Extended VerbaLex Editor Interface Based on the DEB Platform

Dana Hlaváčková, Adam Rambousek

NLP Center
Faculty of Informatics, Masaryk University,
Brno, Czech Republic
{hlavack,xrambous}@fi.muni.cz

Abstract. The paper presents new editing interface for lexicon of verb valencies (VerbaLex). Previous method of editing text files in custom setup of VIM text editor will be replaced by web application based on DEB platform.
New interface will contain more strict control of the valencies format and structure. It should also be easier to learn for new editors. The data will be saved in XML format, thus allowing conversion to many output formats.

Key words: VerbaLex, verb valencies, DEB, dictionary writing systems

1 Introduction

The beginnings of building the verb valency frame dictionary at the Faculty of Informatics at Masaryk University (FI MU) dates back to 1997 [1]. Since then, the dictionary, denoted as Brief, has undergone a long development and has been used in various tools from semantic classification to syntactic analysis of Czech sentence [2].

VerbaLex is a large lexical database of Czech verb valency frames and has been under development at The Centre of Natural Language Processing at the Faculty of Informatics Masaryk University (FI MU) since 2005. The organization of lexical data in VerbaLex is derived from the WordNet structure. It has a form of synsets arranged in the hierarchy of word meanings (hyper-hyponymic relations). For this reason, the headwords in VerbaLex are formed by lemmata in synonymic relations followed by their sense numbers (standard Princeton WordNet notation).

Verbalex is based on three independent resources – electronic dictionaries of verb valency frames:

- BRIEF — a dictionary of 50,000 valency frames for 15,000 Czech verbs, which originated at FI MU in 1997 [11]
- VALLEX – a valency lexicon of Czech verbs based on the formalism of the Functional Generative Description (FGD) developed during the Prague Dependency Treebank (PDT) project [13]

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011, pp. 59–65, 2011.

SYNSET: BAVIT:1, ROZPTÝLIT:2, ROZPTYLOVAT:2
DEFINITION: poskytovat někomu zábavu/make (somebody) laugh

- *passive: yes*
- *meaning: I*
- *class: amuse-31.1-1*
- *impf: bavit:1 pf: rozptýlit:2 impf: rozptylovat:2*

frame: AG <person:1>$^{\text{obl}}_{\text{whoNom}}$ VERB PAT <person:1>$^{\text{obl}}_{\text{whatAccus}}$
ACT <act:2>$^{\text{opt}}_{\text{by doing whatInstr}}$

- *example: impf: bavil děti hrou (he amused the children by playing the game)*
- *attr: use: prim, reflexivity: obj_ak*

Fig. 1. An Example of a Verbalex Valency Frame

- Czech WordNet valency frames dictionary created during the Balkanet project [12] and containing 1,359 valency frames (incl. semantic roles) associated with 824 sets of synonyms (synsets)

The structure of valency information can be seen in Figure 1 and is briefly described below.

The current version of VerbaLex contains 6,360 synsets, 21,193 verb senses, 10,482 verb lemmata and 19,556 valency frames. The valency database is developed in TXT format and available in XML, PDF and HTML formats.

2 The DEB platform for dictionary writing systems

VerbaLex editor is built on the DEB development platform, which allows the system to use many components common to dictionary writing systems. DEB (Dictionary Editor and Browser, `http://deb.fi.muni.cz/`) is an open-source software platform for the development of applications for viewing, creating, editing and authoring of electronic and printed dictionaries. The DEB platform follows the client-server architecture. Most of the functionality is provided by the server side and client side offers graphical interfaces for 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 corpora.

The overall design of the DEB platform focusses on modularity. The data stored in a DEB server can use 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 [5], which is an open-source native XML database providing XPath and XQuery access into a set of document containers. However, it is possible to switch to another database backend easily.

We have experienced issues with the database performance, so we compared several XML databases in series of benchmarks. Database systems working with XML data (both native XML databases and XML enabled relational databases) are already widespread and used in many areas. Their performance was benchmarked by many projects using several benchmarks. However, conclusions of previous publications [6,7,8] do not provide one definitive answer as for the choice of the best XML database. Generally, the results suggest that different XML benchmarks can show different weak and strong points of each database systems. When comparing the two classes of XML databases, i.e. relational databases with XML support and native XML databases, we can see that XML enabled relational databases process data manipulation queries more efficiently, and native XML databases are faster in navigational queries which rely on the document structure.

