DEB Platform Deployment
Current Applications

Aleš Horák and Adam Rambousek

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
hales@fi.muni.cz, xrambous@fi.muni.cz
http://deb.fi.muni.cz/

Abstract. In this paper, we summarize the latest development regarding the client dictionary writing applications based on the DEB II development platform. The DEB II framework is nowadays used in several full grown projects for preparation of high quality lexicographic data created within (possibly distant) teams of researchers.
We briefly present the current list of DEB II applications with the relevant projects and their phases. For each of the applications, we offer display the view of the interface with overview description of the most important features.

Key words: DEB platform, dictionary editor and browser, dictionary writing systems

1 Introduction

The Dictionary Editor and Browser (DEB) was first designed as a standalone program for writing dictionaries. After several problems with adaptation of the tool for coming new requirements, the second version, sometimes referred to as DEB II became a complete rewrite of the system based on open standards.

In the following text, we enlist the current state of DEB II applications. We believe that this list is the best evidence of the qualities of the framework as a whole together with several hundreds of DEB II users all over the world.

2 Current List of Implemented DEB Applications

In the following sections we present summary details of the particular DEB clients that are currently being implemented within the DEB platform.

2.1 DEBDict – General Dictionary Browser

This DEB client demonstrates several basic functions of the system:

– multilingual user interface (English, Czech, others can be easily added)

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007, pp. 3–11, 2007.

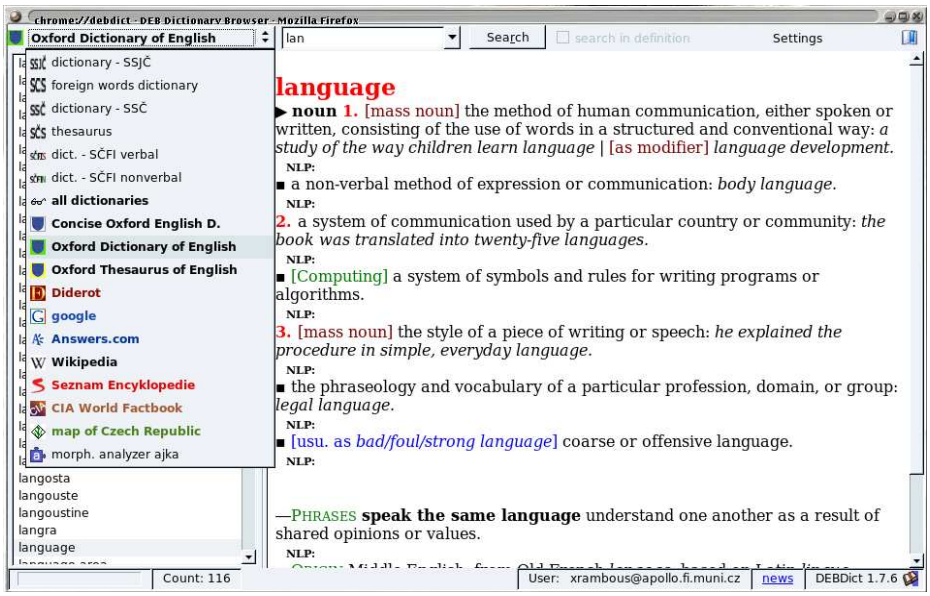

Fig. 1. The DEBDict common interface to several dictionaries with different structures.

- queries to several XML dictionaries (of different underlying structure) with the result passed through an XSLT transformation
- connection to Czech morphological analyzer
- connection to an external website (Google, Answers.com)
- connection to a geographical information system (display of geographical links directly on their positions within a cartographic map) or any similar application

The version of DEBDict that is currently running on our server provides a common interface to 7 dictionaries (see the Figure 1):

- the Dictionary of Literary Czech Language (SSJČ, 180,000 entries)
- the Dictionary of Literary Czech (SSČ, 49,000 entries)
- the Reference Dictionary of Czech Language (PSJČ, 200,000 entries)
- the Dictionary of foreign words (46,000 entries)
- the Dictionary of Czech Synonyms (thesaurus, 23,000 entries)
- two dictionaries of Czech Phrasal Words and Idioms (4,000 and 10,000 entries)
- the Diderot encyclopedia (90,000 entries)

As an addition, DEBDict features an interconnection to several web systems and the geographical system with the list of the Czech towns and cities.

