IJCLA VOL. 2, NO. 1–2, JAN-DEC 2011, PP. 41–51
RECEIVED 01/11/10 ACCEPTED 26/11/10 FINAL 08/02/11

Integrated Lexicographic Platform for the Russian Language

ALEŠ HORÁK,[1] MARIA KHOKHLOVA,[2] ADAM RAMBOUSEK,[1] AND VICTOR ZAKHAROV[2]

[1] *Masaryk University, Czech Republic*
[2] *Saint Petersburg State University, Russian Federation*

ABSTRACT

This paper deals with the first phase of building an integrated lexicographic platform for the Russian language which can be used for various linguistic purposes. The aim of this paper is to present the main ideas of the lexicographic platform and linked projects (digitization of explanatory dictionaries of Russian, design and implementation of a database of citations) and to describe the current state of the data sources and lexicographic tools for Russian. This project is realized in cooperation between the Philological Faculty, St. Petersburg State University, Institute for Linguistic Studies, Russian Academy of Sciences, and the Faculty of Informatics, Masaryk University, Brno, Czech Republic. The ultimate aim is to provide new lexicographic software tools for developing explanatory dictionaries of the Russian language.

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 have become a relevant point in this respect. The corpora stimulated a considerable progress that has been gained in the field of automatization of lexicographic work. This has its own reason. There is no integrated software that enables to work both with traditional dictionaries and new electronic sources of lexical data.

The first explanatory dictionaries of Russian date as back as to the beginning of the 19th century. Among dictionaries of contemporary Russian we can name Ushakov's Dictionary (the Explanatory dictionary of the Russian Language, 1935-1940, the 2. revised edition 1947-1948) [1], Ozhegov's Dictionary (the Dictionary of the Russian Language, the first edition was published in 1949) [2], the Dictionary of the Contemporary Russian Language in 17 volumes (also known as BAS – "Bol'shoj akademicheskij slovar' russkogo jazyka", 1950-1965) [3], the Dictionary of the Russian Language in 4 volumes (also known as MAS – "Malyj slovar' russkogo jazyka", 1957-1961, the 2. revised edition 1981-1984) [4], the Complex Normative Dictionary of the Modern Russian Language ("Komplexsnyj normativnyj slovar' sovremennogo russkogo jazyka") [5], and the Big Academic Dictionary of the Russian Language in 25 volumes (also known as the new BAS – "Bol'shoj akademicheskij slovar' russkogo jazyka", since 2004) [6].

The intention is to collect resources of these dictionaries within 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 been existing for several years and was inspired by several similar projects abroad, as the Celex database [7], and the Czech lexical database [8, 9].

2 DEB PLATFORM

The basis of the new project implementation is formed by the DEB II dictionary writing systems platform developed at the Natural Language Processing Centre, Faculty of Informatics, Masaryk University.

The DEB II system (Dictionary Editor and Browser, `http://deb.fi.muni.cz/`) is an open-source software platform designed for fast development of applications for viewing, creating, editing and authoring of electronic and printed dictionaries. The platform is based on the approach of the client-server architecture (see the DEB II platform schema in Figure 1). 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 pro-

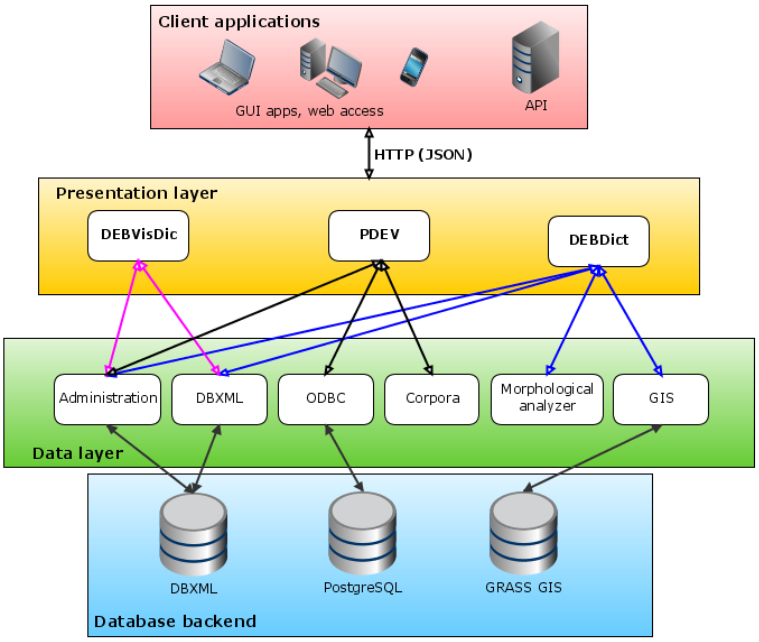

Fig. 1. The DEB II platform schema

vides 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 [10]. 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.

– 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, for example geographic information system or corpus query tools.

The DEB II platform versatility is apparent in more than ten projects based on the platform, ranging from dictionary viewers to complex ontology editors. In the following sections, we provide overview information on the main DEB II applications.

