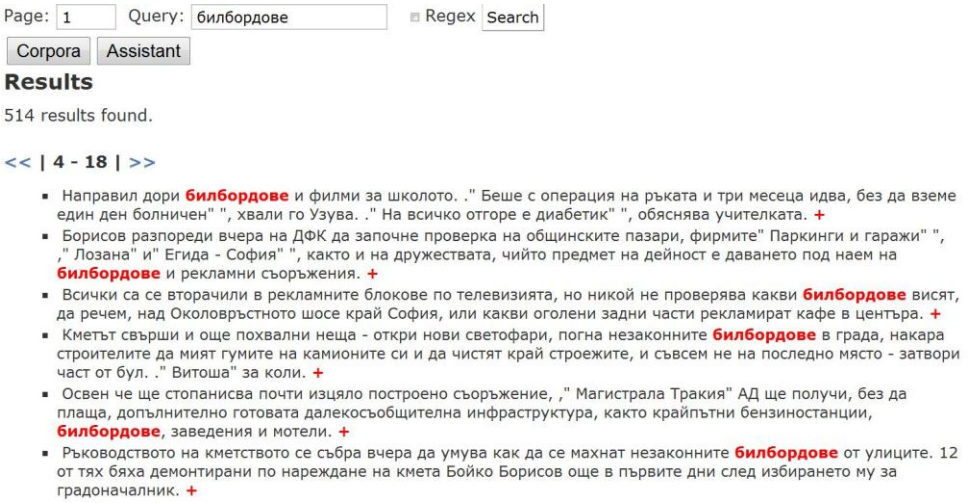

Figure 1. Occurrences of the morphological form *билбордове* in the Corpus

5. When determining the most typical collocations of a given head word

It is well-known that the corpus data helps the lexicographer in determining the most frequent collocations of a given word which should be given in the lexical entry. That information is especially valuable for the new-word lexicography because of the high dynamics of new words and the fact that in a lot of cases their usage is not established.

In the computational lexicography for measuring the significance (salience) of the collocations different statistical techniques are used such as MI, MI3, T-score, log likelihood, minimum sensitivity, log dice, MI.log_f etc. (see Kilgarriff, Tugwell 2001 and the literature quoted there). The corpus query systems usually include one or more of these functions.

For the extraction of the most frequent collocation from the neological sub-corpus we use the corpus query systems developed at the Institute for Bulgarian Language (at the Department of Computational Linguistics). For example, by means of that software tool we determined as the most frequent among the extracted collocations of the new adjective *виртуален* (Engl. *virtual*) the following collocations of that adjective: *виртуална реалност* (Engl. *virtual reality*), *виртуален секс* (Engl. *virtual sex*), *виртуална библиотека* (Engl. *virtual library*), *виртуален свят* (Engl. *virtual world*) (see the Figure 2 below). These collocations were included in the lexical entry of the adjective *виртуален* as the most representative ones.

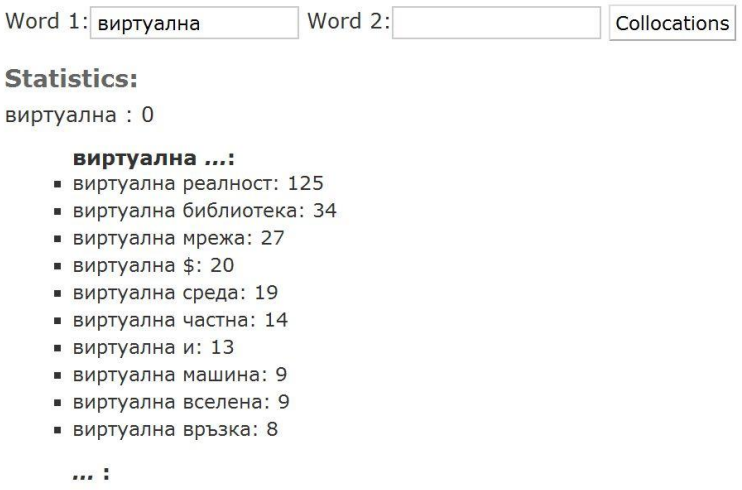

Figure 2. Collocations of the of the adjective *виртуален* 'virtual' (във форма за ж.р. ед. ч.) in the Corpus.

The application of the corpus-based approach in the new-word lexicography leads to crucial improvements in dictionary-making process and ensures objectivity and exhaustiveness in the lexicographical treatment of the lexical innovations.

Note

[1] The research is carried out with partial financial support of Bulgarian Science Fund (Contract DTK 02/52 of 2009)

References

Hanks, P. 2003. 'Lexicography.' In R. Mitkov (ed.), *The Oxford handbook of computational linguistics.* Oxford: Oxford University Press, 48–69.

Janicivic, T. and D. Walker 1997. 'NeoloSearch: Automatic Detection of Neologisms in French Internet Documents.' In *Proceedings of ACH/ALLC'97*. Queen's University, Ontario, Canada, 93–94.

Kilgarriff, A. and D. Tugwell 2001. 'WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography.' In *Proceedings of Collocations workshop*, ACL 2001. Toulouse, 32–38.

Kilgarriff A. and M. Rundell 2002. 'Lexical Profiling Software and its Lexicographic Applications: a Case Study.' In A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002.* Copenhagen: *CST*, 807–919.

O'Donovan, R. and M. O'Neill 2008. 'A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary.' In E. Bernal and J. DeCesaris (eds.), *Proceedings of the XIII Euralex International Congress: Barcelona, 15-19 July 2008.* Barcelona: L'Institut Universitari de Lingüistica Aplicada, Universitat Pompeu Fabra, 571–579.

Rundell, M. and A. Kilgarriff 2011. 'Automating the creation of dictionaries: where will it all end?' In F. Meunier, S. De Cock, G. Gilguin, M. Paquot (eds.), *A Taste for*

Corpora. In honour of Sylviane Granger. [Studies in Corpus Linguistics, 45]. Amsterdam: John Benjamins, 257–282.