Processing Czech Verbal Synsets with Relations to English WordNet

Vašek Němčík, Dana Hlaváčková, Aleš Horák, Karel Pala, and Michal Úradník

NLP Centre, Faculty of Informatics, Masaryk University
Brno, Czech Republic
{xnemcik,hlavack,hales,pala,xuradnik}@fi.muni.cz

Abstract. In the paper, we present the latest results of the process of final work on preparing the second stable version of the large lexicon of Czech verb valencies named VerbaLex.
VerbaLex lexicon is developed at the Masaryk University NLP Centre for the last three years. The current stage of this unique language resource aims at full interconnection with the pivot world semantic network, the Princeton WordNet. The paper describes the techniques used to achieve this task in a semi-automatic way.
We also describe interesting parts of preparation of a printed representative lexicon containing an important subset of VerbaLex with the valency description adapted to human readable forms.

1 Introduction

One of the most difficult tasks of automatic language processing is the analysis of meaning of language expressions or sentences. The central point of every sentence analysis is formed by its predicate part, i.e. by the analysis of the *verb* and *its arguments*. Within the process of design and development of the Czech syntactic analyser [1], this assumption has led us to the creation of a representative lexicon of Czech verbs and verb valency frames containing many pieces of information suitable for computer processing of the verb arguments from the point of view of their syntax as well as semantics (related to the content of the central semantic network of English, the Princeton WordNet [2]).

In the current text, we show the details of discovering relations between more than three thousands of new Czech verbal synsets and the respective English synsets, which finally allows to incorporate all the VerbaLex synsets to one of its original sources, the Czech WordNet [3].

2 VerbaLex Valency Lexicon

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

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 49–55, 2008.

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).

The basic valency frames (BVF) with stressed verb position contain various morpho-syntactic and semantic information. The types of verbal complementation (nouns, adjectives, adverbs, infinitive constructions or subordinate clauses) are precisely distinguished in the verb frame notation. The type of valency relation for each constituent element is marked as obligatory (obl) or optional (opt). BVF is followed by an example of verb usage in a sentence. Semantic information of verbal complementation is represented by two-level semantic roles in BVF. The first level contains the main semantic roles proposed on the 1st-order Entity and 2nd-order Entity basis from the EuroWordNet Top Ontology. The 1st-level semantic roles represent a closed list of 30 semantic tags. On the second level, we use specific selected literals (lexical units) from the set of Princeton WordNet Base Concepts with the relevant sense numbers. We can thus specify groups of words (hyponyms of these literals) suitable for relevant valency frames. This concept allows us to specify valency frames notation with a large degree of sense differentiability. The list of 2nd-level semantic roles is open, the current version contains about 1,000 WordNet lexical unites.

VerbaLex captures additional information about the verbs, which is organized in *complex valency frames* (CVF):

- definition of verb meanings for each synset;
- verb ability to create passive form;
- number of meaning for homonymous verbs;
- semantic classes;
- aspect;
- types of verb use;
- types of reflexivity for reflexive verbs.

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.

3 Linking New VerbaLex Synsets to WordNet

Extending a valency lexicon such as VerbaLex is a complex task. It does not only consist in editing valency frames of the individual verbs, but comprises also linking the data to other linguistic resources, and eliminating duplicates and inconsistencies. In the case of VerbaLex, each of the 3,686 newly added synsets is linked to its counterpart in the English Princeton WordNet (PWN) and to its hypernym in Czech WordNet. The linking procedure cannot be automated reliably and requires very costly human lexicographers to do a great amount of humdrum work.