2.1 *DEBDict*

General dictionary browser, used by more than 700 users to access six electronic dictionaries of Czech and other lexical resources. Thanks to the features of the DEB II platform, DEBDict can check user's access rights and thus provide access to selected dictionaries intended for a specific group of people. For example, if the dictionary copyright does not allow public distribution, the access to the dictionary data may be limited to members of a research team.

2.2 *DEBVisDic*

The specific task of preparation of lexical semantic networks with the structure of the Princeton WordNet [11] requires special tools. During the Balkanet project [12], a wordnet browser and editor VisDic was developed by FI MU and it was used for building several national wordnets. Since 2005, it was replaced by DEBVisDic, a new system based on the DEB II development platform.

The DEBVisDic client application is split to the core module and individual modules for each wordnet. This way, it is possible to define different data structure, workflow, or include data from external sources separately for each (national) wordnet. For example, verbs in the Czech wordnet are connected to the verb valency lexicon VerbaLex [13]. The DEBVisDic server part provides an Application Programming Interface (API) usable by external applications or web services.

DEBVisDic was used as a basis for several multilingual projects – the Global Wordnet Grid [14], aiming to gather freely available wordnets of many languages, Cornetto [15], Dutch lexical semantic database, and KYOTO [16], European project building a multilingual knowledge extraction system.

2.3 *PRALED*

The Prague Lexical Database application (called PRALED) is developed in close cooperation with the linguists of the Institute of Czech Language (ICL), Czech Academy of Sciences. The application is used to build the new complex Czech Lexical Database, combining digitized dictionaries with graphical presentations of original (often hand-written) excerpt cards, several text and spoken corpora, morphological analyzer and other resources.

The PRALED users are divided into two groups: the ICL researchers are able to view and create entries, whereas others (usually reviewers) can only view finished entries. Currently, 25 linguists are using the application, each of them is creating over 200 entries per day. To add word usage evidence from the corpora, PRALED is connected with the Czech National Corpus [17].

2.4 *Art Glossaries*

In a joint project with the Faculty of Fine Arts, Brno University of Technology DEB II platform was employed as a base for the multi-lingual glossary of fine arts terms. Textual information was enhanced with the multimedia files – pronunciation recording, graphic samples, or explanatory animations. The glossary contains approximately 2000 entries and is utilized by the students as a helpful educational resource (currently on-line and textbook publishing is being considered).

Following the success of the fine arts glossary, the tool is now being enhanced for a new project in coopeartion with the Theatre Faculty, Janáček Academy of Music and Performing Arts, Brno.

2.5 *Family Names in UK*

One of the recent projects is the application to compile a database of English surnames developed for the University of the West of England.

The database will contain the meanings and origins of up to 150 000 UK surnames and will be made publicly available on-line.

The application is linked to several related resources to provide as many information as possible – surnames' frequency and location, dictionaries, genealogical sources.

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 [3]): Ushakov's Dictionary, the Dictionary of the Russian Language in 4 volumes, and the Dictionary of the Russian Language of the 18th century [18].

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 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 browser and 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" ("Komplexsnyj normativnyj slovar' sovremennogo russkogo yazyka") [5] and the above mentioned Dictionary of the Russian Language in 4 volumes [4].

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 modern Russia. The

[3] `http://feb-web.ru`

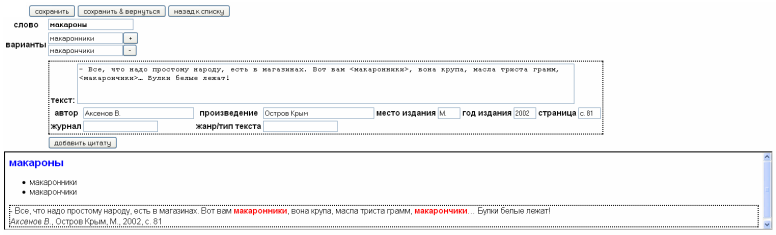

Fig. 2. Example entry (*makarony/pasta*) in the DEB II quotations editor tool

dictionary includes active vocabulary whose selection was based on expert decisions about semantic, grammatical, orthoepic or other features difficult for the 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 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 Database

The Large Card File (LCF) 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 [19]. 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. One comprises about 5.5 million cards (collected from 1886 to 1968), while the other contains more than 2.5 million cards (collected from 1968 to 1994).

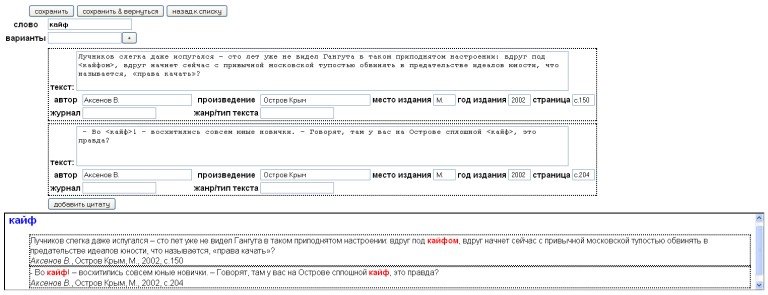

Fig. 3. Example entry (*kajf/pleasure*) in the DEB II quotations editor tool, with multiple quotations

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 LCF should be done electronically.

