Tools for Managing Multilingual Lexical Resources

Aleš Horáček, Karel Pala, and Adam Rambousek
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{hales,pala,xrambous}@fi.muni.cz
http://deb.fi.muni.cz/

Abstract

In the paper we describe the results in the development of the tools for handling diverse multilingual lexical resources such as monolingual and multilingual dictionaries, terminological dictionaries, complex lexicographic databases or WordNet semantic networks. All the presented tools are based on the Dictionary Editor and Browser (DEB) platform which uses standard XML formats. In this direction we strive to standardization of the lexical resources and also their interoperability. All the presented tools are freely available.

We summarize the basic features of the DEB platform as a whole and then concentrate on four applications: DEBDict (a general dictionary browser), DEBTerm (multilingual terminological dictionary editor), PRALED (Czech Lexical Database system) and Visual Browser (graphical semantic network browser).

Keywords: DEB platform, dictionary editor and browser, dictionary writing systems

1 Introduction

The Dictionary Editor and Browser II (DEB II) platform (Horáček et al., 2006b; Horáček and Pala, 2006) offers a development framework for any dictionary editing and browsing system application working with data structures that can be encoded in the XML format structures. In this way the DEB platform offers a ground for the standardized representation of lexical resources.

In the paper we present the dictionary tool DEBDict that is designed and used for handling Czech lexical resources, namely main Czech dictionaries: three explanatory ones, one synonymical and one dictionary of Czech idioms and phraseology. Moreover, DEBDict offers the access to Web resources, two encyclopedias, integrated morphological analyzer of Czech and information from the Czech corpus SYN2000.

Further, we present here also the editor and browser DEBTerm which is a tool for writing and browsing terminological lexical databases. The tool is presently tested with the fine art terminology (in cooperation with the Faculty of Fine Arts,
Brno University of Technology, where it is supported by the Czech Ministry of Education project FRVŠ 2530/2007 F2/d).

We also have to mention the PRALED lexicographical station which has been developed on the DEBII platform and is being used as a main tool for building the Czech Lexical Database in close cooperation with the Institute of Czech Language, Czech Academy of Sciences, Prague (Rangelova and Králík, 2007).

The tool DEBVisDic developed for editing semantic networks (WordNets) and ontologies has also been built on the DEBII platform but we will not deal with it here, it was described and presented elsewhere (Horák et al., 2006a, 2008).

At the end, we also have paid attention to the relations between WordNets and Semantic Web. This interest gives us a strong motivation for studying the properties of the XML data formats and tools for working with them. Moreover, we have developed the Visual Browser (Nevěřilová, 2005) – the tool which visualises semantic networks and can transform WordNet XML formats into RDF and OWL format.

2 Main Features of the DEBII Platform

The most important property of the DEBII system is the client-server nature of all its applications. The ability of distributed authorization allows teams to work fluently on one common data source. The development of applications within the DEB platform can be divided into the server part (the server side functionality) and the client part (graphical interfaces with only basic functionality). The server part is built from small parts, called servlets, which allow a modular composition of all services. The client applications communicate with the servlets using the standard web protocol HTTP.

For the server data storage the current database backend is provided by the Oracle Berkeley DB XML (Chaudhri et al., 2003), which is an open source na-
tive XML database providing XPath and XQuery access into a set of document containers.

The user interface, that forms the most important part of the client application, usually consists of a set of flexible forms that dynamically cooperate with the server parts. According to this requirement, DEB ii has adopted the concepts of the Mozilla Development Platform (Feldt, 2007). Firefox Web browser is one of the many applications created using this platform. The Mozilla Cross Platform Engine provides a clear separation between application logic and definition, presentation and language-specific texts.

The clients communicate with servlets using standard HTTP protocol with the data structures encoded in the JSON format). Any external application able to send and receive such data is able to use server API.

2.1 Assets of the DEB platform

According to our recent experience with the DEB ii platform the following positive features can be mentioned:

- It is the client-server architecture that brings along a lot of benefits. All the data are stored on the server and considerable part of functionality is also implemented on the server, while the client application can be very lightweight.
- This approach provides very good tools for team cooperation; data modifications are immediately seen by all the users. The server also provides authentication and authorization tools.
- Server offers different interfaces using the same data structure. These interfaces can be reused by many client applications. It should be stressed that several client applications can use the same interface to query XML dictionaries with different underlying structures, as it is done, for example, in DEBDict. The client part can be implemented in any way – it may be coded in any programming language or may even look only as a simple web page.
- One of the main benefits of developing DEBDict and DEBTerm on the DEB ii platform is the 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. And of course, any data flaws discovered can be instantly corrected, there is no need to change the client software or provide new data files to each client.
- Of course, a certain drawback of the client-server architecture is that an operating server is necessary for a fully functional application. However, in special situations, the server can be installed within a local environment, or for the possibility of simple off-line dictionary editing, the client may work in a degraded manner without the instant connection to the server.
- DEB ii based tools exploit XML data format thus making the lexical databases more standard and interoperable. Not only that, thanks to the XML data format used and to its dictionary specific configurability the DEB ii platform is already serving for developing various types of dictionaries, i.e. monolingual, translational, thesauri, multilingually linked wordnet-like databases (semantic networks) and ontologies.
• integration with the corpus manager Bonito2 and Word Sketch Engine (Kilgarriff et al., 2004) which allows a user (lexicographer) to obtain the sorted individual contexts including frequencies and statistical distribution parameters (salience).