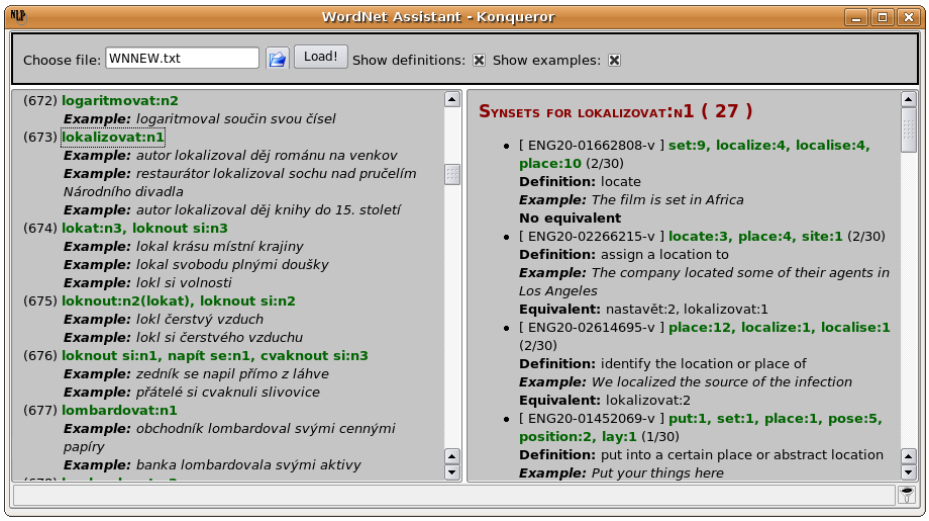

Fig. 1. WordNet Assistant

Based on our earlier experience, when linking new Czech synsets to PWN, the human lexicographers typically look up English translations for the individual synset literals. Subsequently, they use the translations to query the DEBVisDic WordNet browser and search for the corresponding English synset within the query results. To alleviate the human lexicographers from such routine tasks and to speed up the linking process, certain procedures have been semi-automated.

3.1 WordNet Assistant

WordNet Assistant (WNA) is a piece software that carries out certain steps, that would have to be otherwise carried out by a human lexicographer, automatically. This helps the lexicographer concentrate on crucial decisions and thus work more efficiently.

First, WordNet Assistant obtains individual literal translations by looking them up in a multi-lingual dictionary, in our case the GNU/FDL English-Czech Dictionary compiled at the Technical University in Plzeň [4]. Then, it uses these translations to query the English WordNet using the DEB server [5] interface, in order to obtain the relevant English synsets. On top of that, it sorts the English synsets yielded in the previous step according to their estimated relevance. More relevant synsets are presented to the user in a more prominent way (i.e. higher on the list). The user interface of WNA can be seen in Figure 1.

The heuristics used is not entirely accurate. It is based on the assumption that the more translations of literals of the original synset the English synset contains, the more relevant it is.

In addition to the above-mentioned functionality, WNA helps the lexicographer locate the hypernym of the new Czech synset in Czech WordNet. The hypernym can be suggested based on the already determined English counterpart in the following way:

- start at the English synset corresponding to the new Czech synset
- check whether the current synset has a Czech counterpart
- if there is no Czech counterpart, move to the hypernym of the current synset and continue with the previous step
- if there is one, it is the resulting related synset

Like in the case of sorting the English synsets according to their relevance, the suggested synset need not necessarily be the correct one. However, it is more than sufficient when it is close enough to lead the lexicographer to the relevant part of the hyper-hyponymic tree. It seems to be unavoidable anyway that the human lexicographer inspects and compares a number of close synsets before making the linking decision.

Generally, providing a hyper-hyponymic subtree of a reasonable size or a number of synsets, rather than a single one, helps prevent and detect inconsistencies. Given that the English synset may have been already linked with some other Czech synset, the lexicographer may consider revising the linking, merging the Czech synsets in question, or adding some distinctive features that would make it possible to link one of the synsets to a more specific English counterpart.

The work on the final version of linking the current VerbaLex synsets to PWN is reaching its end, however, no precise evaluation is available yet. We plan to analyze the information on the position of the synset chosen by the human lexicographer on the list presented by WNA, and based on that, to study the accuracy of the relevance heuristics. Nevertheless, for our purposes, it is more appropriate to evaluate the system in terms of time saved by the human lexicographers. Certain saving in time contributable to WNA is in principle guaranteed, because the lexicographers need to gather the information computed by WNA anyway.

3.2 Problems in Linking to PWN

The set of synsets newly added to VerbaLex contains a number of not particularly common verbs like “bručet” (“to grumble”) or “nachodit se” (“to have a lot of walking around”), for which it is extremely hard, or even impossible to find an English lexicalized counterpart. These synsets have been marked as “unlinkable”, comprise approximately 15 % of all synsets and can be divided into a number of categories:

- Perfective verbs (usually denoting an end of action)
  doletět (“to arrive somewhere by flying”), *dočesat* (“to finish combing”), *dokrmit* (“to finish feeding”)

- Reflexive verbs
  naběhat se (“to have a lot of running around”), *nalítat se* (“to have a lot of rushing around”), *načekat se* (“to do a lot of waiting”), *maskovat se* (“to disguise oneself”)
- Metaphoric expressions
  nalinkovat (“to line something for somebody” meaning “to plan/determine something for somebody”), *žrát* (“to eat somebody” meaning “to nag at somebody”)
- Expressive verbs
  ňafat se (“to argue”),
- Verbs with no direct English equivalent
  přistavit (“to park/stop a vehicle somewhere”)
- Verbs with no equivalent in PWN
  přepólovat (“to change the polarity of something”)

It should be remarked that similar problems have been already discussed during building the first version of the Czech WordNet in the course of the EuroWordNet project [6]. The issues of translatability between typologically different languages have been also touched in the Balkanet project where the notion of virtual nodes was suggested. They call for a special study.

