Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank

Jan Hajič,[†] Martin Holub,[*] Marie Hučínová,[*] Martin Pavlík,[*]
Pavel Pecina,[*] Pavel Straňák,[*] Pavel Martin Šidák[*]

[†]Institute of Formal and Applied Linguistics, [*]Center for Computational Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25, Prague, Czech Republic
{hajic, holub, hucinova, pavlik, pecina, p.stranak, sidak}@ufal.mff.cuni.cz

Abstract

We give a brief report on our experience with lexico-semantic annotation of a Czech linguistic corpus. We use the Czech WordNet (CWN) as a repository of lexical meanings and we annotate each word which is included in the CWN. The statistics of the annotated data is used as a feedback for validating and improving the coverage and quality of the CWN. We also discuss some methodological questions.

1. Introduction

Generally, the annotation of linguistic corpora usually consists of a sequence of processes corresponding to several levels of annotation. In the Prague Dependency Treebank (PDT; see (Hajič et al., 2001b), (Hajič et al., 2001a)), the annotation can be viewed as a gradual enrichment of text by several types of labels in the following sequence: raw text — tokenized text — morphologically analyzed and lemmatized text — syntactically annotated text — lexico-semantically annotated text.

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. Note that the members of $\mathrm{T}_p(l)$ always make a list of options from which a human annotator selects a correct tag for the lemma l in a given context.

The purpose of lexico-semantic annotation or tagging is to distinguish between different meanings of semantically ambiguous lemmas that can emerge when a lemma is used in different contexts. Undoubtedly, the lexico-semantic information given by correctly assigned semantic tags may be very important for many NLP tasks.

This paper concentrates on our practical experience with lexico-semantic annotation and empirical observations rather than on theoretical questions. At the very beginning, to start the lexico-semantic annotation of the PDT, we had to make two crucial decisions:

1. What system of semantic tags should we use for the lexico-semantic annotation?

   One possibility is to use a well known type of lexical database called WordNet (Fellbaum, 1998). Then, the basic semantic elements are *synsets*, sets of synonyms. As we annotate Czech texts, we decided to employ the Czech WordNet (CWN, (Smrž, 2003)) as a semantico-lexical basis for the annotation even though this choice is not a matter of course.

2. Moreover, it is also problematic how to employ the system of the synsets. In other words, how should we establish $\mathrm{T}_p(l)$ for each relevant lemma using the WordNet?

   For $\mathrm{T}_p(l)$ one can simply take the set of synsets which contains exactly the given lemma, while more complicated solutions permit even various sets of synsets to serve as semantic tags.

   Our current approach described in section 2. is very close to the first option, yet in section 6. we also discuss the latter one as in our opinion it is a way how to eliminate or at least reduce the undesirable impact of high granularity of the WordNet.

The rest of the paper is organized as follows: in section 2. we describe the process of manual annotation, our annotation tool, and how we deal with the CWN. Section 3. first introduces some information about texts we have annotated, and then the statistics of the performed annotation. Two applications are shown in sections 4. and 5. We validate the famous Yarowsky's hypothesis "one sense per collocation" and use the annotated data for validating and improving the CWN. Finally, we discuss the relation between the granularity of semantic tags and the inter-annotator agreement. Section 7. briefly summarizes the main contributions.

[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 (Hajič and Honetschläger, 2003). However (despite lexico-semantic analysis being placed only after the syntactical level), we currently use the lemmas produced by morphological analyzer for various practical or technological reasons.

Incorrect Reflexivity	l is reflexive but CWN knows only its non-reflexive form or vice versa.
Missing Positive Sense	l is positive, but CWN includes only its negative form.
Missing Negative Sense	l is negative, but CWN includes only its positive form.
Incorrect Lemma	The lemma l assigned to the word is incorrect (therefore the synsets proposed are incorrect too).
Figurative Use	The word is used in a metaphorical or other figurative way.
Proper Name	Assigned to proper names not included in the CWN.
Unclear Word Meaning in the Text	The meaning of l is unclear (therefore no synset can be assigned).
Unclear CWN Sense	The meaning of a synset is unclear and no other proposed synset can be used.
Missing More General Sense	At least one of the proposed synsets corresponds to the meaning of l, but is too specific and so expressing only part of l.
Missing Sense	None of the synsets proposed expresses the meaning of l and more specific exceptions can not be used.
Other Problem	Assigned if no other category can be used.

