PRALED – A New Kind of Lexicographic Workstation

Aleš Horák, Adam Rambousek

NLP Center
Faculty of Informatics, Masaryk University,
Brno, Czech Republic
{hales,xrambous}@fi.muni.cz
http://deb.fi.muni.cz

Abstract. This article describes the structure, features and usage of a specialized lexicographic workstation, named PRALED, developed by the Faculty of Informatics, Masaryk University for the purpose of development of new modern lexical database of the Czech language at the Institute of Czech Language, Czech Academy of Sciences.
The PRALED system is based on the Dictionary Editor and Browser (DEB) development platform that is designed for implementations of general dictionary writing systems. The design of the PRALED client and server parts is oriented to fluent editing of very complex dictionary entries. The resulting lexicographic database contains all the morpho-syntactic information of Czech lexical entries in a machine readable form, providing an invaluable resource for both human experts as well as computer applications.
The article describes the DEB platform, as well as some current DEB applications, focusing on the PRALED lexicographic station.

1 Introduction

Preparation of complex lexicographic collections, usually in the form of printed dictionaries, was for long times dependent on large collections of excerpts from literary works that had to be tediously catalogued and organized by lexicographic experts. In the case of the Czech language, [1] and [2] are examples of large Czech dictionaries which were prepared using this technique.

The growing need to handle various lexical resources that take the form of dictionaries, semantic networks, ontologies, valency lexicons, or FrameNets is the cause why researchers seek for software systems that are able to store dictionary-like data in effective data structures. Many dictionary publishing houses operate large systems with the complex functionality of so called lexicographic stations that manipulate XML [3] and several companies offer dictionary writing programs of different complexity [4] or [5]. However, these and similar tools are not always able to efficiently manipulate resources obtained from data-driven NLP applications.

Reflecting the development of information technologies, the Institute of Czech Language (ICL) in Prague has been working on several digitizing projects. See e.g. http://bara.ujc.cas.cz/psjc/ for digitized version of [1] with graphical presentations of original (often hand-written) excerpt cards. ICL had set up a special section

for the computerization of data. Currently, Czech lexicographers have not only access to the electronic version of the digitized excerpts and previously published printed dictionaries, but also to numerous text and spoken corpora created at the Institute of the Czech National Corpus, the NLP Centre of the Faculty of Informatics, Masaryk University (FI MU) in Brno and the Institute of Formal and Applied Linguistics, Charles University in Prague.

Within the research intent *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* [6] (Academy of Sciences research intent AV0Z90610521), all the available resources (corpora, morphological analyzer, digitized dictionaries, ...) are used in preparing the supportive material for the future new lexicographic description of Czech, which will be finally published as an electronic modern monolingual dictionary of the Czech language named LEXIKON 21.

Before the start of the project, several dictionary writing systems were evaluated by ICL (mainly [4], and [5]). It was decided to develop custom application, so called lexicographic workstation that integrates several linguistic and NLP applications in a single environment, for example corpora query tool, dictionary browsing, and database editing. Custom application fits the needs and requirements of the lexicographic project much better than general dictionary writing system. Furthermore, the goal of the project is not the dictionary developement, but the developement of a complex lexical database. However, to save work and not to build the application completely from the ground, it was decided to use DEB (Dictionary Editor and Browser) platform as a base for the lexical database editor.

The lexicographic work has been divided into several phases – in the years 2005–2008 all the lexicographic resources of ICL have been digitized and, together with FI MU, a new system used as a lexicographic station, named PRALED (short for Prague Lexical Database), has been designed and implemented. In the years 2009–2012 PRALED was used for building a complex database of 100 000 lexical units of various types. PRALED application was also actively developed and updated according to requirement changes during the lexicographic research process. The preparation of LEXIKON 21 will be using the results of this second phase of the ICL's research intent.

2 The DEB platform for dictionary writing systems

The PRALED lexicographic station 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 platform is developed directly by the PRALED development team (with the authors as team leaders and main developers) at FI MU, thus many new features that have been implemented for PRALED lexicographers are now generally available for all other DEB applications. The DEB platform follows a client-server architecture – see the DEB platform schema in Figure 1. 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.

