The Lexico-Semantic Annotation of PDT: Some Results, Problems and Solutions

Eduard Bejček, Petra Möllerová, and Pavel Straňák

Institute of Formal and Applied Linguistics,
Charles University, Prague, Czech Republic
{bejcek, mollerova, stranak}@ufal.mff.cuni.cz

Abstract. This paper presents our experience with the lexico-semantic annotation of the Prague Dependency Treebank (PDT). We have used the Czech WordNet (CWN) as an annotation lexicon (repository of lexical meanings) and we annotate each word which is included in the CWN. Based on the error analysis we have performed some experiments with modification of the annotation lexicon (CWN) and consequent re-annotation of occurrences of selected lemmas. We present the results of the annotations and improvements achieved by our corrections.

1 Introduction

In the Prague Dependency Treebank (PDT; see [1,2]), the annotation can be viewed as an iterative analysis of text in the following sequence: raw text — tokenized text — morphologically analysed and lemmatised text — surface syntax (analytical layer) — deep syntax including verb valencies, topic-focus articulation and other features (tectogrammatical layer). It is not just enrichment of the original text with additional information or rather explication of grammatical information contained in the text. The tectogrammatical layer includes all the information needed to generate the surface structure without using the original text or the other layers. For technical reasons lexico-semantic annotations have been added to the morphological level.

Lexico-semantic annotation (if the process is manual, done by humans) or *tagging* (if it is automatic, performed by a machine) means assigning a semantic tag from an a priori given set to *each* relevant lexical unit in a text. Lexical units which we deal with during this process are lemmas of words;[1] the relevant ones are those of the autosemantic parts of speech, namely all nouns, adjectives, verbs, and adverbs.

In this paper, symbol $\mathrm{T}_p(l)$ denotes a set of *possible semantic tags* which can be assigned to lemma l and $\chi \subset \mathrm{T}_p(l)$ is a set assigned.

The purpose of lexico-semantic annotation or tagging is to distinguish between different meanings a semantically ambiguous lemma has in different contexts.

[1] The lemmas at the syntactical level of the PDT form a set of *tectogrammatical lemmas*, which is different from the set of lemmas at the morphological level [3]. However (despite lexico-semantic analysis being placed only after the syntactical level), we currently use the lemmas produced by morphological analyser for various practical or technological reasons.

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 21–28, 2006.

2 The Project Goal

The original goal was to provide training data for word sense disambiguation. For this purpose two annotators begun to annotate in parallel[2]the files of PDT 1.0 [2] with synsets from Czech WordNet. Later it was decided to aim for complete lexico-semantic annotation of PDT, thus enriching it with word sense information. To this day each annotator has processed 34,231 sentences consisting of 431,447 words. Because not all words occurring in the text exist in CWN, 148,774 instances (i.e. words) of 7972 lemmas were actually annotated. Since the morphological layer of PDT 2.0 consists of approximately 2 million words, we have annotated over 20% of PDT 2.0.

3 Annotation Using the Czech WordNet

The CWN consists of 28,392 synsets (including nouns, adjectives, verbs, and adverbs) [4]. We use it to obtain the set of possible semantic tags $T_p(l)$ for each relevant lemma. In the process of annotation, each annotated lemma is assigned the best tag from this set.

Table 1. List of the exceptions ordered by their preference

Incorrect Reflexivity
Missing Positive Sense
Missing Negative Sense
Incorrect Lemma
Figurative Use
Proper Name
Unclear Word Meaning in the Text
Unclear CWN Sense
Missing More General Sense
Missing Sense
Other Problem

The annotators must always assign exactly one synset or exception[3] to each relevant word (i.e. $|\chi| = 1$). In case annotators believe that one synset cannot be assigned, they can either mark the occurrence as vague (Exception 8) or they can say they are missing more general sense in the CWN (Exception 10). These two exceptions are equivalent to situations when we know that $|\chi| = 1$, but we cannot identify χ, and when we know that $|\chi| > 1$, respectively. These exceptions were assigned very rarely and agreed on in 20 and 0 cases, respectively. From this we can conclude that allowing $|\chi| > 1$ would give us nothing and could only hurt interannotator agreement, because $T'_p(l)_{|\chi|>1} = \mathcal{P}(T_p(l)_{|\chi|=1})$.

Annotators are instructed to try to assign a uniliteral synset first. Only if no uniliteral synset is usable, they examine the multiliteral synsets (if present). If and only if no synset

[2] We have used double blind annotation to be able to create a gold standard data by imploying a corrector (=third annotator).

[3] In contrast to SemCor [5] and other similar projects (see Section 8).

from $T_p(l)$ can be assigned, the annotators choose one of the exceptions given in Table 1 (for details see [6]).

