WDS'07 Proceedings of Contributed Papers, Part I, 125–130, 2007. ISBN 978-80-7378-023-4 © MATFYZPRESS

Selected Sense Enumerated Lexical Resources for Czech

E. Bejček
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Abstract. In this paper we present three quite different approaches to word senses description in three particular lexicons. The advantages and disadvantages of these approaches are mentioned. We have done some practical experiments with all of them. These experiments—including machine learning and manual annotation—are briefly described. At the end, we conclude by comparing those three lexicons.

1. Introduction

1.1. Motivation

The matter of our interest is a lexicon describing word senses. This type of lexicon is required for sense tagging task. It is a special case of wider word sense disambiguation (WSD) area of Natural Language Processing (NLP). The goal of such a task is to assign a lexicon entry to every relevant token in a given (possibly plain) text. It is obvious that we need some lexicon to do this.

If this task is well done, it can help in many areas of NLP such as machine translation, information retrieval, it improves results of question answering etc. It is generally supposed in all these tasks that better understanding of text (which requires also word sense discrimination) leads to better efficiency of systems.

1.2. Lexicons and their problems

There are several approaches to automatic WSD (e.g. automatic clustering of word senses without any knowledge of the language), the one we deal with is based on the sense enumerated lexical resources. The main part of this paper consists of description of three different word sense lexicons for Czech language.

Let us say we want to build a statistical system for automatic word sense disambiguation. We need two resources: a lexicon of word senses and annotated data. We need large amount of manually annotated data for training a statistical system. But even if our system is not based on machine learning, we need some annotated data at least for evaluation of the results.

We need to choose a lexicon and let a human annotator annotate certain amount of data on its base. Now we will discuss the problem of choosing the lexicon.

The main problem for all lexicographers is to set the number of senses for every word (of their interest) and moreover to identify the distinction between two related senses. It seems that very fine-grained segmentation used in some comprehensive lexicons misses these hard borders. Thus usage of these lexicons is very difficult in practical disambiguation. On the other hand, there is a number of coarse-grained lexicons, where only sufficiently different meanings with clear borders are considered as two senses. It however often ends in such a mass joining neighbouring senses that the original purpose—to classify all senses—is lost. Words with one or two meanings are the result of it.

In the Section 2 we introduce the Czech WordNet and Word Sense Disambiguation experiments with it. Section 3 is devoted to the PDT-VALLEX and experiments based on valency frames. The fourth Section is focused on the current project Lexemann. In the last section we conclude and show some advantages and disadvantages of each of these three approaches.

2. Czech WordNet

The name WordNet is used for a project, which has been developing from 1978, as well as for the output of this project, an ontology. WordNet is organized by the concept of synonym sets (*synsets*), groups of words that are roughly synonymous in a given context. The glossary definition and the example sentences are shared among all synonyms in a given synset. One synset (usually with more than one word) represents one sense. These synsets are interlinked by several relations, such as hyponymy, hyperonymy, holonymy, meronymy, antonymy and others.

Synsets can be seen as nodes in the space and relationships between pairs as labeled edges. So whole WordNet is a directed graph. It is though very complex graph with more than one component (because of several parts of speech).

Original WordNet was created at the Princeton University (*Fellbaum* [1998]) for English. Many variations (especially for other languages) have been built in the following years. Czech WordNet came into existence as a part of project EuroWordNet.[1] Everything we mentioned about WordNet holds as well for Czech WordNet. There were 17,000 nouns, 4,000 verbs, 2,000 adjectives and 200 adverbs in our version of the Czech WordNet.

2.1. WSD with Czech WordNet

Czech WordNet has been used for manual annotation of the *Prague Dependency Treebank* (PDT, *Hajič et al.* [2001]). PDT is a corpus annotated on morphological, analytic and tectogramatical layer (for more information see *Hajič* [2005]). The data is divided into three segments according to the depth of the manual annotation. The deeper annotation the lesser amount of data is annotated. The annotation project is described in *Hajič et al.* [2004]. Two human annotators independently processed data of PDT and they had to assign WordNet synset for each word that have more senses (i.e. is included in more synsets). Finally, only those occurrences with agreement of both annotators was considered as correct.