Further, additional checks have been performed to detect duplicate VerbaLex synsets. Considering the size of the lexicon and the intricacies of Czech, duplicities cannot be completely prevented. Thanks to the linking of VerbaLex to PWN, it is possible to group semantically related synsets and further improve the quality of the data. Such synsets are for example: *baculatět:N1, buclatět:N1* and *kulatět se:N1, kulatit se:N1*. These synsets are both linked to the PWN synset “round:8, flesh out:3, fill out:6”,they are synonymous and need to be merged.

4 Compressed Human Readable Form of Verb Valency Frames

The process of presentation and checking of the verb valency expressions is a complicated task that can be partly done in an automatic manner, but in case of the need of a small subset with 100 % certainty of correctness the work must be done by human linguistic experts. That is why, for the purpose of preparing a printed book of the most frequent and/or somewhat specific verbs in VerbaLex, we have developed and implemented translation of VerbaLex frames to a compressed human readable form, see an example in the Figure 2. Such translation allows a human reader to grasp the meaning of the valency frame(s) in a fast and intuitive way.

The semantic roles in the frames need to be substituted with words from the particular language (Czech) and inflected in the correct form. This substitution is done by translating each role to Czech obtaining the lemmatized form of the role. The required inflected forms of the roles and verbs are then obtained using the morphological analyzer `ajka` [7]. This tool can generate all forms of an input

One VerbaLex frame for *připustit/připouštět* (admit):

```
ORG(person:1)who_nom VERB STATE(state:4)what_acc
GROUP(institution:1)who_nom VERB STATE(state:4)what_acc
ORG(person:1)who_nom VERB STATE(state:4)that
GROUP(institution:1)who_nom VERB STATE(state:4)that
```

The compressed human readable form:

```
V: člověk/instituce - připustí, připouští - stav
```

person/institution – admits – a state

Fig. 2. An example of a translation of the VerbaLex frames to a compressed human readable form.

word with the corresponding grammatical tags. With this process, the original VerbaLex frame from the Figure 2 is translated to:

```
V: člověk - připustí, připouští - stav
V: instituce - připustí, připouští - stav
V: člověk - připustí, připouští - stav
V: instituce - připustí, připouští - stav
```

This form is already suitable for human reading, however, we can see that there are duplicate and near-duplicate lines in the output. The "compression" of these forms of verb frames then follows the procedure:

1. remove exact duplicate frames;
2. frames containing phraseologisms are gathered in a separate set;
3. all remaining frames are compared 1:1 as candidates for joining a tuple in one new frame;
4. the best tuple is joined and forms a new frame
5. repeat from step 3 until no new frame can be created;
6. the result is formed by the compressed frames and the original set of frames with phraseologic phrases.

As a result of this procedure, we will obtain the compressed human readable frame for all the VerbaLex verb frames.

5 Conclusion

Within the final work on preparation of the second extended and improved version of the VerbaLex verb valency lexicon, we have designed and implemented the WordNet Assistant software tool. WordNet Assistant is aimed at supporting lexicographers in discovering new equivalents between (Czech) VerbaLex synsets and (English) Princeton WordNet synsets. In this paper, we present the process and problems in exploring such interlingual relations as well as their reuse for merging of very similar synsets.

Finally, for the purpose of presentation and checking of the valency frames, we have described an implemented tool that can offer a compressed human readable form of the VerbaLex verb frames.

We believe that the presented resources and implemented techniques prove that we have achieved new and valuable results in the description of the meaning of Czech verbs.

Acknowledgments. This work has been partly supported by the Academy of Sciences of Czech Republic under the projects 1ET100300414 and 1ET100300419 and by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

1. Horák, A.: Computer Processing of Czech Syntax and Semantics. Librix.eu, Brno, Czech Republic (2008).
2. Fellbaum, C., ed.: WordNet. An Electronic Lexical Database. MIT Press, Cambridge (1998).
3. Pala, K., Smrž, P.: Building the Czech Wordnet. Romanian Journal of Information Science and Technology **7**(2–3) (2004) 79–88.
4. Svoboda, M.: GNU/FDL English-Czech dictionary (2008) `http://slovnik.zcu.cz/`.
5. 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.
6. Vossen, P., Bloksma, L., Peters, W.: Extending the Inter-Lingual-Index with new concepts. (1999) Deliverable 2D010 EuroWordNet, LE2-4003.
7. Sedláček, R.: Morphemic Analyser for Czech. Ph.D. thesis, Masaryk University, Brno, Czech Republic (2005).