# Editing of VerbaLex

Dana Hlaváčková and Vašek Němčík

NLP Laboratory, Faculty of Informatics
Masaryk University, Brno, Czech Republic
xnemcik@fi.muni.cz, hlavack@fi.muni.cz

**Abstract.** In this article we point out some problems in editing of the valency database VerbaLex. Editing is mostly manual and time-consuming work, requiring experienced annotators. We propose a solution in the form of a new editing interface, which can be used by new annotators without making unnecessary mistakes. Further, we also mention creating new and modifying existing tools (WordNet Assistant), which can accelerate the work on VerbaLex and eliminate errors in the editing to a minimum.

## 1 Introduction

VerbaLex represents an extensive database of Czech verbs valency frames. The database is currently being created at the Natural Language Processing Centre at the Faculty of Informatics, Masaryk University. VerbaLex is a work which lies between linguistics and the sphere of natural language processing. This is a special type of synchronous dictionary built with the help of computer tools and the use of electronic data sources.

## 2 VerbaLex

The organization of lexical data in VerbaLex is derived from the WordNet [1] 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, henceforth PWN, notation). For each of the synsets, a definition describing the meaning of the verbs is added. At present there are also links between Czech and English equivalent synsets in the PWN [2] (and similar in other languages), which are ensured by the so-called interlingual index.

Information about verbs is captured in VerbaLex in the following way. The head of each entry is constituted by synonyms with the numbers of their verb meanings. The synset notation also bears information about existing aspect pairs, and the abbreviation of verbal aspect (*pf*, *impf*, *biasp*). The valency frames follow with the description of valency and semantic roles for each slot and examples of usage. The entire entry is accompanied by various attributes of verbs. Item *use* shows the use of verbs in natural context – *primary*, *figurative* or *idiomatic*. The

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 15–20, 2010. 

type of reflexivity is marked for reflexive verbs. The last item refers to a system of semantic classes. This numbering is based on semantic classes by B. Levin [3] and was adopted from the project VerbNet [4], formulating a very detailed segmentation of English verb meanings into 395 classes. For our purposes it was reduced to 100 classes of Czech verbs. VerbaLex is available in TXT, PDF, XML and HTML formats.

## 3 Creating and Editing of VerbaLex

VerbaLex has been under development since 2005 and currently contains 10,482 verbs and 19,556 valency frames. 15 annotators and 6 technical support staff members have participated on the creation and editing of VerbaLex.

Originally, the valency database in its basic form could be edited using an interactive tool verbalex.sh, which is based on a well-configurable multi-platform editor GVIM. It allows easy insertion of language data in plain text. The advantage of the editor is syntax highlighting, which in our case, streamlines editing of the entire text and provides immediate control of the accuracy of recorded data (incorrect text is highlighted in color). The editor was specially adapted for creating VerbaLex and offers a variety of other ways to speed up and facilitate the work with a large database. The default format of valency frames was offered to the annotators and their task was to fill it with the actual data.

Creating a database of verb valencies represents a large amount of manual work which is commonly associated with unintentional errors. GVIM editor has been modified to indicate the maximum number of procedural errors. Besides the aforementioned coloring, an automatic check of formal errors in the file processed was added to the editor. It allows each user to individually review the selected section, edit and correct the majority of errors both in the format of an entry, and in the logical structure of the valency frame. Further errors are highlighted during the export of VerbaLex to the XML and HTML format. With the new version of the editor (VIM 7.0), automatic control of Czech spelling was added, which greatly facilitates the correction of errors and typos in the text parts of the database (e.g. definitions and examples).

## 4 Present Situation and Difficulties in Editing

Currently, editing VerbaLex is possible without having to open the large text file containing the database. After the annotator enters the verb, separate windows are opened with only the synsets containing entered verb. In this way we can also add new verbs to the database (when opening a blank window).

In spite of all efforts to ensure the correct editing of VerbaLex, it is still manual work with text files, which are prone to random errors that are difficult to detect. This type of errors cannot be found by the control in GVIM, because it was not designed to detect this type of inconsistencies. The errors are often discovered

only in the final version of VerbaLex, in HTML browser. The most common errors are:

- missing parts of text, such as tag for the verb in valency frame (VERB);
- accidental deletion of parts of text;
- bad copy of parts of the text.

The number of errors increases with each editing by inexperienced and untrained annotators. Moreover, we need to spend lots of time to explain the way of notation in the GVIM editor with formal errors control to new annotators.

For existing and new annotators is is also difficult to add new verbs. It is necessary to observe many rules in terms of both content and in terms of formal notation. Adding new verbs consists of several steps, which are an opportunity for inclusion of new errors. This process is time-consuming work and is performed manually.

For the verb, which is not yet listed in the database, it is necessary to, above all:

- find appropriate synonyms and build synsets;
- write a definition that covers all the meanings of the verbs in synset;
- create a valency frame indicating semantic roles and examples of use (frames are often not valid for the whole synset, but for individual verbs only, that are arranged in the so-called subsynset);
- add any other information for individual verbs (aspect, reflexivity, the type of use);
- include synset to the semantic classes;
- find an English equivalent synset in PWN and verify that it is not already linked to another Czech synset.

Our goal is to simplify the work and shorten the time required to add new verbs. One possibility is the use and adaptation of an existing tool, the WordNet Assistant, for searching English equivalents in PWN. Without its use, the annotators have to:

- translate Czech verb into English;
- find the English equivalent in the PWN (usually occurs in several synsets);
- choose the correct English synset;
- determine whether it is not linked to some other Czech synset (manual search in text version of VerbaLex).