DEBDict is also able to solve common problems with publication copyright for particular dictionaries – DEBDict supports individual user access rights.

This means that it is possible to display selected dictionaries only to a limited group of users.

2.2 DEBVisDic

DEBVisDic was one of the first applications built over the DEB II platform – it was designed as a completely new client-server tool for WordNet browsing and editing.

DEBVisDic uses new versatile interface (see the Figure 2) that allows the user to arrange the work without any limitations. Of course, DEBVisDic contains all the main features that were present in VisDic:

- multiple views of multiple WordNets
- freely defined text views
- synset editing
- hypero-hyponymic tree
- query result lists
- plain XML view of a synset
- synchronization
- inter-dictionary linking
- tree browsing
- consistency checks
- journaling

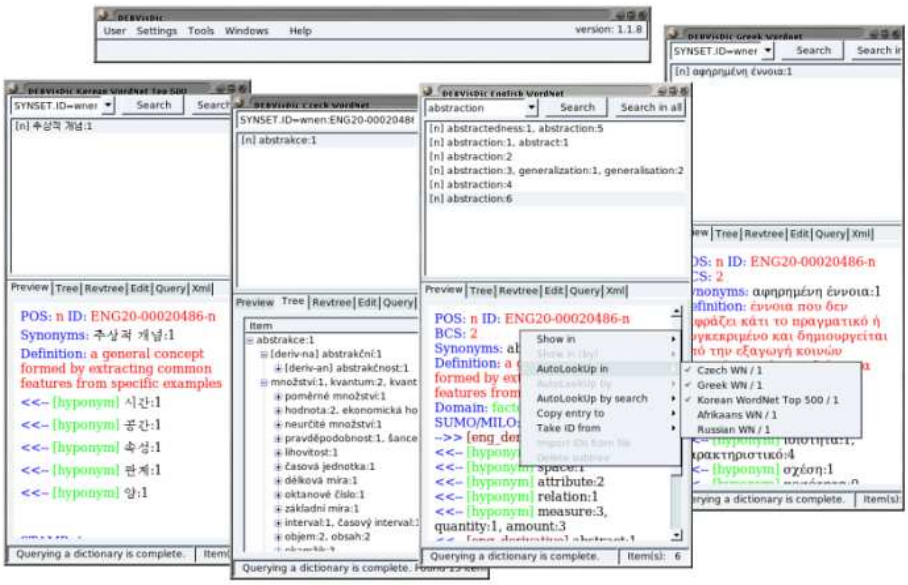

Fig. 2. The DEBVisDic main interface

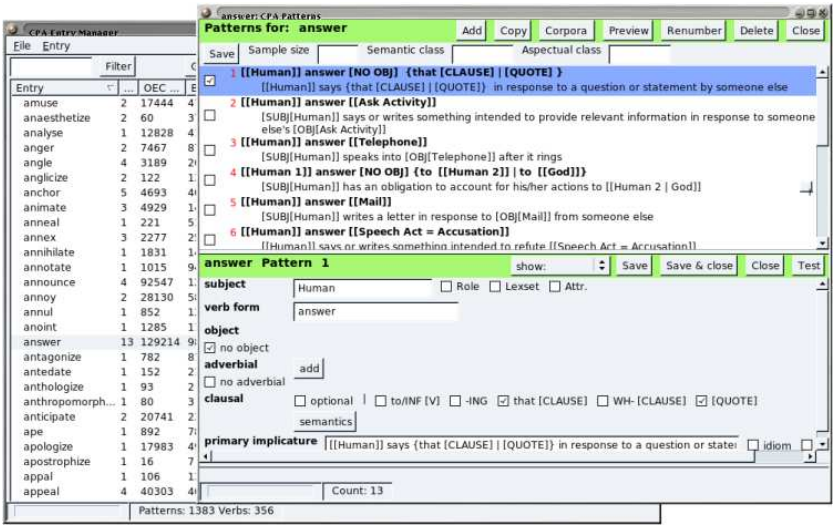

Fig. 3. The DEB CPA tool.