Table 1: List of the exceptions ordered by their preference.

2. Annotation using the Czech WordNet

The CWN was originally developed as a part of the EuroWordNet project (see (EuroWordNet, 2004), (Vossen, 1998)). Since then it was extended and is still being developed as a part of the BalkaNet Project (BalkaNet, 2004); currently, it consists of 28,392 synsets (including nouns, adjectives, verbs, and adverbs) (Smrž, 2003).

We use the CWN to obtain the set of possible semantic tags $\mathrm{T}_p(l)$ for each relevant lemma. In the process of annotation, each annotated lemma is assigned the best tag from this set.

2.1. Semantic tags based on the CWN synsets

In this paper, basic lexical units of the CWN (i.e. elements of synsets) are called *literals*. Literals which consist of exactly one lemma are called *uniliterals*, the other are called *multiliterals*.

Given a lemma l, the members of the set $\mathrm{T}_p(l)$ are

1. all synsets with a uniliteral consisting of l, and
2. some synsets with multiliterals (especially with those containing l) selected by a special procedure based on the CWN hypernymy/hyponymy relation (Pavlík, 2002).

2.2. Annotation environment

We use a graphical annotation tool.[2] The input file is a morphologically annotated text from the PDT with the corresponding $\mathrm{T}_p(l)$ sets encoded. The window of the application is split into four parts (see Fig. 1). When the annotator loads the input file, the text is displayed in the area marked A. In column B the annotator can see the list of lemmas of the words to be annotated. When the annotator chooses a lemma in column B, it is highlighted in the area A and he can see a list of possible tags $\mathrm{T}_p(l)$ in area C. To decide which synset from the offered list best represents the meaning of the word, the annotator can browse the synsets displayed in area C and review their English glosses (if present in the CWN), their hypernym synsets and the glosses of these hypernyms in area D. This way the annotator can see at the same time the annotated word in its full context and all the necessary information about its $\mathrm{T}_p(l)$ to select the best tag.

2.3. Instructions for annotators

The annotators must always assign exactly one synset or exception[3] to each relevant word and they 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 from $\mathrm{T}_p(l)$ can be assigned, the annotators choose one of the exceptions given in Table 1. First eight exceptions should be chosen preferably. Only if none of them is used, exception ‘Missing Sense’ can be assigned. Only if neither of the mentioned options is applicable, the annotator assigns the last exception ‘Other’.

3. Annotation statistics

The long-term goal of our project is the complete annotation of the PDT 1.0 (Hajič et al., 2001a). After one year of annotation we have processed 11,014 sentences containing 125,129 words, mostly from the domain of economics. This is about 15 % of the PDT.

The entire annotation was performed independently by two human subjects (postgradual students with linguistic education) having identical instructions described in section 2. The average time needed for processing a typical document containing about 50 sentences by one annotator was 1 hour. Such a document contains approximately 100 to 280 words to be annotated.

Now we present a summary of the annotated data and some statistics.