Because of the special focus on dictionary writing systems, we ran different test suites designated to both "raw speed" of the database and to specific requirements of knowledge and ontology systems. According to the results of the tests (see [9] for the details of the tests results), none of the available native XML databases can supersede the others for all kinds of operations needed for knowledge and ontology storage and manipulation. Berkeley DB XML cannot efficiently solve the queries involving multiple nodes and full-text queries. The eXist database contains the Lucene module for text search and supports many XML standards, so it can be recommended for deployment where these features are more important than the database performance. On the other hand the MonetDB database can be, according to its specific architecture, conveniently used for when working with very large amounts of XML data. For middle-size data collections, the Sedna database can provide the same performance as MonetDB, while offering richer set of features. The potential drawbacks of Sedna are the need to use special queries for the defined data indexes and the use of commercial tool for optimized full-text queries. However, the full-text queries without this optimization are already comparably fast.

During the testing of both database engines within the DEB platform, we found out that the MonetDB programming interface for the Ruby language used in the DEB platform is not stable enough and not developed actively at the moment. Because of that, MonetDB is not ready yet to be included in the platform. Fortunately, Ruby interface for Sedna is stable and maintained and better suited for DEB platform. That is why Sedna was chosen for the DEB database backend transition. It is now used for all new project and existing projects will be transfered later.

The user interface, that forms the most important part of each dictionary application, usually consists of a set of flexible forms that dynamically cooperate with the server. Most of the DEB client applications are developed using

the Mozilla Development Platform [10]. The 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. Furthermore, it imposes nearly no limits on the computer operating system of the users when accessing the dictionary data – the DEB applications run on MS Windows, Linux or Mac OS.

Thanks to the enhanced features of new HTML standards and their support in modern web browsers, many of the dictionary writing systems can be implemented as web applications.

The main assets of the DEB 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.

3 Current DEB applications

The DEB development platform provides a basis for many different kinds of lexicographic applications. The list of real dictionary systems that was developed on the DEB platform currently contains the following applications:

- DEBDict, a general multiple-dictionary browser
- DEBVisDic, wordnet editor and browser
- TeDi, multilingual terminological dictionary of art terms
- TeA, multilingual terminological dictionary of agriculture terms
- Cornetto, editor and browser of Dutch lexical-semantic database
- Global Wordnet Grid, publicly accessible multilingual wordnet dictionary
- PRALED, complex application for building new Czech lexical database
- KYOTO, backend for wordnet and ontology storage in EU-FP7 project
- PDEV (CPA), Pattern Dictionary of English Verbs, tightly connected with corpora
- Family Names in UK, web editor for Comprehensive Dictionary of English Surnames

The first two applications are widely used with hundreds of users all over the world and with participation in various national and multilingual research

projects. In the following paragraphs, we will provide more details about DEB-Dict and DEBVisDic as well as PDEV and the Dictionary of English Surnames, which are the most interesting (besides PRALED) from the lexicographic point of view. The whole next section will then be devoted to PRALED.

4 New VerbaLex interface

New editing interface was designed to be easily understandable by new editors and also to lower the possibility of editing error (for example, typo in the value from the list). Because of that most of the valency elements allow only the selection from predefined set of values. Usually, editor writes just the verb sense definition and frame examples.

Main difference from the old data format is removal of optional elements, for example it was possible to define combination of several semantic roles for one frame. This feature caused problems in automatic processing and conversions of the lexicon data. In the new format, each frame can contain only one semantic role and frames are duplicated with role changes, if needed. New web editor can be seen in figures 2 and 3.

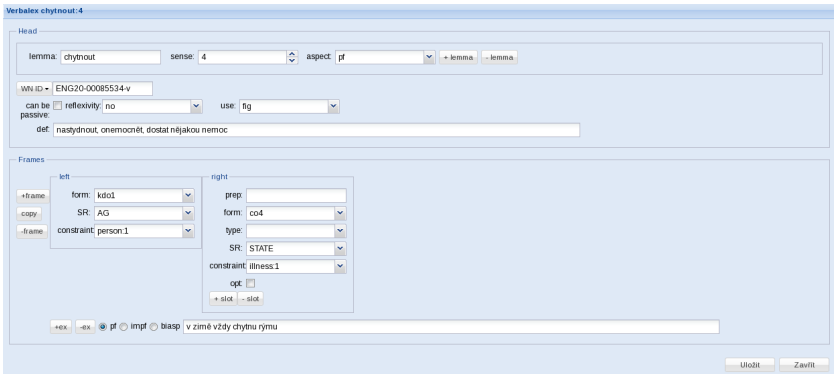

Fig. 2. Example of new VerbaLex editor interface.