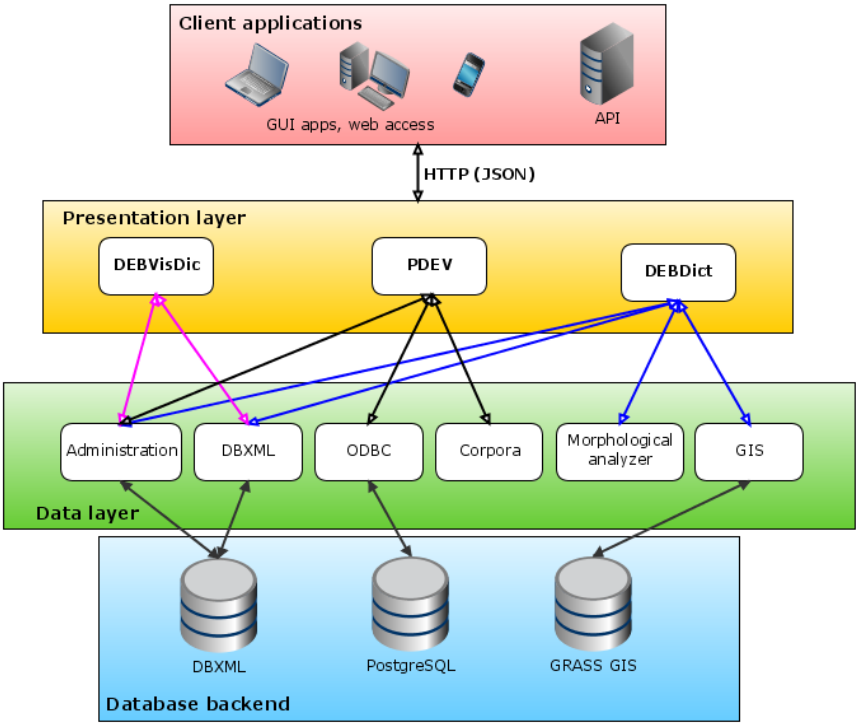

Fig. 1. The DEB platform schema

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 [7], 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.

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 [8]. 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.

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
- DEBTEDI, multilingual terminological dictionary of art 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 DEBDict 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.

3.1 DEBDict

A DEB application with many of active users is a general dictionary and lexical database browser named DEBDict, which is available at `http://deb.fi.muni.cz/debdict/`. It is designed for all users who need to work with various versions of machine readable dictionaries to obtain the necessary lexical information and it allows to work with any number of electronic dictionaries.

At the DEB server at FI MU, DEBDict offers access to all relevant dictionaries of the Czech language. Thanks to the features of the DEB 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.

3.2 DEBVisDic

The specific task of building a lexical semantic network in the form of the Princeton WordNet [10] requires special tools. During the Balkanet project [11], a wordnet browser and editor VisDic was developed by FI MU. VisDic was used for creating several national wordnets. Since 2005, it was replaced by DEBVisDic, a new system based on the DEB development platform.

The DEBVisDic client application is split to the core and the individual modules for each wordnet. This way, it is possible to define different data structure, workflow, or include data from external sources per each of the (national) wordnets. For example, verbs in the Czech wordnet are connected to the verb valency lexicon VerbaLex [12].

Besides the data for the user interface, the DEBVisDic server part provides also application programming interface (API) that is usable by external applications or web services (e.g. the OntoTagger tool from the KYOTO project).

DEBVisDic has been used as a basis for several multilingual projects: the Global Wordnet Grid [13] – aiming to gather freely available wordnets of many languages, Cornetto [14] – Dutch lexical semantic database, and KYOTO [15] – European project building a multilingual knowledge extraction system.

3.3 PDEV – Pattern Dictionary of English Verbs

The *Corpus Pattern Analysis* (CPA) [16] is a new technique for mapping meaning onto words in text. It is currently being used to build a *Pattern Dictionary of English Verbs* (PDEV), which will be a fundamental resource for use in computational linguistics, language teaching, and cognitive science.

The verbs in PDEV are analysed by the verb patterns with links to example sentences from corpora. Lexicographers, when creating new entries in PDEV, divide the concordances of the verb to several groups and create a *verb pattern* for each group, describing the subject, the objects, adverbials and other pattern elements.

The PDEV application is currently also used in projects building Czech, Italian and Spanish pattern dictionaries. All of them share the same base application with custom modifications for each language. The PDEV application is tightly connected with the

Sketch Engine. The users can easily display and edit corpus examples while editing the pattern.

3.4 Family Names in the United Kingdom

A joint project with the University of the West of England called *Family Names in United Kingdom* started in May 2010 and is set to create the largest ever database of the UK's family surnames.