Having this data, we could train an automatic classification of word senses based on decision trees. We could not exploit the richest available annotation of PDT (namely t-layer of PDT), because the intersection of this data with Czech WordNet annotation data was too small. We have used only analytical trees (a-layer of PDT), which nearly quadruplicate the amount of usable data.

We had continuous text with almost 150,000 annotated words, 78 % of them had more than one meaning according to Czech WordNet. As a baseline, we used the most frequent synset for each word. See Table 1. We chose about 40 features as described in *Bejček* [2006]. Then we trained decision trees with these features in C4.5 system[2]. The accuracy was 91.4 % for nouns and 93.9 % for adjectives. Although the baseline was high, these results represent 13.7 % and 10.0 % improvement against the baseline.

Table 1. Results for decision trees on Czech WordNet senses

results	nouns	adjectives
baseline	90.02 %	93.24 %
decision trees	91.39 %	93.91 %
improvement	13.7 %	10.0 %

Table 2. Results for decision trees on PDT-VALLEX frames

results	nouns
baseline	84.92 %
decision trees	86.91 %
improvement	13.20 %

[1] Czech WordNet was developed at the Masaryk University in Brno, see *Smrž* [2003].

[2] Author's homepage: http://www.rulequest.com/Personal

2.2. Improvement of annotation

The inter-annotator agreement mentioned above was not very high, exactly 61.7 %. That provided only 5,000 fully disambiguated sentences.[3] *Bejček, Möllerová, and Straňák* [2006] describe the process of improvement of the quality of annotation. Conflicting lemmas (i.e. lemmas, where annotators had not agreed[4]) were inspected. These lemmas, sorted by their frequency, had exponential character as Figure 1 shows. Hence, 3,700 lemmas out of 4,700 had less than 10 occurrences (right part of the figure) and there were only 25 lemmas on the other end that had more than 200 error occurrences.

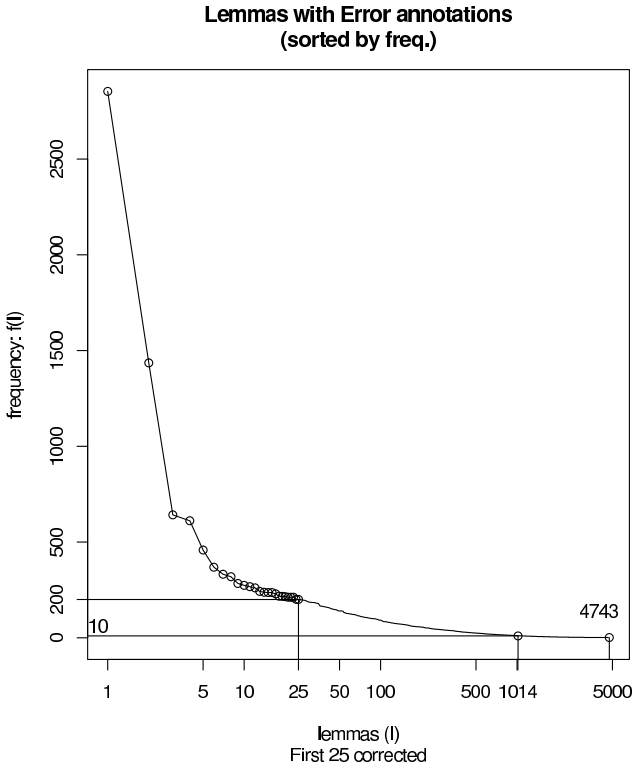

Figure 1. Frequency of errors for separate lemmas

The decision was made, that only these 25 lemmas would be processed: corresponding synsets in the Czech WordNet were improved and all their occurrences in the text were corrected. The third annotator, called *corrector*, provided that. Corrector's task was to check through all non-agreements and choose the correct solution. She had an improved lexicon, both annotators' results, well arranged environment and sentences sorted into exception classes. All these facts speeded up and simplified the correction process.

Let us summarize — by correcting of 0.5 % lemmas, we gained 7.4 % correctly annotated words. That raised the number of fully disambiguated sentences from 5,000 to 7,000.

---

[3]Sentences, where the annotators agreed for all ambiguous words. (It means ambiguous in the Czech WordNet.) These sentences could be processed completely then which is the advantage against separate words.

