Annotation of Multiword Expressions in the Prague Dependency Treebank

Eduard Bejček, Pavel Straňák and Pavel Schlesinger
Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
{bejcek, stranak, schlesinger}@ufal.mff.cuni.cz

Abstract

In this article we want to demonstrate that annotation of multiword expressions in the Prague Dependency Treebank is a well defined task, that it is useful as well as feasible, and that we can achieve good consistency of such annotations in terms of inter-annotator agreement. We show a way to measure agreement for this type of annotation. We also argue that some automatic pre-annotation is possible and it does not damage the results.

1 Motivation

Various projects involving lexico-semantic annotation have been ongoing for many years. Among those there are the projects of word sense annotation, usually for creating training data for word sense disambiguation. However majority of these projects have only annotated very limited number of word senses (cf. Kilgarriff (1998)). Even among those that aim towards "all words" word-sense annotation, multiword expressions (MWE) are not annotated adequately (see (Mihalcea, 1998) or (Hajič et al., 2004)), because for their successful annotation a methodology allowing identification of new MWEs during annotation is required. Existing dictionaries that include MWEs concentrate only on the most frequent ones, but we argue that there are many more MWEs that can only be identified (and added to the dictionary) by annotation.

There are various projects for identification of named entities (for an overview see (Ševčíková et al., 2007)). We explain below (mainly in Section 2) why we consider named entities to be concerned with lexical meaning. At this place we just wish to recall that these projects only select some specific parts of text and provide information only for these. They do not aim for full lexico-semantic annotation of texts.

There is also another group of projects that have to tackle the problem of lexical meaning, namely treebanking projects that aim to develop a deeper layer of annotation in adition to a surface syntactic layer. This deeper layer is generally agreed to concern lexical meaning. Therefore the units of this layer cannot be words anymore, they should be *lexias*.

Lexia is defined by Filipec and Čermák (1986) as equivalent to a "monosemic lexeme" of (Filipec, 1994) or a "lexical unit" of (Cruse, 1986): *"a pair of a single sense and a basic form (plus its derived forms) with relatively stable semantic properties"*.

We work with the Prague Dependency Treebank (PDT, see Hajič (2005)), which has in addition to the morphemic and the surface syntactic layers also the tectogrammatical layer. The latter has been construed as the layer of the (literal) meaning of the sentence and thus should be composed of lexias (lexical units) and the relations between their occurrences.[1]

On the tectogrammatical layer only the autosemantic words form nodes in a tree (t-nodes). Synsemantic (function) words are represented by various attributes of t-nodes. Each t-node has a lemma: an attribute whose value is the node's basic lexical form. Currently t-nodes, and consequently their t-lemmas, are still visibly derived from the morphological division of text into tokens. This preliminary handling

[1] With a few exceptions, such as personal pronouns (that corefer to other lexias) or coordination heads.

has always been considered unsatisfactory in FGD.[2] There is a clear goal to distinguish t-lemmas through their senses, but this process has not been completed so far.

Our project aims at improving the current state of t-lemmas. Our goal is to assign each t-node a t-lemma that would correspond to a lexia, i.e. that would really distinguish the t-node's lexical meanings. To achieve this goal, in the first phase of the project, which we report on in this paper, we *identify multiword expressions and create a lexicon of the corresponding lexias*.

2 Introduction

We annotate all occurrences of MWEs (including named entities, see below) in PDT 2.0. When we speak of **multiword expressions** we mean "idiosyncratic interpretations that cross word boundaries" (Sag et al., 2002). We understand multiword expressions as *a type of lexias*. We distinguish also a special type of MWEs, for which we are mainly interested in its type, rather than individual lexias, during the annotation: **named entities (NE)**.[3] Treatment of NEs together with other MWEs is important, because syntactic functions are more or less arbitrary inside a NE (consider an address with phone numbers, etc.) and so is the assignment of semantic roles. That is why we need each NE to be combined into a single node, just like we do it with MWEs in general.

For the purpose of annotation we have built a repository of lexias corresponding to MWEs, which we call SemLex. We have built it using entries from some existing dictionaries and it is being enriched during the annotation in order to contain every lexia that was annotated. We explain this in detail in Section 4.1.