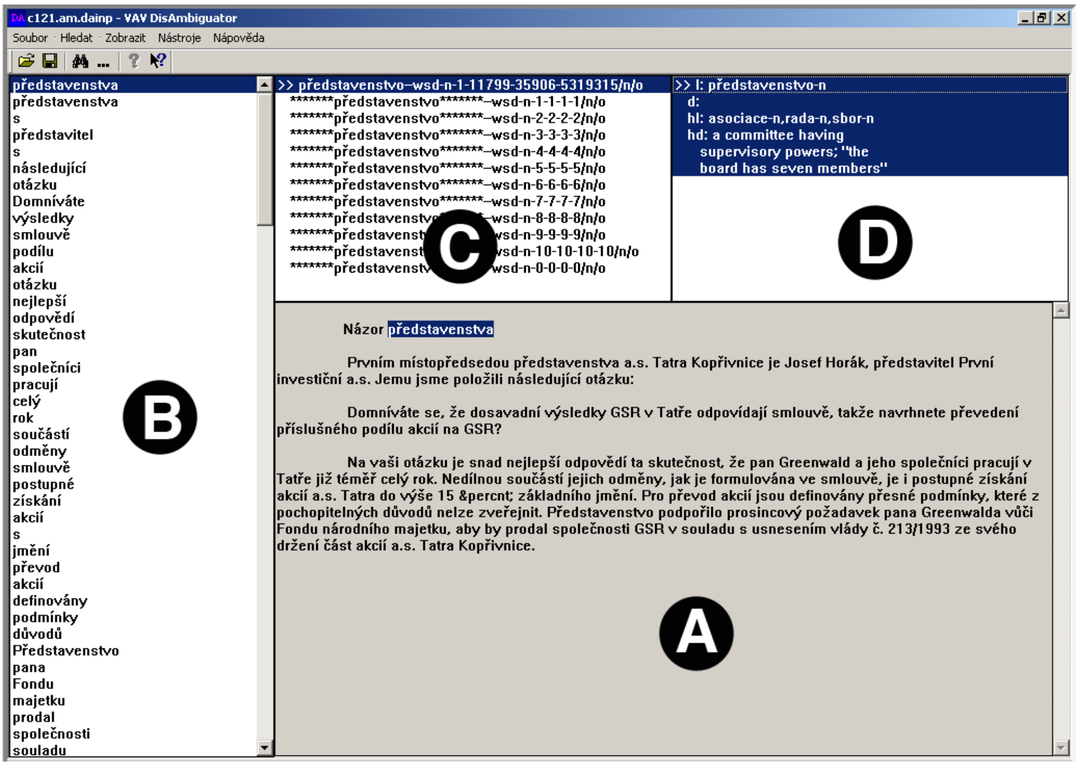
3.1. Summary of the data distribution

In terms of lexical semantics, only *autosemantic* words (nouns, adjectives, verbs, and adverbs)[4] can be the subject of semantic tagging. There were 69 % such words in the annotated text. However, only words present in the CWN

[2] The program called DA was designed and implemented by Jiří Hana.

[3] In contrast to SemCor (Landes et al., 1998).

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



```
představenstva
představenstva
s
představitel
s
následující
otázku
Domníváte
výsledky
smlouvě
podílu
akcií
otázku
nejlepší
odpovědí
skutečnost
pan
společníci
pracují
celý
rok
součástí
odměny
smlouvě
postupné
získání
akcií
s
jmění
převod
akcií
definovány
podmínky
důvodů
Představenstvo
pana
Fondu
majetku
prodal
společnosti
souladu
```

```
>> představenstvo–wsd-n-1-11799-35906-5319315/n/o
   *******představenstvo*******–wsd-n-1-1-1-1/n/o
   *******představenstvo*******–wsd-n-2-2-2-2/n/o
   *******představenstvo*******–wsd-n-3-3-3-3/n/o
   *******představenstvo*******–wsd-n-4-4-4-4/n/o
   *******představenstvo*******–wsd-n-5-5-5-5/n/o
   *******představenstvo*******–wsd-n-6-6-6-6/n/o
   *******představenstvo*******–wsd-n-7-7-7-7/n/o
   *******představenstv        –wsd-n-8-8-8-8/n/o
   *******představenst          wsd-n-9-9-9-9/n/o
   *******představenst          wsd-n-10-10-10-10/n/o
   *******představenstv         wsd-n-0-0-0-0/n/o
```

```
>> l: představenstvo-n
   d:
   hl: asociace-n,rada-n,sbor-n
   hd: a committee having
     supervisory powers; "the
     board has seven members"
```

Názor představenstva

Prvním místopředsedou představenstva a.s. Tatra Kopřivnice je Josef Horák, představitel První investiční a.s. Jemu jsme položili následující otázku:

Domníváte se, že dosavadní výsledky GSR v Tatře odpovídají smlouvě, takže navrhnete převedení příslušného podílu akcií na GSR?

Na vaši otázku je snad nejlepší odpovědí ta skutečnost, že pan Greenwald a jeho společníci pracují v Tatře již téměř celý rok. Nedílnou součástí jejich odměny, jak je formulována ve smlouvě, je i postupné získání akcií a.s. Tatra do výše 15 &percnt; základního jmění. Pro převod akcií jsou definovány přesné podmínky, které z pochopitelných důvodů nelze zveřejnit. Představenstvo podpořilo prosincový požadavek pana Greenwalda vůči Fondu národního majetku, aby by prodal společnosti GSR v souladu s usnesením vlády č. 213/1993 ze svého držení část akcií a.s. Tatra Kopřivnice.

