A. Horák, M. Khokhlova, A. Rambousek, V. Zakharov

LEXICOGRAPHIC PLATFORM FOR DICTIONARIES AND CORPORA[1]

1. Introduction

The information society has become very quickly a computerized one. Constantly, new technologies come to new spheres of human activity. The arrival of corpus linguistics and corpora has become a relevant point in this respect. The corpora stimulated a considerable progress that has been gained in the field of lexicographic work. This has its own reason. There is no integrated software that enables to work both with traditional dictionaries and electronic sources of lexical data.

The intention is to collect resources of modern Russian dictionaries within the one framework. All these data will be converted into a well-structured format (e.g., XML format) and concentrated in a unified database. Such a database will be prepared for all kinds of linguistic research.

The idea has existed for several years and was inspired by several similar projects abroad, as the Celex database[2], and the Czech lexical database.

2. DEB platform

The basis of the new project implementation is formed by the DEB II dictionary writing systems platform developed at the Natural

[1] The work is supported by the grant of the Russian Fund for Basic Researches No. 10-07-00563 and by the grant of the Russian Humanitarian Scientific Fund No. 10-04-12135.

[2] Celex lexical database. URL: **http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html**.

Language Processing Centre, Faculty of Informatics, Masaryk University (Czech Republic).

The DEB II (Dictionary Editor and Browser, http:// deb.fi.muni.cz/) is an open-source software platform designed for the fast development of applications for viewing, creating, editing, and authoring electronic and printed dictionaries. The platform is based on the approach of the client-server architecture. Most of the functionality is provided by the server side, and the client side offers (computationally simple) graphical interfaces to users. The client applications communicate with the server using the standard web HTTP protocol.

The server part is built from small reusable parts, called servlets, which allow a modular composition of all services. Each servlet provides different functionality such as database access, dictionary search, morphological analysis or a connection to various corpora.

The overall design of the DEB II platform focuses on modularity. The data stored in a DEB II server can employ any kind of structural database and combine the results in answers to user queries without the need to use specific query languages for each data source. The main data storage is currently provided by the Oracle Berkeley DB XML[1]. However, it is possible to switch to another database backend easily, without any changes to the client parts of the applications.

The main assets of the DEB II development platform can be characterized by the following points:

- All the data are stored on the server and a considerable part of the functionality is also implemented on the server, while the client application can be very lightweight.

[1] *Chaudhri A.B., Rashid A., Zicari R., eds*. XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional, 2003.

- Very good tools for team cooperation; data modifications are immediately seen by all the users. The server also provides authentication and authorization tools.
- Server may offer different interfaces using the same data structure. These interfaces can be reused by many client applications.
- Homogeneity of the data structure and presentation. If an administrator commits a change in the data presentation, this change will automatically appear in every instance of the client software.
- Integration with external applications.

The DEB II platform versatility is apparent in more than ten projects based on the platform, ranging from dictionary viewers to complex ontology editors.

3. Electronic Dictionaries of Russian

Nowadays many dictionaries of the Russian language (including explanatory ones) exist in an electronic form. But usually these are scanned texts in either graphical or text formats. Lack of structuring makes it difficult to search in them and combine them effectively with other language resources.

Several Russian explanatory dictionaries are available on-line (through Feb-web: Fundamental Electronic Library [1]): Ushakov's Dictionary, the Dictionary of the Russian Language in 4 volumes[2], and the Dictionary of the Russian Language of the 18th century[3].

There is an option to look up only in one dictionary at the same time and browse in it but not to use it as a database. Because entries of

---

[1] **http://feb-web. ru**

[2] *Ushakov D.N., ed*. Tolkovyj slovar' russkogo jazyka v 4 tomakh. 1935–1940.

[3] *Sorokin J.S., ed*. Slovar' russkogo jazyka XVIII veka. Leningrad-St. Petersburg (since 1984).

different dictionaries have various structures that makes it hard to work with the data.

This raises the question of one integrated structure of Russian explanatory dictionaries and their conversion to this structure. Moreover, this also leads to the question of developing one tool that could be used both as a browser and an editor.

For the first stage of the project, we have chosen two dictionaries of Russian. They are the "Complex Normative Dictionary of the Modern Russian Language" ("Komplcxsnyj normativnyj slovar sovremennogo russkogo yazyka")[1] and the Dictionary of the Russian Language in 4 volumes[2].

The "Complex Normative Dictionary of the Modern Russian Language" as well as the "Normative Explanatory Dictionary of the Live Russian Language" are being compiled at the Laboratory of Computational Lexicography of the Philological Faculty of St.Petersburg State University (Russia) under the guidance of Prof. G.N. Sklyarevskaya. It is intended for users to provide them with information on correct word usage of latest and newest terms and concepts of contemporary Russian. The dictionary includes an active vocabulary whose selection was based on expert decisions about semantic, grammatical, orthoepic or other features difficult for language users. The usage of these words has to be normalized. The data is being actively revised and supplemented on the basis of corpus examples, Internet data, various terminological or explanatory dictionaries, and linguistic studies. Dictionary word list is compiled on the data of the Fund of Modern Russian (approximately 17 million tokens).