The editor of the surname database is a web-based application, which extremely simplifies the installation process for users. All the surnames are divided to groups of surnames connected by variant spelling or references – the lexicographers then work with these groups. The application supports several functions to make the editing easier, for example automatic opening of new surnames added to the group, or moving the explanation between surnames, references are checked and fixed automatically. Other functionalities are designed to help the editors – e.g. quick inserting of special symbols (Greek alphabet etc.), reference documents searching, or templates for frequently used texts.

4 PRALED

The Prague Lexical Database application (called PRALED) is developed in close cooperation with linguists and users from the Institute of the Czech Language. The design of PRALED is based on the DEB platform.

Since the beginning of the project, the application and the user interface is continuously updated according to the changes in the research data and the needs of the lexicographic team. Thanks to the design of the DEB platform and the Mozilla Development Platform, it is possible to prepare prototypes of new versions in short periods of time.

The design of the PRALED client and server parts is oriented to fluent editing of very *complex dictionary entries*. The resulting lexicographic database contains all the morpho-syntactic information of Czech lexical entries in a *machine readable form*, providing an invaluable resource for both human experts as well as computer applications.

The PRALED users can be divided into two groups: the ICL researchers are able to view and create entries, whereas others (usually reviewers) can only view the finished entries. During the editing phase, 25 linguists were using the application, each of them creating or extending over 200 entries per day. During the last year of the project (2011), over 10 reviewers were evaluating completed entries for the purpose of the project final report. As of July 2012, several researchers and reviewers are evaluating the data in the preparation of the continuation project.

The client application consists of two parts – the *entry listing*, and the *complex editing form*. After successful login to the application, the entry listing window is displayed (see Figure 2). The dictionary is organized by headwords, or lexical entries – a single word, or multiple word expression, that form a dictionary entry with several meanings. With the basic filtering, a user can search for entries by headword, or by a piece of text from the definition. In the advanced search it is possible to freely combine any

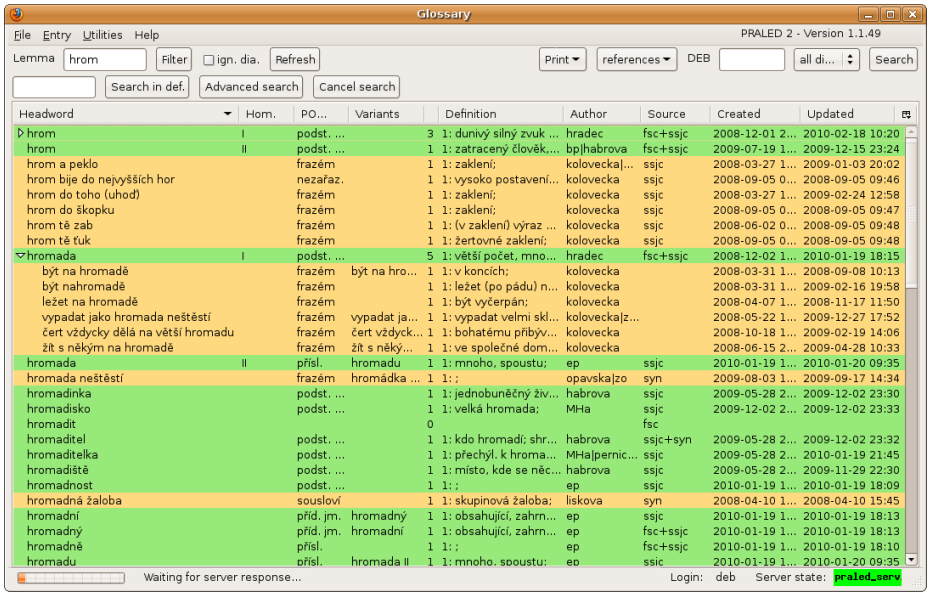

Fig. 2. PRALED: The List window

criteria from the entry data (for example, entries edited by a selected author in January 2011). The client application then translates the user selection to an XPath query that is executed on the server.

The list of selected entries is provided by the server in the RDF format. Thanks to this format, the user can view the results sorted by different fields (headword, author, last change date, etc.) and it is also possible to nest linked entries together (collocations are linked to the main headword). The resulting list can be printed in several output formats. To distinguish the entries in the list for user, the rows have different colours according to the entry type (single word, collocation, abbreviation).