4 Annotation Statistics

4.1 Summary of the Data Distribution

In terms of lexical semantics, only *autosemantic* words (nouns, adjectives, verbs, and adverbs)[4] can be the subject of semantic tagging. There were 70% such words in the annotated text. However, only words present in the CWN were annotated because they have at least one possible tag to be assigned. 35% of all words fullfiled this condition but only 24% were ambiguous (i.e. had more than one possible tag). This implies that only about 1/2 of all autosemantic words in a given text can be subject of automatic word sense disambiguation and only 1/3 are really ambiguous (according to the CWN). Detailed counts are given in the Table 2.

69% of annotated words were nouns, 21% were verbs, and 10% were adjectives. Since the CWN version we worked with does not contain any adverbial synsets, no adverbs were annotated.

Only 67% of nouns, 26% of adjectives, and 49% of verbs occur at least in one synset and thus could be processed by annotators. Now let us see how difficult this work was.

As described in section 3, there are three types of semantic tags used for annotation: uniliteral synsets, multiliteral synsets, and exceptions. A typical annotated word had 3 possible uniliteral and 6 multiliteral synsets in the set of possible tags $T_p(l)$. Considering only those words with more than one possible tag, they have 3.8 uniliteral synsets and 8 multiliteral ones. Multiliteral synsets appeared almost exclusively in the tag sets of nouns.

4.2 Inter-annotator Agreement

All kinds of linguistic annotation are usually performed by more than one annotator. The reason is to obtain more reliable and consistent data. In order to learn this reliability we can measure inter-annotator agreement, a relative number of cases when selections of the annotators were identical. This number gives also evidence of how difficult the annotation is. Manually annotated data is often used to train systems for automatic assigning relevant tags (tagging). Inter-annotator agreement gives an upper bound of accuracy of such systems.

Table 3 shows the inter-annotator agreement measured from various points of view. Basic agreement on selection of uniliteral synsets was 61.5%. If we consider both uniliteral and multiliteral synsets the inter-annotator agreement increases only by 0.2%. Overall inter-annotator agreement on all possible types of tags is 74.6% – 1/4 of all processed words are not annotated reliably. This number varies depending on POS: verbs were significantly more difficult to assign a correct uniliteral synset.

Generally speaking, the inter-annotator agreement is relatively low but it does not necessarily imply that annotators had problems to distinguish word meanings. They rather had problems to select the most suitable options that would correspond to their opinion.

[4] Numerals are sometimes considered autosemantic words too, but usually they are not the subject of semantic annotation.

According to the CWN, some words occurring in the annotated texts had up to 18 senses. Surprisingly, the inter-annotator agreement does not depend on the degree of ambiguity. It ranged from 15% to 80% regardless of the number of possible tags. We can conclude that the size of word tag sets is probably not what causes the low inter-annotator agreement.

5 Discussion on Semantic Tags and the Inter-annotator Agreement

There are two basic situations when the annotators can hardly generate the desired results, i.e. choose both the same synset: a) if they for some reason do not understand the meaning of the word to be annotated in the text, or b) if they understand the text and the word meaning, but they are unable to choose the desired meaning from proposed $T_p(l)$.

If we wanted to tackle the first source of non-agreement, we could allow the annotators to choose more than one synset, to address the vagueness of meaning. Such a change would however result in much bigger $T_p(l)$. Our experience shows us that if the choices are too many, the annotators make more mistakes and the work is slow and therefore expensive. Because $T_p(l)$ would be enlarged for every word, but the vague contexts are very rare, we have decided against this option.

6 Corrections

When we have analysed the inter-annotator agreement and the exception annotations, we have found that significant number of non-agreements is caused by several highly frequent lemmas that are not treated well in the CWN (see 6).

The inter-annotator agreement on a synset (i.e. both annotators assigned the word the same synset) was 61.5%. In 25.7% at least one annotator assigned an exception and in the remaining 12.8% both annotators assigned a synset but they disagreed. This gives us 38.5% of non-agreement.

The non-agreement here means anything but the agreement of both annotators on assigning the same synset to a word. We have split the non-agreements into 2 classes:

a) At least one annotator assigned a synset (i.e. disagreement on synset or synset / exception disagreement)
b) Both annotators assigned an exception (i.e. agreement or disagreement on exception)

The analysis of the synsets for words with frequent non-agreement showed that the annotators had

Table 2. Word counts in annotated text

All words	431 447	100.0%	
Autosemantic words	300 725	69.7%	100.0%
Annotated words	148 744	34.5%	49.5%
Ambiguous words	101 703	23.6%	33.8%

Table 3. Inter-annotator agreement (in %) on selection of the same: uniliteral synset (U); uniliteral or multiliteral synset (UM); uniliteral or multiliteral synset or exception (UME)

POS	U	UM	UME
N	65.6	66.0	74.1
V	44.8	44.8	75.4
A	67.0	67.0	76.1
All	61.5	61.7	74.6

1. Either little or no information for choosing correct synset from a range of choices (synsets were missing definitions and examples), so they basically had to choose randomly, or
2. The correct meaning of the word was not in CWN (missing synset).