– user configuration

With the help of the DEB platform reusability, DEBVisDic will be supplemented with many new features that are currently accessible only as separate tools or resources. This functionality includes:

– connection to a morphological analyzer (for languages, where it is available)
– connection to language corpora, including Word Sketches statistics
– access to any electronic dictionaries stored within the DEB server
– searching for literals within encyclopedic web sites
– and many others

Currently, DEBVisDic is also used for preparation of new Polish, Hungarian, Slovenian, Dutch (within the Cornetto project), Nepalese and Afrikaans WordNets and it is proposed as the main tool for the prepared Global WordNet Grid.

2.3 DEB CPA Editor and Browser

Corpus Pattern Analysis (CPA, [1]) is a new technique for mapping meaning to words in text. No attempt is made in CPA to identify the meaning of a verb or noun directly, as a word in isolation. Instead, meanings are associated with prototypical sentence contexts. Concordance lines are grouped into semantically motivated syntagmatic patterns. Associating a "meaning" with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns.

CPA editing tool (see the Figure 3) displays the list of verb entries, along with the information who and when updated the entry. Each entry consists of several patterns (the number of patterns is not limited) and it is possible to freely modify their order and content. The main part of the tool, the pattern editing window, allows to enter and modify all the information about one pattern. The form is very versatile, e.g. it allows to add any number of subject/object alternations. The tool is connected to an on-line resource – it is possible to look up subject and object semantic type in Brandeis Semantic Ontology [2] which is hosted on a web server at Brandeis University. Examples documenting the pattern are taken from BNC using a modified version of Bonito2 corpus manager that is integrated to the DEB CPA tool.

2.4 DEB TEDI Terminological Dictionary Tool

The DEB TEDI client is the main tool used for preparation of a new terminological dictionary of 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 project is to build a terminological database consisting of about 5 000 dictionary entries which are classified into categories and supplemented with term definitions, translations info English, German and French, and with Czech usage examples. The resulting dictionary will be offered as a publicly available application directed especially to fine arts students.

2.5 The PRALED Lexicographic Station

This client is designed for the development of the Czech Lexical Database (CLD, denoted also as LEXIKON 21 [3]) and it serves as a main tool in preparation of the new comprehensive and exhaustive database of lexicographic information for the Czech language. The user's part of the PRALED tool is presently under the development in the Institute of Czech Language (ICL), Czech Academy of Sciences, Prague.

The PRALED system offers the following functionality:

- queries to several XML dictionaries (of different underlying structures), particularly to all relevant Czech dictionaries, i.e. SSJČ, SSČ, PSJČ, SCS, SČFI and DIDEROT (see [4,5,6,7,8]),
- editing existing or writing new dictionary entries. A lexicographer can use a set of forms which define the structure of the entry and fill in all relevant fields (see the Figure 4) which presently are:
  - 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)

