VisDic – Wordnet Browsing and Editing Tool

Aleš Horák and Pavel Smrž

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {hales,smrz}@fi.muni.cz

Abstract. This paper deals with wordnet development tools. It presents a designed and developed system for lexical database editing, which is currently employed in many national wordnet building projects. We discuss basic features of the tool as well as more elaborate functions that facilitate linguistic work in multilingual environment.

1 Introduction

Princeton WordNet became one of the most popular language resources. It is currently used in many areas of natural language processing such as information retrieval, automatic summarization, document categorization, question answering, machine translation etc. To integrate into the applications, many researchers work with the Princeton database and transform data to their own proprietary formats.

The Princeton team also developed a data browser for WordNet which can be downloaded together with English data from the web page `http://www.cogsci.princeton.edu/~wn/` both for Windows and UNIX platform. No WordNet editing tools are provided as the only instruments for majority of the lexicographic work in Princeton are standard text editors. The consistency of data is not therefore checked during the editing process itself, it is postponed to later phases.

Year by year the number of Princeton WordNet clones and WordNet-inspired initiatives increased. In 1998–1999 the EU project EuroWordNet 1 and 2 [1] took place, in which multilingual approach has dominated and WordNets for 8 European languages, particularly for English, Dutch, Italian, Spanish, French, German, Czech and Estonian, have been developed. The Interlingual Index (ILI), Top Ontology, set of Base Concepts and set of Internal Language Relations have been introduced as well [2]. These changes also led to the design and development of the new database engine for EuroWordNet and it resulted in the editing and browsing tool called Polaris [3].

In 2001 the EU project Balkanet [4] has been launched which can be viewed as a continuation of EuroWordNet project. It has been conceived as a multilingual as well and within its framework WordNets for 6 languages are being presently developed, particularly for Greek, Turkish, Romanian, Bulgarian, Serbian and

Czech. Before Balkanet has started it had already been obvious that Polaris tool had no future because its development had been closed and as a licensed software product (by Lernout and Hauspie) it had been rather expensive for most of the research institutions involved (typically universities). Moreover, the system had been provided only for MS Windows platform.

As the developers of Czech WordNet within EuroWordNet 2 project we came to the conclusion that a new tool for WordNet browsing and editing has to be developed rather quickly. At the same time we realized that it was necessary to look for the solution that would also support establishing the necessary standards for WordNet like lexical (knowledge) databases. Thus we decided to develop a new tool for WordNets based on XML data format, which can be used for lexical databases of various sorts. The tool is called VisDic and it has been implemented recently in Natural Language Processing Laboratory at Faculty of Informatics, Masaryk University for both Windows and Linux platform.

2 Basic Functionality

VisDic was developed as a tool for presentation and editing (primarily WordNet-like) dictionary databases stored in XML format. Most of the program behaviour and the dictionary design can be configured. With these capabilities, we can adopt VisDic to various dictionary types—monolingual, translational, thesaurus or generally linked wordnet lexicons.

2.1 Multiple Views of Multiple Wordnets

The main working window is divided into several dictionary panels. Each panel represents a place for entering queries and browsing context of one specified wordnet dictionary. The panels can display different wordnets as well as multiple contexts of the same dictionary.

The contents of a panel offers, besides the query input and matching results list, a set of overlapping notebooks tabs each of which represents one kind of view of the same entry from the list of results. The order, the type and even the content of each notebook tab is specified by the user in the configuration files (see 3.6). The main types of views are described in the following sections.

2.2 Freely Defined Text Views

The content of the Text View notebook tab is entirely built from the user definition that follows the XML structure of the wordnet entry. The editor can thus present an easily readable view of the entry with highlighting important parts of the entry content (see the Figure 1).