3 Current state of MWEs in PDT 2.0

During the annotation of valency that is a part of the tectogrammatical layer of PDT 2.0 the t-lemmas that correspond to lexias have been basically identified for all the verbs and some nouns and adjectives. The resulting valency lexicon is called PDT-VALLEX (Hajič et al., 2003) and we can see it as a repository of lexias based on verbs, adjectives and nouns in PDT that have valency. [4]

This is a starting point for having t-nodes corresponding to lexias. However in the current state it is not fully sufficient even for verbs, mainly because parts of MWEs are not joined into one node. Parts of frames marked as idiomatic are still represented by separate t-nodes in a tectogrammatical tree. Verbal phrasemes are also split into 2 nodes, where the nominal part is governed by the verb. Non-verbal idioms have not been annotated at all.

Below we give an example of the current state: an idiom meaning "in a blink (of an eye)" – literally "*what not-see" (Figure 1).

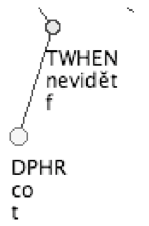

Figure 1: "Co nevidět" (in a blink)

4 Methodology

4.1 Building SemLex

Each entry we add into SemLex is considered to be a **lexia**. We have also added 9 special entries to identify NE types, so we do not need to add the expressions themselves. These types are derived from NE classification by (Ševčíková et al., 2007). Some frequent names of persons, institutions or other objects (e.g. film titles) are being added into SemLex during annotation (while keeping the information about a NE type), because this allows for their following occurrences to be pre-annotated automatically (see Section 5). For others, like addresses or bibliographic

[2] Functional Generative Description (FGD, (Sgall et al., 1986; Hajičová et al., 1998)) is a framework for systematic description of a language, that the PDT project is based upon. In FGD units of the t-layer are construed equivalently to monosemic lexemes (lexias) and are combined into dependency trees, based on syntactic valency of the lexias.

[3] NEs can in general be also single-word, but in this phase of our project we are only interested in multiword expressions, so when we say NE in this paper, we always mean multiword.

[4] It is so because in PDT-VALLEX valency is not the only criterion for distinguishing frames (=meanings). Two words with the same morphological lemma and valency frame are assigned two different frames if their meaning differs. Thus the PDT-VALLEX frames correspond to lexias.

entries, it makes but little sense, because they most probably will not reappear during the annotation.

Currently (for the first stage of lexico-semantic annotation of PDT) SemLex contains only lexias corresponding to MWEs. Its base has been composed of MWEs extracted from Czech WordNet (Smrž, 2003), Eurovoc (Eurovoc, 2007) and SČFI (Čermák et al., 1994).[5] Currently there are over 30,000 multi-word lexias in SemLex and more are being added during annotations.

The entries added by annotators must be lexias as defined above. Annotators define their "sense" informally (as much as possible) and we extract an example of usage and the basic form from the annotation automatically. The "sense" information shall be revised by a lexicographer, based on annotated occurrences.

4.2 Annotation

PDT 2.0 uses PML (Pajas and Štěpánek, 2005), which is an application of XML that utilises a stand-off annotation scheme. We have extended the PDT-PML with a new schema for so-called s-files. We use these files to store all of our annotation without altering the PDT itself. These s-files are very simple: basically each of them consists of a list of s-nodes. Each s-node corresponds to an occurrence of a MWE and it is composed of a link to the entry in SemLex and a list of identifiers of t-nodes that correspond to this s-node.

Our annotation program reads in a tectogrammatical representation (t-file) and calls TrEd (Pajas, 2007) to generate plain text. This plain text (still linked to the tectogrammatical representation) is presented to the annotator. While the annotator marks MWEs already present in SemLex or adds new MWEs into SemLex, tree representations of these MWEs extracted from underlying t-trees are added into their SemLex entries via TrEd scripts.

5 Pre-annotation

Because MWEs tend to occur repeatedly in a text, we have decided to test pre-annotation both for the speed improvement and for improving the consistency of annotations. On the assumption that *all occurrences of a MWE share the same tree structure*, while there are no restrictions on the surface word order other than those imposed by the tree structure itself we have decided to employ four types of pre-annotation:

A) External pre-annotation provided by our colleague (see Hnátková (2002)). With each MWE a set of rules is associated that limits possible forms and surface word order of parts of a MWE. This approach was devised for corpora that are not syntactically annotated.

B) Our one-time pre-annotation with those lexias from SemLex that were already used in annotation, and thus have a tree structure as a part of their entry.

C) Dynamic pre-annotation as in B, only with the SemLex entries that have been recently added by the annotator.