Figure 1: A screenshot of the annotation tool DA.

were annotated because they have at least one possible tag to be assigned. 34 % 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 following table.

All words	125 129	100.0 %	
Autosemantic words	85 965	68.7 %	100.0 %
Annotated words	42 900	34.3 %	49.9 %
Ambiguous words	30 091	24.0 %	35.0 %

Table 2: Word counts in annotated text.

70 % of annotated words were nouns, 20 % were verbs, and 10 % were adjectives. Since the CWN version we worked with does not contain any adverbial synsets, no adverbs were annotated.

Detailed summary of part-of-speech (POS) distribution is given in Table 3. The relative counts are with respect to counts of autosemantic words. These numbers refer to "coverage" of annotated texts with words from the CWN. Generally, the coverage is poor, but varies strongly depending on POS.

POS	Autosemantic		Annotated		Ambiguous	
N	43 315	100 %	30 184	70 %	22 294	51 %
A	16 519	100 %	4 272	26 %	3 107	19 %
V	18 421	100 %	8 444	46 %	4 690	25 %
D	7 710	100 %	0	0 %	0	0 %

Table 3: Absolute and relative word counts per POS.

Only 70 % of nouns, 26 % of adjectives, and 46 % 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 2., there are three types of semantic tags used for annotation: uniliteral synsets, multiliteral synsets, and exceptions. The average numbers of tags of different types which could be selected for one word are in Table 4. A typical annotated word had 3 possible uniliteral and 7 multiliteral synsets in the set of possible tags $\mathrm{T}_p(l)$. Considering only those words with more than one possible tag, they have 3.8 uniliteral synsets and 9 multiliteral ones. Multiliteral synsets appeared almost exclusively in the tag sets of nouns.

POS	Annotated words			Ambiguous words		
	U	M	E	U	M	E
N	2.8	9.8	11	3.5	12.1	11
A	3.0	0.1	11	4.7	0.1	11
V	3.8	0.0	11	4.9	0.0	11
All	2.9	6.9	11	3.81	9.0	11

Table 4: Average numbers of possible tags of all types for annotated and ambiguous words with respect to their POS, and in total. (U stands for uniliterals, M for multiliterals, and E for exceptions.)

Although multiliteral synsets appeared in sets $\mathrm{T}_p(l)$ very often, annotators used them rather rarely (0.6 % of words), which is in accordance with their instructions (see section 2.3.). Uniliteral synsets were assigned to 82 % of all annotated words. 17.4 % of words were tagged by an exception. See details for relevant POS in Table 5.

POS	U	M	E
N	85.8	1.2	13.0
V	62.9	0.0	37.1
A	90.9	0.0	9.1
All	82.0	0.6	17.4

Table 5: Average usage (in %) of uniliteral synsets (U), multiliteral synset (M), and exceptions (E) per POS and in total.

3.2. 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.

POS	U	UM	UME
N	64.7	65.1	70.9
V	44.5	44.5	63.8
A	71.0	71.0	74.6
All	61.4	61.6	69.9

Table 6: Inter-annotator agreement (in %) on selection of the same: uniliteral synset (U); uniliteral or multiliteral synset (UM); uniliteral or multiliteral synset or exception (UME).

Table 6 shows the inter-annotator agreement measured from various points of view. Basic agreement on selection of uniliteral synsets was 61.4 %. 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 69.9 % – almost 1/3 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.

According to the CWN, some words occurring in the annotated texts had up to 18 senses (see Table 7). 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.

Ambiguity	Words	Agreement (%)
1	12809	79
2	11154	75
3	7071	70
4	5466	54
5	2270	56
6	1034	51
7	819	39
8	547	53
9	329	63
10	162	72
11	612	80
12	69	52
13	68	38
14	90	41
15	13	15
16	369	60
17	18	0
18	72	50

Table 7: Overall inter-annotator agreement in relation to degree of word sense ambiguity in the CWN.