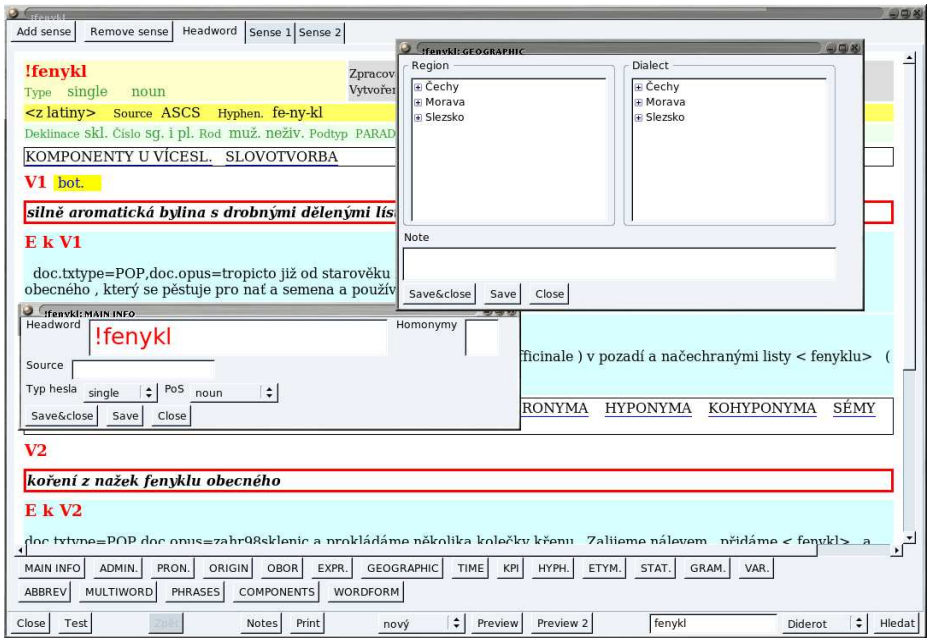

Fig. 4. The PRALED user interface

- etymological information
- integration with Czech morphological analyzer
- connection to an external website (Google, Answers.com)
- remarks and additional comments
- integration with the corpus manager Bonito2 and Word Sketch Engine [9], which allows a lexicographer to obtain the sorted individual word contexts including frequencies and statistical distribution parameters (salience).

2.6 Cornetto

The Cornetto project (STE05039) is funded by the Nederlandse Taalunie in the STEVIN framework. The goal is to build a lexical semantic database for Dutch, covering 40K entries, including the most generic and central part of the language. Cornetto will combine the structures of both the Princeton WordNet and FrameNet for English [10], by combining and aligning two existing semantic resources for Dutch: the Dutch WordNet [11] and the Referentie Bestand Nederlands [12]. The Dutch WordNet (DWN) is similar to the Princeton WordNet for English, and the Referentie Bestand (RBN) includes frame-like information as in FrameNet plus additional information on the combinatoric behaviour of words in a particular meaning. The combination of the two lexical resources will result

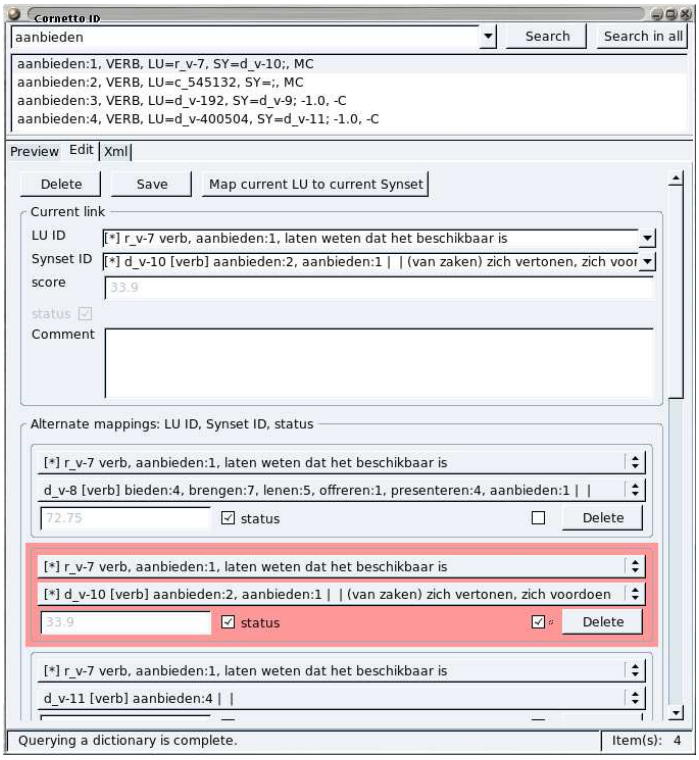

Fig. 5. Cornetto Identifiers window, showing the edit form with several alternate mappings

in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation and language-generation systems. In addition to merging the WordNet and FrameNet-like information, the database is also mapped to a formal ontology to provide a more solid semantic backbone.