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 directly in the electronic database. During the testing stage, software tools can be enhanced to meet the needs of the users and project. All cards will be digitized, if the evaluation of testing stage is successful and additional funding allows it.

The quotation database 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 Figures 2 and 3.

During the development of the database, 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 database is connected with the Russian National Corpus and the DEBDict 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 database, new quotations were tentatively collected in text files, these were converted to the XML format and imported into the database. Currently, the database 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.

Acknowledgements

The work is supported by the grant of the Russian Foundation for Basic Research No. 10-07-00563A and by the grant of the Russian Foundation for Humanities No. 10-04-12135B.

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the projects P401/10/0792.

References

1. Ushakov, D.N., ed.: Tolkovyj slovar' russkogo jazyka v 4 tomakh. (1935-1940)

2. Ozhegov, S.I.: Slovar' russkogo jazyka, Moscow (1949) Ed. by S. P. Obnorskij.
3. Babkin, A.M., Barkhudarov, S.G., Philin, P.P., eds.: Slovar' sovremennogo russkogo literaturnogo jazyka v 17 tomakh, Moscow, Leningrad (1950-1965) Widely used abbreviation BAS.
4. Jevgen'jeva, A.P., ed.: Slovar' russkogo jazyka v 4 tomakh, Moscow (1957-1961) Widely used abbreviation MAS.
5. : Komplexsnyj normativnyj slovar' sovremennogo russkogo jazyka (2010)
6. Gerd, A.S., ed.: Bol'shoj Akademicheskyj slovar' russkogo jazyka Rossijskoj akademii nauk, St. Petersburg (since 2004) volume 1-14.
7. Celex: Celex lexical database (2004) `http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html`.
8. Klímová, J., Oliva, K., Pala, K.: Czech lexical database – first stage. In: Short Proceedings of Complex Conference 2005, Budapest, Hungary (April 2005)
9. Rangelova, A., Králík, J.: Wider Framework of the Research Plan Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century. In: Proceedings of the Computer Treatment of Slavic and East European Languages 2007, Bratislava, Slovakia (2007) 209–217
10. Chaudhri, A.B., Rashid, A., Zicari, R., eds.: XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional (2003)
11. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
12. Horák, A., Smrž, P.: VisDic – wordnet browsing and editing tool. In: Proceedings of the Second International WordNet Conference – GWC 2004, Brno, Czech Republic (2003) 136–141 `http://nlp.fi.muni.cz/projekty/visdic/`.
13. Hlaváčková, D., Horák, A.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Proceedings of the Slovko Conference, Bratislava, Slovakia (2005)
14. Horák, A., Pala, K., Rambousek, A.: The Global WordNet Grid Software Design. In: Proceedings of the Fourth Global WordNet Conference, Szegéd, Hungary, University of Szegéd (2008)
15. Horák, A., Vossen, P., Rambousek, A.: A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In: Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing, Haifa, Israel, Springer-Verlag (2008) 1–15
16. Vossen, P.: KYOTO Project (ICT-211423), Knowledge Yielding Ontologies for Transition-based Organization (2008) `http://www.kyoto-project.eu/`.
17. ICNC: Czech National Corpus - SYN2000 (2000) Accessible at WWW: `http://www.korpus.cz`.
18. Sorokin, J.S., ed.: Slovar' russkogo jazyka XVIII veka, Leningrad-St. Petersburg (since 1984)

19. Rogozhnikova, R.P.: Sokrovishchnitsa russkogo slova. Istoriia bolshoi slovarnoi kartoteki Instituta lingvisticheskikh issledovanii RAN, Saint-Petersburg, Russian Federation (2003)

ALEŠ HORÁK
FACULTY OF INFORMATICS,
MASARYK UNIVERSITY,
BOTANICKÁ 68A, 602 00 BRNO, CZECH REPUBLIC
E-MAIL: <HALES@FI.MUNI.CZ>

MARIA KHOKHLOVA
PHILOLOGICAL FACULTY,
SAINT PETERSBURG STATE UNIVERSITY,
UNIVERSITY EMB. 11, 199034 ST. PETERSBURG
RUSSIAN FEDERATION
E-MAIL: <KHOKHLOVA.MARIE@GMAIL.COM>

ADAM RAMBOUSEK
FACULTY OF INFORMATICS,
MASARYK UNIVERSITY,
BOTANICKÁ 68A, 602 00 BRNO, CZECH REPUBLIC
E-MAIL: <XRAMBOUS@FI.MUNI.CZ>

VICTOR ZAKHAROV
PHILOLOGICAL FACULTY,
SAINT PETERSBURG STATE UNIVERSITY,
UNIVERSITY EMB. 11, 199034 ST. PETERSBURG
RUSSIAN FEDERATION
E-MAIL: <VZ1311@YANDEX.RU>