3.3. Sense Distribution

In Table 4 we show the average word sense ambiguity in our text according to the CWN. Although this number is relatively high (3 uniliteral plus 9 multiliteral synsets), the real average sense ambiguity of words according annotators is only 1.47. Put differently, all annotated words were assigned only 1.47 different tags in average.

Omitting the cases of disagreement, 62.4 % of all annotated words were always assigned only one synset.

Some more details are given in Table 8.

Amb	N	V	A	Total
1	61.2	56.4	73.2	62.4
2	28.7	28.4	19.5	27.3
3	7.9	10.7	0.7	7.2
4	0.7	4.1	2.6	1.4
5	1.0	0.3	4.0	1.4
6	0.5	0.0	0.0	0.3

Table 8: Word sense distributions in relation to degree of ambiguity.

4. Related experiments

Manual semantic annotation (and also other types of manual annotation) is a time-consuming and therefore expensive process. One way to make this work easier is to use a user-friendly application providing a comfortable environment for annotator's decision making and tag assignment.

Another (but disputable) method is to preprocess unannotated text and automatically tag unambiguous phenomena or prepare the most likely tags for each word occurrence. This approach has two problematic aspects: usually, automatic annotation is not perfect and annotator should review computer's results; but then the annotator can excessively incline to computer's preferred selections.

An example of the latter method is an application of Yarowsky's hypothesis "One sense per collocation" (Yarowsky, 1995) saying that all occurrences of a word in the same collocation have the same meaning. Thus, annotators could process only the first occurrence of each collocation and then this choice would be automatically assigned to all the other occurrences of this collocation.

We obtained a list of significant collocations occurring in the PDT more than 5 times (for the method see (Pecina and Holub, 2002)) and extracted those collocations that appear in our semantically annotated text. There were 3,741 such collocations, 964 unique.

First we have separately validated this hypothesis on the texts annotated by each annotator, and then only on words that were assigned the same tag by both annotators.

Semantic annotation	a)	b)
Annotator A	86.22	77.25
Annotator B	86.42	71.03
Annotator A+B agreement	97.88	96.24

Table 9: Validity (in %) of Yarowsky's hypothesis "One sense per collocation" for words in collocation occuring a) at least once and b) at least twice in the annotated text.

Results of this experiment can be found in the column a) of Table 9. Considering only the reliable annotation from both annotators, the hypothesis is valid for 97.88 % of words and this fully corresponds to Yarowsky's observation on English.

We obtained worse results on all annotated words – taking separately from both annotators – only about 86 %, which however coresponds to the low inter-annotator agreement. The annotators had difficulties to select appropriate tags, consequently they sometimes annotated words with the same meaning with different synsets (low consistency of annotation).

Results in column b) of Table 9 are from experiments using words occurring in the text more than once. They are unsurprisingly lower.

5. Validating and improving the Czech WordNet

Based on our experience with semantic annotation we point out some issues concerning the coverage and quality of the CWN:

- Less than 50 % of nouns, adjectives and verbs in annotated texts occur in the CWN.
- Only 30 % of all nouns, adjectives and verbs were succesfully annotated with a CWN synset.
- Some of very common meanings of frequent words are not covered by the CWN.
- Only 12 % of all CWN synsets were assigned to a word.

These facts give us evidence of (i) uneven distribution of the CWN synsets and (ii) insufficient word coverage.

One of the important outcomes of our work is valuable information which can lead to quality improvement of the CWN and that cannot be obtained in other way. We can provide the authors of the CWN with

- distribution of synset elements for individual synsets;
- distribution of synsets for individual words;
- more or less specific information about missing synsets, percentage and specification of their types (which correspond to the kinds of the exceptions, see Table 1.).

5.1. Comparing two CWN versions

The CWN version 1.2a, which we have been using, has 24,855 synsets, whereas the newly developed version 1.8d has 28,392 synsets. 3537 synsets were added in total, but more importantly many synsets were verified and changed, some wrong synsets were deleted and new once added, some of them based on our feedback.

Valency frames were also added to many verb synsets, which should simplify annotator's decisions and improve consistency of annotations. Most importantly, CWN 1.2a did not include any adverbial synsets. Consequently none of the 7710 adverbs in our texts has been annotated. The version we have been using does not include Czech glosses and not all synsets have an English gloss. Some English glosses also do not fit the Czech synsets. In contrast, CWN 1.8d includes many Czech glosses that fit the synsets and also includes example sentences.