A separate window with the preview of all the information of a lexical entry is opened for each edited headword. Users are able to show or hide the data as they need for current task. The entry window corresponds to the dictionary XML structure and is divided to general entry information and specific information for each sense. Linguists in ICL usually concentrate on one feature of the word and several people together work on each entry. To make this task easier, users can select which information they want to edit (for example, morphology) and a separate tool is opened to edit just the part of entry structure. Figure 3 shows the entry editing window for the word *jazyk* (*tongue*) with the general information at the top – about the word itself (frequency, PoS, grammar etc.), the editors, and possible derivated words or collocations. Next section displays the data for each sense of the word (user can also select to edit only one sense) – definition, usage evidence from corpora, synonyms etc. The application handles team cooperation and ensures that several users do not overwrite changes of one another.

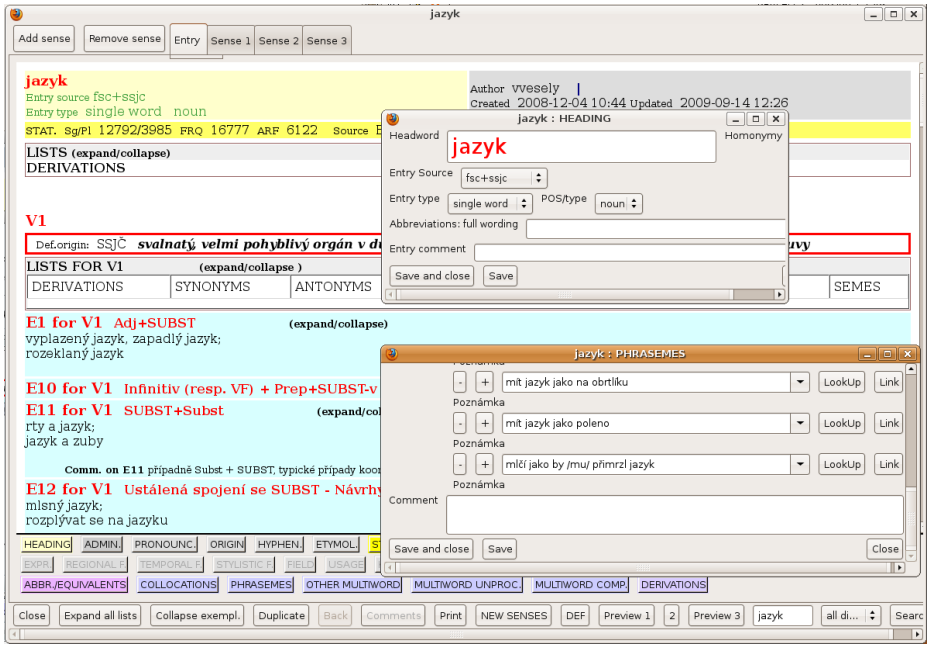

Fig. 3. PRALED: Entry view with two editor windows for the headword *jazyk* (*tongue*)

The PRALED client application currently allows to edit all the structural elements of the complex dictionary entry:

- part of speech and detailed information about the headword type
- orthoepy (spelling)
- pronunciation, hyphenation
- morphological properties (for the given POS) with the possibility to get these information from the morphological analyzer
- etymology and word origin
- statistical information, automatically inserted from corpora
- linked entries: abbreviations, collocations, phrases, components, synonyms, antonyms, hyponyms
- meaning explanation
- domain, temporal and spatial properties
- examples and corpora concordance

It is possible to link entries together, for example to refer to dialectic variants, related collocations, phrases or hyponyms. Users can easily open linked entries while editing the entry, they can also select whether to open the entries in the preview or edit mode.

To add the word usage evidence from the corpora, PRALED is connected with the Czech National Corpus [22]. Linguists are able to select several examples from the corpora and insert them to the edited entry, see Figure 4. Editor may also easily check

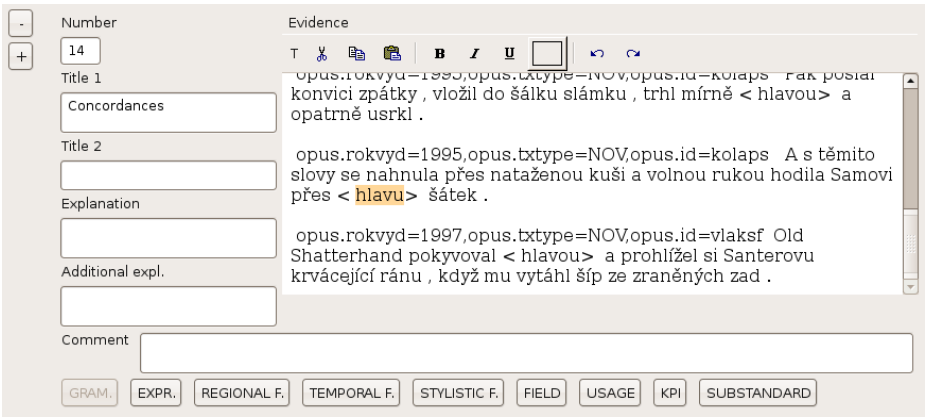

Fig. 4. PRALED: Corpora concordances for an entry

several Czech dictionaries and compare the information about the lexical entry from different sources.