This part of work has been simplified by creating a new version of the WordNet Assistant. Its use eliminates translation errors in selecting the right verbs and equivalents in PWN.

## 5 WordNet Assistant

Extending a valency lexicon such as VerbaLex is a complex task that cannot be performed automatically reliably enough. For humans, however, challenging and interesting as it is, larger-scale lexicographic work is rather tedious and humdrum. Human work may therefore often be rather slow and prone to mistakes.

This led us to implement a web-based application, the WordNet Assistant (henceforth WNA). It aims at assisting the human lexicographer at making various decisions common when adding new words or whole synsets to VerbaLex and thus speeds up the editing process. Moreover, it helps discovering and preventing certain types of inconsistencies.

The original form of WNA was designed to assist at adding whole new synsets by suggesting their probable English counterparts in the PWN. More information about it can be found in [5].

The currently relevant form of WNA concerns individual words and it gives the lexicographers the opportunity to view them in a broader context. Given a Czech word (and optionally its part-of-speech), it presents a list of relevant English synsets in PWN, supplemented by references to existing VerbaLex synsets.

The computation proceeds in two main steps:

- Czech-English dictionary lookup,
- PWN lookup using the DEB server.

The dictionary lookup is carried out using the GNU/FDL English-Czech Dictionary compiled at the Technical University in Plzeň [6]. In principle, any Czech-English dictionary available can be used in this step. The translations found in the dictionary are presented, and only those selected by the user are used in the subsequent step. This manual step accounts for disambiguation problems and possible noise in dictionary data.

The English translations resulting from the previous steps are passed as a query to PWN, with use of the DEB server [7]. A list of synsets containing the individual translations as literals is presented, synsets containing translations of more input words ranked higher on the list. The synsets on the list represent the range of relevant senses to be straightforwardly found in WordNet.

The synsets on the list are not accompanied only by their definition and examples, further, information about their respective counterparts in VerbaLex is included. This information is of great help when looking for inconsistencies, such as VerbaLex synsets that need to be split or merged. It also facilitates locating VerbaLex synsets the given word may be added to, or related senses that haven't been covered in VerbaLex yet.

The functionality and information provided by the application will be adapted based on the current needs and experience of the lexicographers.

## 6 New Editing Interface for VerbaLex

The question is, how to avoid errors in editing the database and speed up the work when adding new verbs. One possibility is to create a new user interface for editing of VerbaLex, which will be sufficiently clear and user-friendly for new and inexperienced annotators. User interface should be easily accessible and manageable (e.g. through a web interface), and should meet certain basic requirements that will ensure faster and more efficient editing of the database.

The requirements are:

1. clear graphic design of the interface with a simple navigation for users;
2. clearly separate the attributes of verbs and attributes of synsets;
3. add option to edit the individual verbs (with all their attributes), not all synsets only;
4. reduce manual text input to a minimum, leave this option only for definitions and examples (if possible, implement spell checking for these parts);
5. inputting attributes of verbs and synsets allowed only by selecting default options;
6. separate the left-side and right-side parts of valency frames;
7. check blank parts of the interface form;
8. preview of the finished entry;
9. record of the annotator's name and date of editing;
10. setting different levels of access rights (e.g. view, edit, add new entry).

This is a summary of the basic requirements, which should minimize the random formal errors that occur when handling a text file. The interface allows entering free text only in case of synset definitions and usage examples for valency frames, other items will be selected from a fixed menu of options. Transparent record of the names and dates allow us to check the work of annotators. Distinction of the access rights prevents accidental and unauthorized access to the database. When designing the actual interface we probably design other options that will ensure a simple and clear VerbaLex editing.

## 7 Conclusions and Further Work

A new editing interface should help to eliminate formal errors resulting from inattentive editing and accelerate the work on VerbaLex. With regards to this, it will also be necessary to change the existing XML format and a web interface, which could be supplemented by more options for searching information in VerbaLex. The WNA tool now allows efficient search of English equivalents in PWN and simplifies the addition of new verbs. The question remains, how to find and correct various inconsistencies in the database contents. It is often impossible to identify them automatically and they can corrected only by manual browsing of VerbaLex. One of our other tasks is identification of irregularities in the database content and finding ways of remove them efficiently.

**Acknowledgements**

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536.

## References

1. Fellbaum, C.: Wordnet. An electronic lexical database (1998).
2. Fellbaum, C.: English verbs as a semantic net. In: Five papers on WordNet, Princeton University (1990) Technical Report 43, Cognitive Science Laboratory.
3. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. The University of Chicago Press, Chicago (1993).
4. Palmer, M., Rosenzweig, J., Dang, H.T., Kipper, K.: Investigating regular sense extensions based on intersective levin classes. In: Coling/ACL-98, 36th Association of Computational Linguistics Conference, Montreal (1998) 293–300.
5. Němčík, V., Pala, K., Hlaváčková, D.: Semi-automatic linking of new Czech synsets using Princeton Wordnet. In: Intelligent Information Systems XVI, Proceedings of the International IIS '08 Conference, Warszawa, Academic Publishing House EXIT (2008) 369–374.
6. Svoboda, M.: GNU/FDL English-Czech dictionary (2008) `http://slovnik.zcu.cz/`.
7. Horák, A., Pala, K., Rambousek, A., Povolný, M.: DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In: Proceedings of the Third International WordNet Conference – GWC 2006, Brno, Czech Republic, Masaryk University (2005) 325–328.