3 DEBDict – general dictionary browser

The most used client on the DEBII platform we are going to present is a general dictionary and lexical databases browser named DEBDict\(^1\). It is designed for all the users who need to work with the various versions of Machine Readable Dictionaries to obtain the necessary lexical information and allows to work with literally any number of the electronic dictionaries without modifying them in a special way.

Presently, the queries to several XML dictionaries (of different underlying structures with different DTDs) are possible, particularly to all relevant Czech dictionaries, i.e.

- Slovník spisovného jazyka českého (SSJČ, Dictionary of Literary Czech),
- Slovník spisovné češtiny (SSČ, Dictionary of Written Czech),
- Slovník cizích slov (SCS, Dictionary of Foreign Words),
- Slovník českých synonym (SCS, Dictionary of Czech Synonyms),
- Slovník českých frazeologismů a idiomů (SCFI, Dictionary of Czech Phraseologies and Idioms),
- Příruční slovník spisovného jazyka českého (PSJČ, Reference Dictionary of Literary Czech).

Further, the DEBDict can work with encyclopedic resources like Encyclopedia Diderot, Seznam Encyclopedia, Wikipedia, CIA World Factbook, it can access the searching tool Google and it is linked to the Czech morphological analyzer Ajka.

Moreover, the DEBDict displays also multilingual features, i.e. besides the Czech dictionaries DEBDict is already used with three English dictionaries and one Slovenian dictionary.

DEBDict checks user access rights, some of the dictionaries may be available to all users, some of them only to a restricted group depending on what kind of the copyright applies.

The DEBDict client displays the following basic functions (see Fig. 1):

- multilingual user interface (English, Czech, other language can be easily added, e.g. Slovak, Croatian or Polish). In this respect we are ready to offer DEBDict tool to Polish and other colleagues and, in fact, to anybody who may be interested in testing and using it,
- queries to several XML dictionaries (having different underlying structure) with the result displayed via XSLT transformations,
- access to two encyclopedias: Diderot and Seznam,

\(^1\)Available at http://deb.fi.muni.cz/debdict/
connection to a morphological analyzer,
• connection to an external website (Google, Answers.com, Wikipedia),
• connection to a geographical information system (GIS) which can display geographical data directly on their positions within a cartographic map.

DEBDict typically offers to a user a complete entry structure (from various dictionaries) with all relevant fields which typically include:

• orthoepy (spelling),
• morphological properties (POS, the respective grammatical categories),
• description of the meaning (entry definition),
• word formation nest (subnet),
• syntactic properties (most often valencies),
• stylistic, domain and regional features,
• semantic relations to other entries (cross-references),
• etymological information,
• integration with Czech morphological analyzer,
• remarks and additional comments.

A relevant feature of the DEBDict is its integration with the corpus manager Manatee/Bonito2 which allows user to see the frequency of the headword in SYN2000, CNK) and obtain its Word Sketch table (Kilgarriff et al., 2004) – this allows a user (lexicographer) to obtain the individual contexts sorted according to frequencies in corpus and statistical distribution parameter such as salience.
4 DEBTerm

The DEBTerm client is a main tool used for preparation of a new terminological dictionary of the Czech art terms. This work is a joint project of the Faculty of Fine Arts, Brno University of Technology and Masaryk University. The aim of the first stage is to build a multilingual terminological database consisting of more than 1300 glossary entries. This glossary is available on-line to all the Fine Art Faculty students and it will be also published as a textbook.

All the entries include a given term in four languages (Czech, English, German and French), domain categorization, definition, usage examples and morphological information. Where possible the information about hyperonyms is included thus making it suitable for preparation domain ontologies. Many entries are enriched with the recorded pronunciation and/or photography, animations and video clips.

Right now, the glossary is undergoing review process. Project continuation is very likely with the aim to enlarge the glossary to several thousands entries and enhance more entries with multimedia resources.