Demonstration and more information about the project background are available at `http://deb.fi.muni.cz/praled/` and `http://lexiko.ujc.cas.cz/`.

5 Conclusions

We have described the design and implementation of a system for development of complex lexicographic database, denoted as PRALED lexicographic workstation. The system is actively used by tens of linguistic experts for preparation of a new modern electronic dictionary of the Czech language, which will form an invaluable resource both for human experts and for automatic computer processing.

PRALED is built on the DEB development platform, which provides many common features of dictionary writing systems in more than ten applications currently developed and used by linguistic experts as well as general public from all over the world. The freely available DEB server is currently installed in ten institutions from three continents and the main DEB server in Brno has more than 600 registered users from 19 countries.

The PRALED lexicographic workstation is oriented to complex processing of one language, the application design is however completely multilingual. Most of the PRALED features can be used without modifications for other languages. Language-specific features, like grammar information or morphology analysis, need to be modified, but fortunately changes to the application are easy and fast. Thanks to the modular design of the DEB platform, generic functions from PRALED, such as e.g. the Sketch Engine interface, are also used as servlet modules in other applications that are currently being developed and used for lexicographic work with tens of languages ranging through nearly all continents.

Acknowledgements

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 and 102/09/1842.

References

1. Havránek, B., et al., eds.: Příruční slovník jazyka českého (Reference Dictionary of Czech Language, PSJČ). Státní pedagogické nakladatelství/SPN, Praha (1935–1957) electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno.
2. Havránek, B., et al.: Slovník spisovného jazyka českého (Dictionary of Written Czech, SSJČ). 1^{st} edn. Academia, Praha (1960–1971)
3. McNamara, M.: Dictionaries for all: XML to final product. In: XML Conference 2003, Philadelphia, USA (2003)
4. Joffe, D., de Schryver, G.M.: TshwaneLex – professional off-the-shelf lexicography software. In: Third International Workshop on Dictionary Writing Systems: Program and List of Accepted Abstracts, Brno, Czech Republic, Masaryk University, Faculty of Informatics (2004) `http://tshwanedje.com/tshwanelex/`.
5. Erlandsen, J.: iLex - new DWS. In: Third International Workshop on Dictionary Writing Systems: Program and List of Accepted Abstracts, Brno, Czech Republic, Masaryk University, Faculty of Informatics (2004) `http://www.emp.dk/ilexweb`.
6. 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
7. Chaudhri, A.B., Rashid, A., Zicari, R., eds.: XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional (2003)
8. Feldt, K.: Programming Firefox: Building Rich Internet Applications with XUL. O'Reilly (2007)
9. Sedláček, R.: Morphemic Analyser for Czech. PhD thesis, Masaryk University, Brno, Czech Republic (2005)
10. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
11. 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/`.
12. Hlaváčková, D., Horák, A.: Verbalex – new comprehensive lexicon of verb valencies for czech. In: Proceedings of the Slovko Conference, Bratislava, Slovakia (2005)
13. 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)
14. 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
15. Vossen, P.: KYOTO Project (ICT-211423), Knowledge Yielding Ontologies for Transition-based Organization (2008) `http://www.kyoto-project.eu/`.

16. Hanks, P.: Corpus pattern analysis. In: Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud (2004)
17. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud (2004) 105–116
18. Fillmore, C., Baker, C., Sato, H.: Framenet as a 'net'. In: Proceedings of Language Resources and Evaluation Conference (LREC 04). Volume vol. 4, 1091-1094., Lisbon, ELRA (2004)
19. Herbst, T., Uhrig, P.: Erlangen Valency Patternbank. http://www.patternbank.uni-erlangen.de/ (2009)
20. Reaney, P., Wilson, R.: A Dictionary of English Surnames. Oxford University Press, Oxford Oxfordshire (1997)
21. Hanks, P., Hodges, F.: A Dictionary of Surnames. Oxford University Press, Oxford Oxfordshire (1988)
22. ICNC: Czech National Corpus - SYN2000 (2000) Accessible at WWW: http://www.korpus.cz.