D) When an annotator tags an occurrence of a MWE in the text, other occurrences of this MWE in the article are identified automatically.[6]

(A) was executed once for all of the PDT. (B) is performed each time we merge lexias added by annotators into the main SemLex. We carry out this annotation in one batch for all PDT files remaining to annotate. (C) should be done for each file while it is being opened in LexemAnn GUI. (D) happens each time the annotator adds a new lexia into SemLex and uses it to annotate an occurrence in the text. In subsequent files instances of this lexia are already annotated in step (C), and later even in (B).

After the pilot annotation without pre-annotation (D) we have compared instances of the same tags and found that 10.5% of repeated lexias happened to have two different trees. After closer examination this 10.5% group is negligible because these cases are caused by ellipses, variations in lexical form such as diminutives etc., or wrong lemmatisation, rather than inconsistencies in the tree structure. These cases show us some issues of PDT 2.0, for instance:

- *jižní* × *Jižní Korea* [southern × South Korea] – wrong lemmatisation

[5] Slovník české frazeologie a idiomatiky (Dictionary of Czech Phraseology and Idiomatics)

[6] This is exactly what happens: 1) Tree structure of the selected MWE is identified via TrEd 2) The tree structure is added to the lexeme's entry in SemLex 3) All the sentences in the given file are searched for the same MWE using its tree structure (via TrEd) 4) Other occurrences returned by TrEd are tagged with this MWE's ID, but these occurrences receive an attribute "auto", which identifies them (both in the s-files and visually in the annotation tool) as annotated automatically.

- *obchodní ředitel* × *ředitelka* [managing director – man × woman] – in future these should have one t-lemma and gender should be specified by an attribute of a t-node.

We have not found any case that would show that there is such a MWE that its structure cannot be represented by a single tectogrammatical tree. 1.1% of all occurences were not connected graphs, but this happened due to errors in data and to coordination. This corroborates our assumption that (disregarding errors) all occurrences of a MWE share the same tree structure. As a result, we started storing the tree structures in the SemLex entries and employ them in pre-annotation (D). This also allows us to use pre-annotations (B) and (C), but we have decided not to use them at the moment, in order to be able to evaluate each pre-annotation step separately. Thus the following section reports on the experiments that employ pre-annotation (A) and (D).

6 Analysis of Annotations

Two annotators already started to use (and test) the tool we have developed. They both have got the same texts. The text is generated from the t-trees and presented as a plain text with pre-annotated words marked by colour labels. Annotators add their tags in the form of different colour labels and they can delete the pre-annotated tags. In this experiment data consists of approx. 120,000 tokens that correspond to 100,000 t-nodes. Both annotators have marked about 15,200 t-nodes (~15%) as parts of MWEs. annotator A has grouped them into 7,263 MWEs and annotator B into 6,888. So the average length of a MWE is 2.2 t-nodes.

The ratio of general named entities versus SemLex lexias was 52:48 for annotator A and 49:51 in case of annotator B. Annotator B used 10% more lexias than annotator A (3,279 and 3,677), while they both used almost the same number of NEs. Some comparison is in the Table 1.

type of MWE	A	B
SemLex lexias	3,677	3,279
Named Entities	3,553	3,587
- person/animal	1130	1137
- institution	842	772

Table 1: Annotated instances of significant types of MWEs

Both annotators also needed to add missing entries to the originally compiled SemLex or to edit existing entries. annotator A added 722 entries while the annotator B added 861. They modified 796 and 809 existing entries, respectively.

6.1 Inter-anntator Agreement

In this section our primary goal is to assess whether with our current methodology we produce reliable annotation of MWEs. To that end we measure the amount of inter-annotator agreement that is above chance. There are, however, a few sources of complications in measuring this agreement:

- Each tag of a MWE identifies a subtree of a tectogrammatical tree (represented on the surface by a set of marked words). This allows for partial agreement of tags at the beginning, at the end, but also in the middle of a surface interval (in a sentence).
- A disagreement of the annotators on the tag is still an agreement on the fact that this t-node is a part of a MWE and thus should be tagged. This means we have to allow for partial agreement on a tag.
- There is not any clear upper bound as to how many (and how long) MWEs are there in texts.
- There is not a clear and simple way to estimate the amount of the agreement by chance, because it must include the partial agreements mentioned above.