We expect that using the new CWN version will lead to an improvement of the inter-annotator agreement by eliminating some sources of common errors. However, the high granularity of the WordNet senses, which also often causes inter-annotator disagreement, is a problem sui generis.

6. Discussion on semantic tags and the inter-annotator agreement

We have mentioned two main issues related to our work: insufficient quality of the CWN and poor inter-annotator agreement. The latter one can be tackled by changing our annotation methodology.

As mentioned in the introduction, one of the fundamental questions is what system of semantic tags (i.e. $\bigcup \mathrm{T}_p(l_i)$) should be used for the lexico-semantic annotation. This is closely related to the problem of granularity.

High granularity of the WordNet senses, i.e. the fact that words in the WordNet often have too many senses with only fine distinctions, is probably the most usual argument against the WordNet.

To reduce the impact of this undesirable granularity we can allow the annotators:

(i) assign more than one proposed synset or

(ii) assign a hypernym of a proposed synset.

The option (i) would probably worsen the inter-annotator agreement on synsets and exceptions (currently 69.9 %), yet it would also reduce the number of words annotated with exceptions (24.6 %), so the impact on agreement on synset selection is unclear. The option (ii) states the question how general hypernyms we should allow to be used as semantic tags (since the more general the tag the less information provided).

7. Summary

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.

2. The process of annotation provides a substantial feedback to the authors of the CWN and significantly helps to validate and improve its quality.

   To our best knowledge, the only comparable annotated corpus that can be used for WordNet validation is English SemCor (Landes et al., 1998), cf. also (Stevenson, 2003); as for the other languages, our project seems to be unique.

8. Acknowledgments

This work has been supported by the Ministry of Education, project Center of Computational Linguistics (project LN00A063). Here we would also like to express our grateful thanks to Karel Pala, the head of the CWN development team, for his kind cooperation, helpful commentaries, and fruitful suggestions.

9. References

BalkaNet, 2004. Project website. http://www.ceid.upatras.gr/Balkanet/.

EuroWordNet, 2004. Project website. http://www.illc.uva.nl/EuroWordNet/.

Fellbaum, Christiane (ed.), 1998. *WordNet, An Electronic Lexical Database*. Cambridge: MIT Press, 1st edition.

Hajič, Jan, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká, 2001a. Prague dependency treebank 1.0 (Final Production Label). Published by Linguistic Data Consortium, University of Pennsylvania.

Hajič, Jan and Václav Honetschläger, 2003. Annotation lexicons: Using the valency lexicon for tectogrammatical annotation. *Prague Bulletin of Mathematical Linguistics*, (79–80):61–86.

Hajič, Jan, Barbora Vidová-Hladká, Eva Hajičová, Petr Sgall, Petr Pajas, Veronika Řezníčková, and Martin Holub, 2001b. The current status of the prague dependency treebank. In R. Mouček K. Taušer V. Matoušek, P. Mautner (ed.), *TSD2001 Proceedings, LNAI 2166*. Berlin Heidelberg New York: Springer-Verlag.

Landes, Shari, Claudia Leacock, and Randee I. Tengi, 1998. Building semantic concordances. In Christiane Fellbaum (ed.), *WordNet, An Electronic Lexical Database*, chapter 8. Cambridge: MIT Press, 1st edition, pages 199–216.

Pavlík, Martin, 2002. *Semantic Disambiguation of Terms in DIS (in Czech)*. Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague.

Pecina, Pavel and Martin Holub, 2002. Semantically significant collocations (in czech). Technical Report TR-2002-13, UFAL/CKL.

Smrž, Pavel, 2003. Quality Control for Wordnet Development. In Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen (eds.), *Proceedings of the Second International WordNet Conference—GWC 2004*. Brno, Czech Republic: Masaryk University.

Stevenson, Mark, 2003. *Word Sense Disambiguation: The Case for Combinations of Knowledge Sources*. CSLI Studies in Computational Linguistics. Stanford, California: CSLI Publications.

Vossen, Piek (ed.), 1998. *Introduction to EuroWordNet*. Computers and the Humanities. Kluwer Academic Publishers, 1st edition.

Yarowsky, David, 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*. Cambridge, MA.