We have decided to try and correct the non-agreement cases in two rounds:

1. Have a corrector (3rd annotator) look at the choices of both annotators in cases of a) and try to decide whether one of them is right. In cases where 2 of 3 agree we would consider the word successfully annotated.
2. In cases of b) and in cases from the first round where the corrector can't find a reason to agree with 1st or 2nd annotator we would:
   - Check all the meanings (synsets) of a word in CWN, merge, divide, add or clarify the synsets as needed to give annotators a clear guideline for decision.
   - Re-annotate the occurrences of the corrected lemma.

For our experiment we have taken the lemmas with the frequency of non-agreement $\geq$ 200. This resulted in 25 lemmas as given in Table 4 and marked by circles in Figure 1.

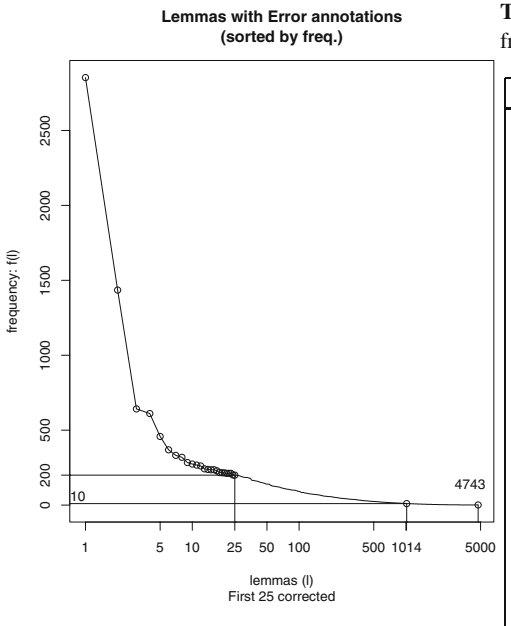

Fig. 1. Corrected lemmas

Table 4. Lemmas with highest non-agreement frequency

Lemma	Agr.	Non-agr.	Total
čas (time)	41	242	283
část (part)	23	238	261
cena (price)	85	642	727
člověk (human)	326	267	593
dát (to give)	17	332	349
den (day)	159	236	395
dobrý (good)	252	217	469
dostat (to get)	20	261	281
fax (fax)	49	212	261
místo (place)	136	274	476
mít (to have)	0	2853	2853
moci (to can)	9	1435	1444
návrh (offer)	49	229	278
podnik (business)	20	458	478
práce (work)	193	215	408
právo (law)	29	201	230
řada (row)	16	216	232
říkat (say)	1	212	213
rok (year)	1629	611	2240
stát (to stand)	152	369	521
stát (state)	136	318	454
svět (world)	59	236	295
systém (system)	32	212	244
uvést (to state)	38	284	322
vysoký (high)	282	200	482

First Round of Corrections. For each lemma we have taken all cases of non-agreement where at least one annotator chose a synset and extracted all the occurrences from the original files. Each occurrence had a context of at least 20 words on each side.

These lists of snippets were further divided to group similar cases in order to simplify the work of the corrector (3rd annotator). The division was as shown in Table 5 according to choice of annotators.

Most of the possible list were empty, only 6 in average existed for each lemma. Each of the resulting lists was added a choice of annotator A and B respectively to each lemma occurrence.

The corrector then had at most three options:

1. agree with A (if he chose a synset)
2. agree with B (if he chose a synset)
3. don't agree with either A or B

First two options meant the word was considered successfully annotated, the third one added this occurrence to those already prepared for the second round.

The corrector was also able to add notes to the word in general or to each occurrence separately for use in the second round.

Although the corrector agreed with A or B sometimes, each lemma of our chosen 25 was in the end sent into the second round. This meant that the CWN synsets will be edited and all the occurrences will have to be re-annotated or at least checked again. Nevertheless crucial data for editing CWN ware gathered.

Second Round of Corrections. First the notes from the 1st round were gathered and compared to the CWN we have been using. We have also checked the most recent version of CWN in order to see if the problems have been resolved. Various Czech printed dictionaries as well as the Princeton WordNet were also consulted. The new sense distinctions were kept as simple as possible. We have identified the basic synset and distinguished a different meaning (created a synset) only if we were able to precisely specify a difference. This in many cases resulted in merging existing CWN synsets. At the same time new synsets for missing senses were sometimes added. For these cases the annotations with the exception number 10 (missing sense) proved valuable. Each synset was also enriched by the sort definition and the example sentence (usually from our data). For editing the synsets we have used the wordnet browser and editor VisDic [4].

