

Semi-automatic Linking of New Czech Synsets Using Princeton WordNet

Karel Pala, Dana Hlaváčková, and Vašek Němčík

Natural Language Processing Center, Masaryk University, Brno, Czech Republic

Abstract

In this paper we discuss extending Czech WordNet with verb synsets coming from the Verbalex database of Czech valency frames. One of the main tasks involved is linking newly added verb synsets to their Czech hypernyms and their English counterparts in Princeton WordNet. To spare the human lexicographers from tedious work, and to make the task more efficient, we have developed WordNet Assistant, a software tool that helps locate the relevant synset(s) in the already existing structures. According to our experience so far, this tool is of great advantage when incorporating new synsets into Czech WordNet, and we regard it as a worthwhile facility when extending WordNets also for other languages.

Keywords: WordNet, Verbalex, Czech, linking, translation, dictionary

1 Introduction

WordNet (henceforth WN) is one of the most widely known and universally used linguistic resources. Like every such resource, it has to be constantly improved and extended to fit the growing needs of contemporary NLP applications.

Extending WN poses several challenges. The main one lies in how to link new synsets to the already existing structures. This has two aspects. We may either wish just to link the new synset to its hypernym, or, when building WN for a language other than English, we may wish to link the new synset to its English counterpart, usually in Princeton WordNet (henceforth PWN). This work has to be done manually, is humdrum and very costly.

Ironically, in many cases, the most difficult phase of adding a new synset lies in locating the relevant part of the hyper-hyponymic tree. In other words, it may take more time to “start searching” than to make the actual linking decision.

To prevent this problem when extending the Czech WordNet, we decided to create a software tool, the WordNet Assistant (henceforth WNA), which helps locate the relevant synset(s) already existing in WordNet. It aims at alleviating the tedious part of the task the lexicographers need to address, so that they can concentrate on the actual linking decision, where human judgement is irreplaceable.

The next section introduces particular Czech lexical resources relevant to our work. Then, the following section describes the tool and methodology used to facilitate extending them. In the final section we sum up the ideas presented in this paper and sketch the future directions of our work.

SYNSET: BAVIT:1, ROZPTÝLIT:2, ROZPTYLOVAT:2
DEFINITION: poskytovat někomu zábavu/make (somebody) laugh

- *passive: yes*
- *meaning: I*
- *class: amuse-31.1-1*
- *impf: bavít:1 pf: rozptýlit:2 impf: rozptylovat:2*

frame: AG <person:1>^{obl}_{whoNom} VERB PAT <person:1>^{obl}_{whatAccus}
 ACT <act:2>^{opt}_{by doing whatInstr}

- *example: impf: bavil děti hrou (he amused the children by playing the game)*
- *attr: use: prim, reflexivity: obj_ak*

FIGURE 1: An Example of a Verbalex Valency Frame

2 Verbalex and Czech WordNet

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. Verbalex is based on three independent resources – electronic dictionaries of verb valency frames:

- BRIEF - a dictionary of 50,000 valency frames for 15,000 Czech verbs, which originated at FI MU in 1997 (Pala and Ševeček, 1997)
- VALLEX – a valency lexicon of Czech verbs based on the formalism of the Functional Generative Description (FGD) developed during the Prague Dependency Treebank (PDT) project (Žabokrtský, 2005)
- Czech WordNet valency frames dictionary created during the Balkanet project (Balkanet, 2001–2004) and containing 1,359 valency frames (incl. semantic roles) associated with 824 sets of synonyms (synsets)

The organization of lexical data in Verbalex is derived from the WordNet structure (Fellbaum, 1998), that is, it has a form of synsets arranged in the hierarchy of word meanings (hyper-hyponymic relations). The individual valency frames contain various syntactic and semantic information. For instance, they contain semantic annotation of each valency slot by means of a PWN synset, specifying that the valency slot in question can be filled only by its hyponym. An example of a Verbalex valency frame is given as Figure 1. Further details about the information contained in Verbalex valency frames were presented by Hlaváčková and Horák (2005).

The current version of Verbalex contains 7,063 synsets, 23,461 verb senses, 10,596 verb lemmata and 21,100 valency frames. Out of the total 7,063 synsets, 4,274 (i.e. 61 %) were added during building this valency database, the rest are already existing synsets adopted from the Czech WN.

The second resource relevant to our work is Czech WordNet. Its core was developed within the EuroWordNet project (Vossen *et al.*, 1998) and it was further extended in the course of Balkanet project (Balkanet, 2001–2004).

The hyper-hyponymic and most other relations in Czech WordNet have been built semi-automatically on the basis of Princeton WordNet (PWN). Throughout the development of Czech WN, it has been necessary to pay attention to the conceptual differences and lexical gaps between Czech and English. For instance, due to the rich formal and derivational morphology of Czech, there are many words with no straightforward equivalents in English. Especially the following phenomena contribute largely to this fact:

- verb aspect;
- reflexive verbs;
- verb prefixation (single, double, triple);
- diminutives (noun derivation by suffixation);
- move in gender (noun derivation by suffixation).