The DEBDict server was installed at the Institute for Linguistic Studies and the two dictionaries were imported. Although each of them

---

[1] Komplexsnyj normativnyj slovar sovremennogo russkogo jazyka. SPb., 2010.

[2] *Jevgen'jeva A.P., ed.* Slovar' russkogo jazyka v 4 tomakh, M., 1957–1961.

is represented with different XML structure, users are presented with the data in a unified form. Thus lexicographers have obtained access to a valuable research resource, forming the first basic part of the new lexicographic platform.

4. Quotation Corpus

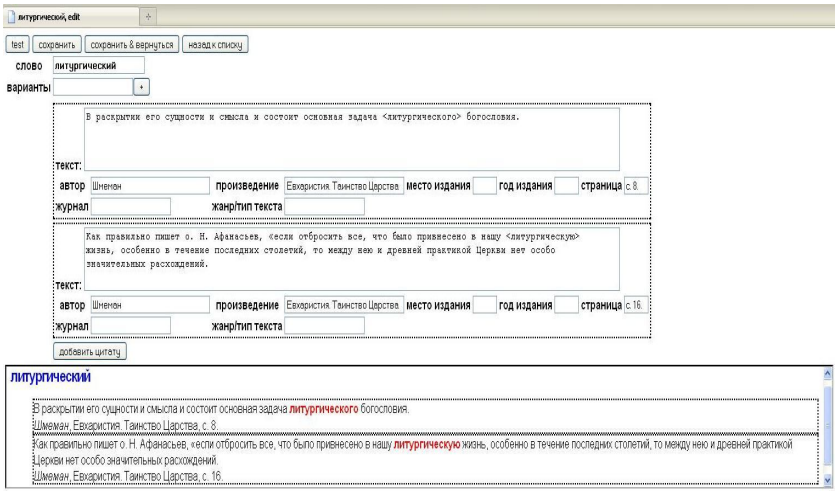
The Large Card File of the Institute for Linguistic Studies of the Russian Academy of Sciences, containing about 8 million of systematized cards with citations, allows for various types of lexicographical and philological research[1]. Its stock was used by lexicographers while compiling a great number of dictionaries and grammars of Russian. Many researchers both from Russia and from abroad use the Large Card File in their investigations on various topics.

The Large Card File was established in the 19th century and currently consists of two parts. The one comprises about 5.5 million cards (collected from 1886 till 1968), while the other contains more than 2.5 million cards (collected from 1968 till 1994).

At its present form the card file is not representative enough. This can be accounted for by both its inherent defects (as during the Soviet time a number of authors and works could not be included due to ideological reasons), and by lack of finance as a consequence for the last 15 years very small amount of new entries have been added to it. It is obvious that only cutting edge information technologies, i.e. electronic libraries, text corpora, programs for lexicographical tasks, can take care of current lexicography needs. Thus, further development and expansion of the Large Card File should be done electronically.

---

[1] *Rogozhnikova R.P.* Sokrovishchnitsa russkogo slova. Istoriia Bolshoi slovarnoi kartoteki Instituta lingvisticheskikh issledovanii RAN. SPb., 2003.



Example entry in the DEB II quotations editor tool.

The final aim is to digitize the content of all the cards in graphical form and build an electronic index of the quotations to help with searching for the headwords, authors etc.

However, digitization of the whole card file is expensive and time-consuming. In the first stage, the newly acquired quotations will be entered in the electronic quotation corpus. During the testing stage, software tools can be enhanced to meet the needs of the users and project.

The quotation corpus is implemented on the DEB II platform. The user interface is formed by a web application, thus the users do not need to install any special extensions. See the interface example in figure.

During the development of the corpus, the DEB II platform was also enhanced with new features needed for the Russian lexicographic tools. A new method of user interface localization was implemented that allows easy updates of the texts in any language and any character set. All the interface texts are stored in a XSLT file which is

transformed into several formats and included in the set of XSLT templates, JavaScript files and internal templates.

The corpus is connected with the Russian National Corpus and the DEB-Dict service with Russian electronic dictionaries, taking a step further to the desired lexicographic platform. Linguists do not need to run several applications, they can work with several resources within one tool.

Before the development of the corpus, new quotations were tentatively collected in the text files, these were converted to the XML format and imported into the corpus. Currently, the corpus contains over 2200 quotations for 2000 words.

5. Conclusion

In this paper, we have presented the results of the first phase of the development of new lexicographic platform for the Russian language. The final aim of this project is to fill in the gap in providing complex software tools based on standard technologies which offer the unified presentation of current Russian lexicographic resources.

The developed platform is based on the DEB II framework, which has currently been used in several international projects for preparing new specialized applications for presentation and editing of lexicographic resources of various kinds and purposes. We believe that the resulting system will enhance the Russian lexicographic work by processing the current rich set of resources with specialized language technologies.