7 Results of Corrections

After the CWN was modified we have started the re-annotation of the data with the new synsets. Although this part has not yet been finished, we can calculate the result. When all the occurrences of our 25 lemmas will be successfully annotated, the improvements of annotated data will be as shown in Table 6.

We have corrected 25 of 4,738 lemmas for which there are cases of non-agreement. We gave gained 10,971 new words annotated with the synsets. This means that by correcting 0.5% of problematic lemmas we have gained 7.4% improvement with respect to annotated words.

Table 5. Classes of files according to annotators' choices (A and B are original annotators)

A	B
uniliteral-x	uniliteral-y
uniliteral	multiliteral
multiliteral	uniliteral
multiliteral-x	multiliteral-y
exception	uniliteral
exception	multiliteral
uniliteral	exception
multiliteral	exception

Table 6. Annotation with uniliteral synset (U) (in %)

POS	U
N	70.3 (+4.7)
V	63.6 (+18.8)
A	69.7 (+2.7)
All	68.9 (+7.4)

It is also interesting to look at the sentences that are fully disambiguated with respect to our CWN. This means that in such sentence all the annotated words are annotated correctly: before our corrections there were 5,111 sentences fully disambiguated, after the corrections it is 6,941 sentences. This means 35.8% improvement with respect to data that can be used for "all words" word sense disambiguation.

Corrections took aproximatelly 320 hours to the corrector who decided on the changes to CWN and annotated the data and 215 hours to the programmer who created the annotation data sets and scripts, implemented changes to CWN and processed the data as needed.

8 Summary

To our best knowledge, there are three similar projects: English SemCor [5], cf. also [7], Spanish Cast3LB [8] and recent Basque corpus annotation [9]. All of these efforts are smaller[5] and they differ in important methodological aspects; most prominently, both Spanish and Basque projects use transversal annotation (word-type by word-type) and they (as well as SemCor) allow arbitrary subset of $T_p(l)$ (i.e. $\chi : |\chi| > 1$) to be assigned as the final tag. We have implored linear process, because, as we have explained earlier in Section 2, we wanted to obtain training data for all words WSD. As for allowing $|\chi| > 1$, we put forward our reasons against it in Section 3.

Our semantic annotation of the PDT has two major applications:

1. Lexico-semantic tags are a new kind of labels in the PDT and will become a substantial part of a complete resource of training data, which can be exploited in many fields of NLP.

   We have shown above that the recent corrections improved significantly the number of sentences that are fully lexico-semantically annotated with respect to our current annotation lexicon.
2. The process of annotation provides a substantial feedback to the authors of the CWN and significantly helps to validate and improve its quality. In process of the corrections we have also begun improving CWN on our own.

[5] Basque: cca 300,000, Cast3LB: 125,000 vs. PDT: cca 2,000,000 tokens.

Acknowledgments

This work has been supported by grant 1ET201120505 of Grant Agency of the Czech Republic, and project MSM0021620838 of the Ministry of Education.

References

1. Hajič, J., Vidová-Hladká, B., Hajičová, E., Sgall, P., Pajas, P., Řezníčková, V., Holub, M.: The current status of the prague dependency treebank. In Matoušek, V., Mautner, P., Mouček, R., Taušer, K., eds.: TSD2001 Proceedings, LNAI 2166, Berlin Heidelberg New York, Springer-Verlag (2001) pp. 11–20.
2. Hajič, J., Hajičová, E., Pajas, P., Panevová, J., Sgall, P., Vidová-Hladká, B.: Prague dependency treebank 1.0 (Final Production Label) (2001) Published by Linguistic Data Consortium, University of Pennsylvania.
3. Hajič, J., Honetschläger, V.: Annotation lexicons: Using the valency lexicon for tectogrammatical annotation. Prague Bulletin of Mathematical Linguistics (2003) 61–86.
4. Smrž, P.: Quality Control for Wordnet Development. In Sojka, P., Pala, K., Smrž, P., Fellbaum, C., Vossen, P., eds.: Proceedings of the Second International WordNet Conference—GWC 2004, Brno, Czech Republic, Masaryk University (2003) 206–212.
5. Landes, S., Leacock, C., Tengi, R.I.: Building semantic concordances. In Fellbaum, C., ed.: WordNet, An Electronic Lexical Database. 1st edn. MIT Press, Cambridge (1998) 199–216.
6. 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).
7. Stevenson, M.: Word Sense Disambiguation: The Case for Combinations of Knowledge Sources. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, California (2003).
8. Navarro, B., Civit, M., Martí, M.A., Marcos, R., Fernández, B.: Syntactic, semantic and pragmatic annotation in cast3lb. Technical report, UCREL, Lancaster, UK (2003).
9. Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Pociello, E., Iruskieta, M.Q.: Improving the basque wordnet by corpus annotation. In: Proceedings of Third International WordNet Conference, Jeju Island (Korea) (2006) 287–290.