[4]It should be mentioned that there were many types of non-agreement, because the annotators could use also ten exceptions for cases such as "incorrect reflexivity", "figurative use", "missing more general sense" or "unclear Czech WordNet sense". Either if they had agreed on the type of an exception, it could not be considered as an agreement. All combinations between exceptions, between exceptions and synsets, and between two different synsets are marked as *non-agreement*.

3. PDT-VALLEX

The second lexicon we have worked with is valency lexicon of Czech verbs PDT-VALLEX introduced in *Hajič et al.* [2003]. *Valency* of a verb is the range of syntactic elements either required or specifically permitted by it. Each such element can be labeled with its function in a sentence, e.g. *actor*, *patient*, *effect*, *obstacle* etc. Following example illustrates this.

1. John$_{\mathrm{ACT}}$ stood$_{verb}$ $_{(}$his toy soldiers$_{)\ \mathrm{PAT}}$ $_{(}$against the wall$_{)\ \mathrm{LOC}}$.
2. John$_{\mathrm{ACT}}$ stood$_{verb}$ $_{(}$on the road$_{)\ \mathrm{LOC}}$ (...hitchhiking).

In the first example (which is first meaning of the word "stand" at once), patient is obligatory, whilst there is no place for any patient in the second example. Roughly speaking, sets such as ACT, PAT, (LOC) in (1.) and ACT, (LOC) in (2.) are called *valency frames*. We can see that different valency frames usually represent different senses. That is the main idea of the valency lexicon PDT-VALLEX and its usage as a sense enumerated lexicon.

The PDT-VALLEX lexicon was created during the annotation of PDT 2.0. Therefore it contains every verb and verbal noun and adjective, which occurred in t-layer of PDT and every such word has several (1.5 in average) entries in the lexicon according to its different valency frames. Moreover, the valency frame of those lexical entries is sufficient enough for distinguishing between two occurrences of the word in the text—at least in the text of t-layer of PDT. Hence, the annotation should be explicit.

3.1. WSD with PDT-VALLEX

We carried a WSD experiment on this data, too. As a method, we again used supervised learning and decision trees. But we examined only nouns this time. We did it so, because the experiment with verbs was shown in *Semecký* [2005] and a representation of adjectives was too small to train anything on it (there was 100 % baseline). The results was similar to the previous experiment (as can be seen in Table 2).

4. Project Lexemann

These problems lead us to start a project of "lexico-semantic annotation" or also "lexeme annotation", Lexemann for short. See *Straňák and Bejček* [2007] for details. It is, by some means, a descendant of the project described in Section 2.1.

4.1. Goal

The goal of the Lexemann Project is to obtain annotated data together with useful lexicon of annotated senses. At the end every autosemantic word or whole phraseme should be assigned an entry from the new lexicon called *SemLex*. But in the current step we concern only on multi-word expressions.

The SemLex lexicon is supposed to contain—at the end of the first step—every multi-word expression found in the data as at least one entry, i.e. in at least one meaning. Every lexicon entry consists of its basic form (e.g. "pevná půda pod nohama", lit. "immovable$_{\mathrm{fem}}$ soil beneath feet$_{\mathrm{dual,\ instr}}$"); lemmatized form, in which every word from the expression is lemmatized separately ("pevný půda pod noha", lit. "immovable$_{\mathrm{lemma}}$ soil beneath foot$_{\mathrm{lemma}}$"); part of speech of whole phraseme; a morphological tagset for each word; and a structure describing a tectogramatical subtree, which represents the phraseme. The last named item—a tectogramatical subtree—is heavily used in autoannotation (see bellow). Then there are some other items in the lexicon entry, such as gloss, example, modifier of an entry (i.e. the person, who last edited it), synonyms etc.

4.2. Approach

Since manual annotation of lexical meanings is a hard and time-consuming problem (as was verified in annotation project mentioned in Section 2.1), we decided to start with a smaller part of a problem: multi-word expressions. It is useful to separate just this part, because the range of this task is easy to specify. And anyhow, multi-word expressions must be solved first, because otherwise they complicate the rest of annotation.

Besides multi-word phrasemes we also annotate multi-word named entities, such as names of companies or structured addresses. There are two reasons for this decision. Firstly, both company name and address should be processed as a single unit, so our annotation provides a kind of a container for it. Secondly, company name could be considered as a phraseme of its kind. But all those company names cannot be placed into the lexicon; instead of it, they could be processed automatically on the basis of this manual annotation of named entities.