5 Conclusions

New editing interface is already used to add new verbs and missing verb senses to VerbaLex lexicon. Over a hundred of verb senses was edited, the interface and XML format for the data was tweaked during the testing phase. Existing data will be converted to the new XML format and checked during the process.

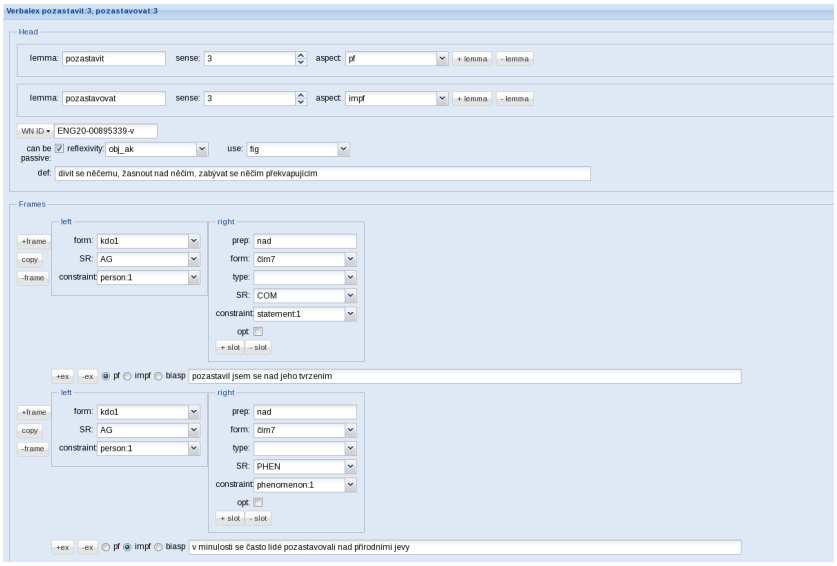

Fig. 3. Example of new VerbaLex editor interface.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and LINDAT-Clarin project LM2010013 and by the Czech Science Foundation under the project P401/10/0792.

References

1. Pala, K., Sevecek, P.: Valence Českých sloves (Valencies of Czech Verbs). In: Proceedings of Works of Philosophical Faculty at the University of Brno, Brno, Masaryk University (1997) 41–54
2. Smrz, P., Horak, A.: Determining type of TIL construction with verb valency analyser. In: Proceedings of SOFSEM'98, Berlin, Springer-Verlag (1998) 429–436
3. : Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`. (2002)
4. Stranakova-Lopatkova, M., Zabokrtsky, Z.: Valency dictionary of czech verbs: Complex tectogrammatical annotation. In M. Gonzalez Rodriguez, C.P.S.A., ed.: LREC2002, Proceedings. Volume III., ELRA (2002) 949–956
5. Chaudhri, A.B., Rashid, A., Zicari, R., eds.: XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional (2003)
6. Böhme, T., Rahm, E.: Multi-user evaluation of XML data management systems with XMach-1. Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web (2008) 148–159
7. Nambiar, U., Lacroix, Z., Bressan, S., Lee, M., Li, Y.: Efficient XML data management: an analysis. E-Commerce and Web Technologies (2002) 261–266
8. Lu, H., Yu, J., Wang, G., Zheng, S., Jiang, H., Yu, G., Zhou, A.: What makes the differences: benchmarking XML database implementations. ACM Transactions on Internet Technology (TOIT) **5**(1) (2005) 154–194

9. Bukatovič, M., Horák, A., Rambousek, A.: Which XML storage for knowledge and ontology systems? In: Knowledge-Based and Intelligent Information and Engineering Systems, Springer (2010) 432–441
10. Feldt, K.: Programming Firefox: Building Rich Internet Applications with XUL. O'Reilly (2007)
11. Pala, K., Ševeček, P.: Valence českých sloves, In: Proceedings of Works of Philosophical Faculty at the University of Brno, MU, Brno (1997) 41–54.
12. BalkaNet project website (2001–2004) http://www.ceid.upatras.gr/Balkanet.
13. Žabokrtský, Z.: Valency Lexicon of Czech Verbs. Ph.D. thesis, MFF UK, Prague (2005)