2.3 Edit

The editing capabilities allow to give the user a full control over the content and linking of each entry in the wordnet hierarchy. To prevent the user from

POS: n ID: ENG171-12836307-n
Synonyms: sunset:1, sundown:1
Definition: the time in the evening at which the sun begins to fall below the horizon
-->> [hypernym] *[n] hour:2, time of day:1
-->> [holo_part] *[n] evening:1, eve:4, eventide:1
-->> [near_antonym] [n] dawn:1, dawning:1, morning:3, aurora:1, first light:1, daybreak:1, break of day:1, break of the day:1, dayspring:1, sunrise:1, sunup:1, cockcrow:1
<<-- [near_antonym] [n] dawn:1, dawning:1, morning:3, aurora:1, first light:1, daybreak:1, break of day:1, break of the day:1, dayspring:1, sunrise:1, sunup:1, cockcrow:1

Fig. 1. An example of freely defined text view of wordnet entry

moving the entry as an object in the spider web of the linkage relations, the linguist rather specifies all the links in a textual dialog, where all the bindings are displayed in one place with consistency checks after each change request.

The actual contents of the Edit notebook tab is also entirely driven by the user instructions in the configuration, where each editing field is named and assigned to an XML tag in the entry.

2.4 Tree and RevTree

The wordnet dictionaries are specific by a heavy network of various kinds of relations between the dictionary entries with the function to capture the ontology relations on the underlying natural language.

The navigation in such environment is thus a crucial point of a successful linguistic work with wordnet data. Since the linkage relations generally do not need to obey any rules, that could make the resulting structure to be an arbitrary directed acyclic graph, or DAG. VisDic implements a browsing mechanism for general graphs. The navigation process works with two interconnected notebook tabs, which always both start at the same dictionary entry and display its position in the graph represented as a breadth-first path trees of all the linkage relations that lead from the entry to other entries in the dictionary. Each of the notebook tabs displays mutually opposite linkage relations, allowing the user to choose the direction of graph navigation in every step.

To facilitate the orientation and to help to position the entry in the wordnet hierarchy, the navigation also displays the path from the entry to its top in the hyper-hyponymical relation tree (see the Figure 2). For more advanced navigation the linguist may also use advanced tree browsing techniques (described in 3.3).

2.5 Query Result and External File Lists

Common actions in the wordnet creation and editing often include processing of a subset of entries based on certain criteria. VisDic offers a suitable kind of views for this situation, which allow to prepare a notebook tab with a list of entries

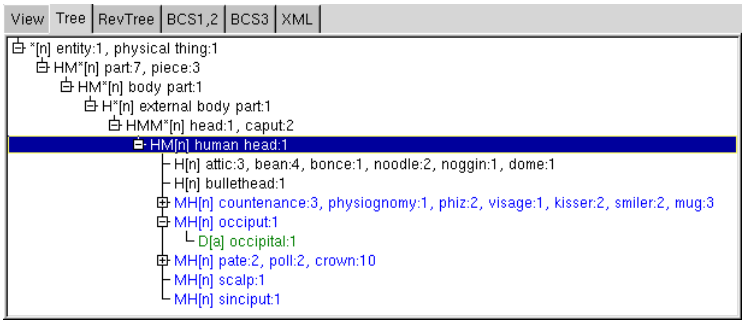

Fig. 2. The tree-like navigation in the wordnet linkage relations graph

matching any user specified query or a list of entries identified by entry-IDs gathered in a plain text file.

2.6 Plain XML View

Sometimes users need a thorough view into the data contained in the dictionary entry. XML View notebook tab offers this possibility. In this view, the user can see a graphically structured XML text, which represents the entry structure as it is stored in the dictionary.

3 Advanced Functionality

The basic functionality described in the previous section generally conforms to any XML based dictionary. However, linguistic work specialized to wordnet creation and editing requires some more specific and more sophisticated functions in the editor.

3.1 Synchronization

Within the creation of a national (e.g. Czech) wordnet, which would correspond to the English wordnet as a primary reference, one of the most frequent operation is a lookup of a dictionary entry (synset) from one wordnet in another dictionary. Such lookup uses either the SYNSET.ID tag (as a direct equivalent) or one of the, so called, equivalence tags (or attributes) defined in the configuration. An example of such tag may be REVMAP or MAPHINT used to help the linguist to process ambiguous link references between various versions of English wordnet.