With regard to this situation, derivational nests, which can be considered as special subnets, have been added to Czech WN. This is a very effective approach to enriching Czech WN. We extended it in such a way by approximately 30,000 synsets (Pala and Hlaváčková, 2007).

The current version of Czech WordNet comprises 56,104 synsets (31,029 nominal, 5,158 verbal, 19,670 adjectival and 247 adverbial) and is being further extended, for instance, by synsets created within the Verbalex project.

3 WordNet Assistant

WordNet Assistant is a tool designed to help the lexicographers link new synsets created within the Verbalex project to the already existing structures. Above all, each new synset has to be linked to its hypernym in Czech Wordnet and to the corresponding synset in the English Princeton WordNet. This involves a considerable amount of time the lexicographers spend doing humdrum things which can be done automatically, or at least semi-automatically. We believe there is a worthwhile potential of speeding up the editing process.

According to our experience, the typical routine activity the human lexicographers do when linking new Czech synsets to PWN, is looking 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.

WordNet Assistant is a piece software that performs these steps for the lexicographer automatically. It queries a multi-lingual dictionary, in our case the GNU/FDL English-Czech Dictionary compiled at the Technical University in Plzeň (Svoboda, 2008). Next, it uses the resulting translations to query the English WordNet using the DEBServer (Horák *et al.*, 2005) interface to obtain the relevant English synsets. Moreover, the synsets are sorted according to their estimated relevance – more relevant synsets are presented to the user in a more prominent way (i.e. higher on the list).

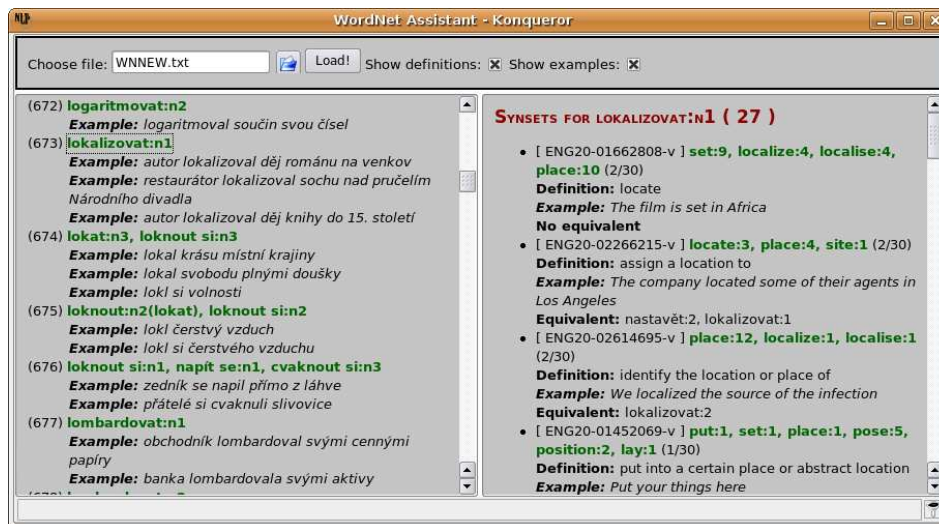


FIGURE 2: WordNet Assistant

Of course, the computed order is only approximate. The underlying heuristics is based on the assumption that a synset containing translations of more literals of the original synset is more likely to be related to it. We plan to improve on this heuristics, nevertheless, even this setting allows the human lexicographer to make the linking decision in much shorter time than with manual search. The user interface of WNA can be seen in Figure 2.

In addition to the above-mentioned functionality, WNA assists the lexicographer with the follow-up linking of the new Czech synset to its hypernym in Czech WordNet. This is done based on the chosen PWN synset. We believe that given the English counterpart of the synset, the most relevant synset in Czech WordNet (which is not as large as PWN) can be obtained 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

The situation is illustrated by Figure 3. The resulting synset is very likely the closest more general concept related to the new Czech synset. However, it need not necessarily be. 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. It may reveal for instance that the English synset corresponding to the newly

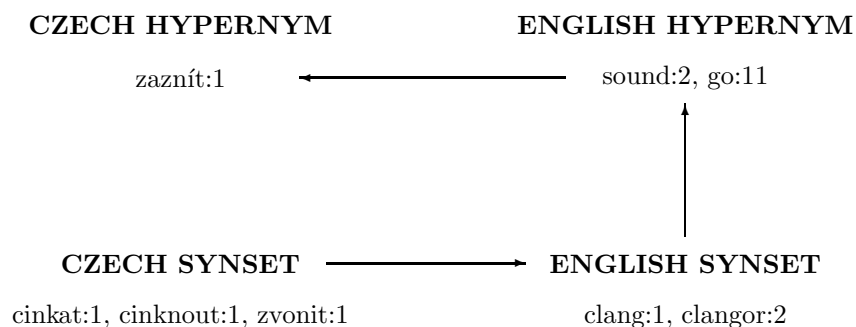


FIGURE 3: Approximating the hypernym of a Czech synset

added synset is already linked to some other synset in Czech WordNet. Such a situation may lead the lexicographer to the decision to revise the existing link, merge synsets, or on the contrary, split it into more distinct ones.