Three present lexicons for Czech were merged together as a first source of SemLex entries: Czech WordNet (*Smrž* [2003]), Eurovoc (*Eurovoc* [URL]) and SČFI (*Čermák et al.* [1994]). The entries were slightly modified to suit the SemLex format.

Project Lexemann employs two annotators who go through a text, annotate it (like in project mentioned in Section 2.1.) and build a lexicon in accordance to the text (like in project PDT-VALLEX in Section 3.) One difference is that they assign a lexicon entry to every autosemantic phraseme (whilst in Section 2.1 they were limited by the Czech WordNet and in Section 3. by the scope of valency—i.e. only verbs and some nouns).

Annotators use special tool with GUI, which enables to show surface structure of the sentence, but creates annotation marks on the tectogramatical structure. It also provides a log of separate annotator's operations, persistent undo/redo, and some auto-annotation based on searching for the known tree structures from SemLex in the newly opened document.

Projects Lexemann is a work-in-progress now, therefore no statistical experiments were tried on it yet.

5. Conclusion

WordNet is a large lexicon with many words of four parts of speech, but also with many unclarities and errors. Czech WordNet as a lexicon for annotation was developed before the annotation project started and was not changed in any way during the annotation itself.

PDT-VALLEX with its precise methodology to distinguish frames, i.e. senses, has relatively clear borders when judging among frames. Unfortunately, this methodology suits only for verbs and other verbal words.

Each of these lexicons has its disadvantages. Czech WordNet has problems when we want to use its poor information to judge among more possible synsets. PDT-VALLEX is more precise in this way; its weakness is that its methodology can be used only for verbs and other verbal words—that means small subset of all autosemantic words we need to process.

During the work with **SemLex**, annotators have explicit guidelines telling them how to annotate. Entries in the lexicon are as clear as possible. It could join advantages from both WordNet and PDT-VALLEX.

But it suffers a little from the method of creation. At the present time, SemLex is still a huge lexicon contaminated with many marginal phrasemes from Czech WordNet, Eurovoc and SČFI (e.g. "Transport Department", "Department of the Exchequer", "Defense Department" etc.) It is questionable if they should rest in the lexicon.

Acknowledgments. The present work was supported by the MŠMT, Object of research (Výzkumný záměr) MSM0021620838.

References

Bejček, E., Automatické přiřazování významu –"Sense-tagging". Master thesis. Charles University, Prague, 2006.

Bejček, E., Möllerová, P., and Straňák, P., The Lexico-Semantic Annotation of the Prague Dependency Treebank: Some results, problems and solutions. In: *Proceedings of the 9th International Conference, TSD 2006*, Springer-Verlag Berlin Heidelberg, pp. 21-28, 2006.

Čermák, F., Červená, V., Churavý, M., Machač, J., Slovník české frazeologie a idiomatiky. Academia, 1994.

Eurovoc, http://europa.eu/eurovoc.

Fellbaum, Ch., WordNet: *An Electronic Lexical Database*. MIT Press, 1998.

Hajič, J., Complex Corpus Annotation: The Prague Dependency Treebank. In: *Insight into Slovak and Czech Corpus Linguistics*, pp 54-73, Veda Bratislava, Slovakia, 2005.

Hajič, J., Holub, M., Hučínová M., Pavlík, M., Pecina, P., Straňák, P., Šidák, P. M., Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank In: *LREC 2004*, Lisbon, 2004.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P., PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pp 57-68, Vaxjo, Sweden, 2003.

Hajič, J., Vidová-Hladká, B., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Prague Dependency Treebank 1.0 (Final Production Label) In: *CDROM*, 2001.

Semecký, J., On Automatic Assignment of Verb Valency Frames in Czech. In: *Proceedings of Language Resources and Evaluation*, LREC 2004, pp. 1941-1944, 2006.

Smrž, P., Quality Control for Wordnet Development. In: *Proceedings of the Second International WordNet Conference—GWC 2004*, pp 38-41, Cairo, 2004.

Straňák, P., Bejček, E., Annotation of Multiword Expressions in the Prague Dependency Treebank. 2007. (manuscript)