The lookup function in VisDic can work in two modes: as an instant (one time) lookup — the *Show (by)* operation, and also as a firmly established link between two notebook tabs called the *AutoLookUp (by)*. In case of AutoLookUp, any move to another dictionary entry in the source notebook tab leads to an automatic lookup of the new entry in the destination tab. VisDic allows to have any acceptable combination of autolookups among all the notebook tabs.

3.2 Editing Support

The efforts of unifying national wordnets based on the English wordnet in many cases lead to copying of synset information between different language dictionaries. Such functionality in VisDic is splitted into two common situation — either the SYNSET.ID of an existing synset is to be unified with the ID of the English synset (*Take key from* operation) or a whole new entry is to be copied to another dictionary (*Copy entry to*).

3.3 Tree Browsing

The basic navigation in related synsets (in some cases reduced to the hyper- and hyponymical relations tree) is supplemented with two important wordnet operations — *Topmost entries* and *Full expansion*.

The Topmost entries operation identifies all synsets, which are (in the tree subset of linkage relations) found as the roots of relational hierarchy, i.e. are not hung below some other synset. This helps the linguist to identify the level 1 entries as well as so far unfiled entries.

The Full expansion allows the user to see all possible descendants of a selected synset in the linkage relations graph. During the operation cycle detection techniques check the violations of tree properties in the graph. Some relations can be also configured to be left out from the full expansion process.

3.4 Consistency Checks

Semi-automatic processing, which often takes part in the national wordnets creation, as well as common human processing of the data inevitably brings in the possibility of mistakes. The inconsistencies, which may be revealed as a duplicity, are controlled by VisDic consistency checks, which contain

- check duplicate IDs
- check duplicate literals and senses
- check duplicate synset literals
- check duplicate synset links

These checks allow the linguist to identify the most common errors e.g. after merging data from various sources.

3.5 Journaling

The work on a large and representative national wordnet usually employs more than one linguist working on the data. The synchronization of the resulting dictionary is made possible in VisDic with the usage of *journaling*.

During the work with VisDic, any changed to the data is marked in a journal file. Each journal file is specific to one dictionary and one user at a time. Such journal file can then be "applied" to the dictionary data and merged with the original. In this way, the simultaneous work of several linguists can be easily interchanged with a common data source.

3.6 XML configuration

Most of the functionality in the VisDic wordnet editor can be adopted to the local needs by means of its configuration files. All settings for the VisDic application are stored in several XML files.

The main configuration file (`visdic.cfg`) serves for global application data storage such as the list of dictionaries, the list of views, fonts, colors, histories, etc.

Besides this, each wordnet dictionary has its special configuration file (*dictionary*`.cfg`), which enables the linguist to set up most of the texts displayed in the application as well as the content of notebook tabs specific to the particular dictionary with respect to the XML structure of the entries.

4 Conclusions and Future Directions

VisDic, during its rather short history, has already proved its suitability for lexical database creation. The main power of VisDic manifests itself especially in development of highly interlinked databases such as wordnet. Its unique features have assured VisDic the leading role in many wordnet editing projects.

The development of such tool is never really closed. The future directions of our work will concentrate at specific support for linguists, improvements in the customization and user interface and team cooperation functionality. Entirely new horizons appear in the ongoing development of VisDic successor, the client-server lexical database editor DEB [5].

Acknowledgements

This work was supported by Ministry of Education of the Czech Republic Research Intent CEZ:J07/98:143300003 and by EU IST-2000-29388.

References

1. Eurowordnet project website, `http://www.illc.uva.nl/EuroWordNet/`.
2. Piek Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.
3. Louw M. Polaris user's guide. Technical report, Belgium, 1998.
4. Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`.
5. Pavel Smrž and Martin Povolný. Deb - dictionary editing and browsing. In *Proceedings of the EACL03 Workshop on Language Technology and the Semantic Web: The 3rd Workshop on NLP and XML (NLPXML-2003)*, pages 49–55, Budapest, Hungary, 2003.