The main idea behind WNA is not limited to Czech and English and it would be worthwhile to employ it within projects for other suitable language pairs. The only requirements are that the languages in question should not be too different and that a bilingual dictionary that can be plugged into the system is available for them. The development of WNA is still in progress and we experiment with it at the moment. Consequently, no precise evaluation is available yet.

As the adding of new synsets to WN proceeds, we plan to log the synsets chosen by the human lexicographers together with the position on the list presented by WNA. Analysis of this data will reveal more precisely how accurate the relevance heuristics is. Nevertheless, for our purposes, it is more appropriate to assess the system in terms of time the human lexicographers save. Significant saving in time contributable to WNA is in essence guaranteed, because the lexicographers need to gather the information computed by WNA anyway. Moreover, the straightforward alternative, manual search, is very slow and does not yield the necessary information in a form as transparent as the WNA output. This further slows down the subsequent decision-making process.

As we are dealing with not particularly common verbs like “bručet” (“to grumble”) or “nachodit se” (“to have a lot of walking around”) at the moment, it sometimes occurs that WNA does not yield any results. The reason is that either none of the literals of the initial synset is present in the dictionary, or due to lexical or conceptual differences between Czech and English, it is not possible to reach an English counterpart by means of a straightforward translation. Such delicate cases need to be handled manually anyway, and the failure of WNA does not hold up the lexicographer for more than a couple of seconds. This is negligible compared with the amount of humdrum work and time WNA saves for other synsets.

Overall, our experience so far indicates that employing WNA is very promising with regard to the efficiency of the human work involved in extending Czech WN.

4 Conclusions and Future Work

In this paper we have discussed extending Czech WordNet with new verb synsets taken from the Verbalex project. We have presented WordNet Assistant, a software tool that for each new synset helps locate the relevant synset(s) in PWN and the already existing part of Czech WN. We have sketched the functionality of this tool and discussed the impact on the efficiency of the human work involved in extending Czech WN.

As incorporating the verbal synsets from Verbalex to Czech WN is still under progress, we are gathering feedback from the lexicographers, and based on that, plan to improve the tool. Further, we plan to experiment with approximating the linking decision done presently by humans automatically using various heuristics.

Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the project 1ET200610406, by the Ministry of Education of CR within the Center of Computational Linguistics LC536, and in the National Research Programme II project 2C06009.

References

- BALKANET (2001–2004), Balkanet project website, <http://www.ceid.upatras.gr/Balkanet>.
- Christiane FELLBAUM, editor (1998), *WordNet. An Electronic Lexical Database*, MIT Press, Cambridge.
- Dana HLAVÁČKOVÁ and Aleš HORÁK (2005), Verbalex – New Comprehensive Lexicon of Verb Valencies for Czech, in *Computer Treatment of Slavic and East European Languages, Third International Seminar*, pp. 107–115, VEDA, Bratislava.
- Aleš HORÁK, Karel PALA, Adam RAMBOUSEK, and Martin POVOLNÝ (2005), DEBVis-Dic - First Version of New Client-Server Wordnet Browsing and Editing Tool, in *Proceedings of the Third International WordNet Conference - GWC 2006*, pp. 325–328, Masaryk University, Brno, Czech Republic.
- Karel PALA and Dana HLAVÁČKOVÁ (2007), Derivational Relations in Czech WordNet, in *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pp. 75–81, Association for Computational Linguistics, Prague, Czech Republic.
- Karel PALA and Pavel ŠEVEČEK (1997), Valence českých sloves, in *Proceedings of Works of Philosophical Faculty at the University of Brno*, pp. 41–54, MU, Brno.
- Milan SVOBODA (2008), GNU/FDL English-Czech Dictionary, <http://slovník.zcu.cz/>.
- Piek VOSSSEN, Laura BLOKSMA, Horacio RODRIGUEZ, Salvador CLIMENT, Nicoletta CALZOLARI, Adriana ROVENTINI, Francesca BERTAGNA, Antonietta ALONGE, and Wim PETERS (1998), The EuroWordNet Base Concepts and Top Ontology, in *Technical Report Deliverable D017, D034, D036, WP5 EuroWordNet, LE2-4003*, University of Amsterdam, Amsterdam.
- Zdeněk ŽABOKRTSKÝ (2005), *Valency Lexicon of Czech Verbs*, Ph.D. thesis, MFF UK, Prague.