The resulting data structure is stored in a database that keeps separate collections for lexical units (mainly derived from RBN), synsets (derived from DWN) and a formal ontology (SUMO/MILO plus extensions [13]). These 3 semantic resources represent different view points and layers of linguistic, conceptual information. The alignment of the view points is stored in a separate mapping table. The database is itself set up so that the formal semantic definition of meaning can be tightened for lexical units and synsets by exploiting the semantic framework of the ontology. At the same time, we want to maintain the flexibility to have a wide coverage for a complete lexicon and encode additional linguistic information. The resulting resource will be made available in the form of an XML database.

The Cornetto database (CDB) consists of 3 main data collections:

1. Collection of Lexical Units, mainly derived from the RBN
2. Collection of Synsets, mainly derived from DWN
3. Collection of Terms and axioms, mainly derived from SUMO and MILO

In addition to the three data collections, a separate table of so-called Cornetto Identifiers (CIDs) is provided, see the Figure 5. These identifiers contain the relations between the lexical units and the synsets in the CDB but also to the original word senses and synsets in the RBN and DWN.

Since one of the basic parts of the Cornetto database is the Dutch WordNet, we have decided to use DEBVisDic as the core for Cornetto client software. We have developed four new modules, described in more details below. All the databases are linked together and also to external resources (Princeton English WordNet and SUMO ontology), thus every possible user action had to be very carefully analyzed and described.

3 Conclusions

During the last three years, DEB II has been going through rapid development and several real applications for electronic dictionaries have been built. The free access to the Brno DEB server is nowadays in use by more than 250 users from 14 countries.

The DEB II server part is also available for download and is currently installed on 7 servers worldwide (Brno, Prague, Amsterdam-UvA, Amsterdam-VU, Poznan, Johannesburg and Budapest), where the DEB applications are used for national research projects.

Acknowledgments

This work has been partly supported by the Academy of Sciences of Czech Republic under the project T100300419, by the Czech Science Foundation under the project 407/07/0679 and by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Hanks, P.: Corpus Pattern Analysis. In: Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud (2004).
2. Pustejovsky, J., Havasi, C., Littman, J., Rumshisky, A., Verhagen, M.: Towards a Generative Lexical Resource: The Brandeis Semantic Ontology. In: Proceedings of LREC 2006, Genoa, Italy (2006) demo.
3. 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.

4. Petr, J., et al.: Slovník spisovného jazyka českého (Dictionary of Written Czech, SSJČ). 1^{st} edn. Academia, Praha (2002) electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno.
5. Filipec, J., et al.: Slovník spisovné češtiny (Dictionary of Literary Czech, SSČ). 1^{st} edn. Academia, Praha (1995) electronic version, LEDA, Praha.
6. Havránek, B., ed.: Příruční slovník jazyka českého (Reference Dictionary of Czech Language, PSJČ). Státní nakladatelství/SPN, Praha (1933–1957).
7. Kraus, J., Petráčková, V., et al.: Akademický slovník cizích slov (Academic Dictionary of Foreign Words, SCS). Academia, Praha (1999) electronic version, LEDA, Praha.
8. Čermák, F., et al.: Slovník české frazeologie a idiomatiky I-IV (Dictionary of Czech Phraseology and Idioms, SČFI). Academia, Praha (1983).
9. 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.
10. 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).
11. Vossen, P., ed.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks for European Languages. Kluwer Academic Publishers, Dordrecht (1998).
12. Maks, I., Martin, W., de Meerseman, H.: RBN Manual. (1999).
13. Niles, I., Pease, A.: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (2003) 412–416.