What we see as a very positive result is the fact that DEBTerm can be used as a tool for building literally any other terminology database. Thanks to client/server architecture the necessary changes for another domain terminology can be done rather quickly (in the term of weeks).
The PRALED system is specifically designed for preparation of a complex structured lexical database of the Czech language named LEXIKON 21 (Rangelova and Králík, 2007). PRALED is being developed in close cooperation between the Institute of Czech Language (ICL), Czech Academy of Sciences and the NLP Centre, Faculty of Informatics, Masaryk University, where ICL specifies the user interface requirements and NLP Centre takes care of the actual software development.

The design of the PRALED lexicographic station solves the problem of editing very complex dictionary entries. Lexicographers can see the visual representation of the dictionary entry as a whole and at the same time they are able to edit any selected details in a separate subform:

- headword variants
- definitions of senses
- pronunciation
- etymology
- domain
- expressiveness (emotional attitudes)
- geographic (regional) information
- archaic or present-day expressions
- statistical information from corpora
- grammatical categories
- abbreviations, multi-word expressions and idioms
- corpus exemplification and Word Sketches
The system is currently exploited for Czech, but with appropriate modifications it can serve for building lexical database of any other language.

6 Visual Browser

The last tool we are going to present is the Visual Browser\textsuperscript{2} that is able to visualise data written in RDF and OWL notation. As with previous tools XML standards by W3C are used as the preferred format for data exchange and storage, making it possible to import such data into visualisation tools. XML formats allow for a clear separation of the form and the representation. This feature allows applications to represent the same data in several ways including rich visual forms.

Visual Browser is a Java application that can visualise various data based on RDF/RDFS notation (formats). At the moment we are using Visual Browser for making some perspectives to view the EuroWordNet data as it can be seen in Fig. 5. In this data model, WordNet literals form RDF resources. Due to this feature, polysemous words are displayed only once, but connected to all their meanings.

It can be seen that Visual Browser is a suitable tool for visualizing the entries from Machine Readable Dictionaries, lexical databases, semantic networks (Wordnets) and ontologies since all of them use various kinds of graphs (trees, nets, oriented graphs).

\textsuperscript{2}Available at \url{http://nlp.fi.muni.cz/projects/visualbrowser/}.  

\begin{figure}
\centering
\includegraphics[width=\textwidth]{visual-browser.png}
\caption{Visual Browser showing WordNet relations}
\end{figure}
7 Multilinguality and Interoperability

We have already indicated that the described DEB\textsuperscript{ii} tools are built as language independent. In practice, this means that the lexical data of any language can processed by them if they exist in the XML formats. Thus we can speak about a family of XML formats that are suitable for the representation of the lexical data which can be used for many languages. In this way the tools fulfill the requirement of multilinguality.

If interoperability means that we want to be able to exchange quickly and easily lexical data and lexical resources in XML formats the DEB\textsuperscript{ii} tools certainly contribute to this goal in a large extent. Some of the dictionaries we work with under the DEB\textsuperscript{ii} platform partially conform with the TEI dictionary Guidelines (Sperberg-McQueen et al., 1994) and others can be easily converted into TEI formats. They yield the possibility to work with the families of XML formats for which the conversion scripts can be written rather easily thus making the manipulation with lexical data faster, more economical, and last but not least, also more comfortable.

8 Conclusions

We have presented a set of the tools based on the DEB\textsuperscript{ii} platform, in particular DEBDict, DEBTerm and PRALED, which are able to handle any dictionary of any language. The only condition has to be fulfilled: the dictionaries have to be in XML format or in a similar one (HTML, SGML, eventually even LaTeX format). The integration of a morphological analyzer, however, depends on the fact whether it exists for a given language or not. The development of the DEB\textsuperscript{ii} based tools is continuing and we also will pay attention to even closer integrating the described tools to our corpus tools (corpus manager Manatee/Bonito2 including Work Sketch Engine).

Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the project T100300419, by the Ministry of Education of CR in the National Research Programme II project 2C06009 and within the Center of Basic Research LC536.

References


Bohumil Havránek, editor (1935-1957), Příruční slovník jazyka českého (Reference Dictionary of Czech Language, PSJC), Státní pedagogické nakladatelství/SPN, Praha, electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno.


Jan Petr et al. (2002), Slovník spisovného jazyka českého (Dictionary of Written Czech, SSJC), Academia, Praha, 1st edition, electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno.


C.M. Sperberg-McQueen, L. Burnard, Association for Literary, Linguistic Computing, University of Virginia, and Library (1994), TEI Guidelines for Electronic Text Encoding and Interchange P3, University of Virginia Library.

F. Čermák et al. (1983), Slovník české frazeologie a idiomaticky I-IV (Dictionary of Czech Phraseology and Idioms, SCFI), Academia, Praha.