Since we want to keep our agreement calculation as simple as possible but we also need to take into account the problems above, we have decided to start from π as defined in (Artstein and Poesio, 2007) and to make a few adjustments to allow for types of partial agreement and estimated maximal agreement.

Because we do not know how many MWEs there are in our texts, *we need to calculate the agreement over all t-nodes*, rather than the t-nodes that "should be annotated". This also means, that the theoretical maximal agreement (upper bound) U, cannot be 1. If it was 1, it would be saying that all nodes are part of a MWE.

Since we know that $U < 1$ but we do not know it's exact value, we use the *estimated upper bound* $\widehat{U}$ (see Equation 1). Because we calculate $\widehat{U}$ over all t-nodes, we need to account not only for agreement on tagging a t-node, but also for agreement, that the t-node is not a part of a MWE, therefore it is not tagged.[7]

If N is the number of all t-nodes in our data and $n_{A \cup B}$ is the number of t-nodes annotated by at least one annotator, then we estimate $\widehat{U}$ as follows:

$$\widehat{U} = \frac{n_{A \cup B}}{N} + 0.052 \cdot \frac{N - n_{A \cup B}}{N} = 0.215 \quad (1)$$

The weight 0.052 used for scoring the t-nodes that were not annotated is explained below. Because $\widehat{U}$ includes all the disagreements of the annotators, we believe that the real upper bound U lies somewhat below it and the agreement value 0.215 is not something that should (or could) be achieved. This is however based on the assumption that the data we have not yet seen have similar ratio of MWEs as the data we have used.

To account for partial agreement we divide the t-nodes into 5 classes c and assign each class a weight w as follows:

$c1$ If the annotators agree on the exact tag from SemLex, we get maximum information: $w = 1$

$c2$ If they agree, that the t-node is a part of a NE or they agree it is a part of some lexia from SemLex, but they do not agree which NE or which lexia, we estimate we get about a half of the information compared to $c1$: $w = 0.5$

$c3$ If they agree that the t-node is a part of a MWE, but disagree whether a NE or a lexia from SemLex, it is again half the information compared to $c2$, so $w = 0.25$

$c4$ If they agree that the t-node is not a part of a MWE, $w = 0.052$. This low value of w accounts for frequency of t-nodes that are not a part of a MWE, as estimated from data: Agreement on not annotating provides the same amount of information as agreement on annotating, but we have to take into account higher frequency of t-nodes that are not annotated:

$$c4 = c3 \cdot \frac{\sum annotated}{\sum not\ annotated} = 0.25 \cdot \frac{12797}{61433} \approx 0.052$$

$c5$ If the annotators do not agree whether to annotate a t-node or not, $w = 0$.

The number of t-nodes (n) and weights w per class c are given in Table 2.

	Agreement				Disagreement
	Agreement on annotation			Not annotation	
	Agreement on NE / lexia				
	Full agreement				
class c	1	2	3	4	5
t-nodes n	10,527	2,365	389	83,287	3,988
weight w	1	0.5	0.25	0.052	0

Table 2: The agreement per class and the associated weights

Now that we have estimated the upper bound of agreement $\widehat{U}$ and the weights w for all t-nodes we can calculate our weighted version of π:

$$\pi_w = \frac{A_o - A_e}{\widehat{U} - A_e}$$

A_o is the observed agreement of annotators and A_e is the agreement expected by chance (which is similar to a baseline). π_w is thus a simple ratio of our observed agreement above chance and maximum agreement above chance.

Weights w come into account in calculation of A_o and A_e.

We calculate A_o by multiplying the number of t-nodes in each category c by that category's weight w, summing these 5 weighted sums and dividing this sum of all the observed agreement in the data by the total number of t-nodes: $A_o = \frac{1}{N} \sum_{c=1}^{5} n_c w_c = 0.160$.

A_e is the probability of agreement expected by chance over all t-nodes. This means it is the sum of the weighted probabilities of all the combinations of all the tags that can be obtained by a pair of annotators. Every possible combination of tags (including not tagging a t-node) falls into one of the categories c and thus gets the appropriate weight w. Calculating the value of A_e depends not only on values of w (see Table 2), but also on the fact that SemLex is composed of 9 entries for NE types and over 30,000 entries for individual lexias. Based on this we have obtained $A_e = 0.047$.

The resulting π_w is then

$$\pi_w = \frac{A_o - A_e}{\widehat{U} - A_e} = \frac{0.160 - 0.047}{0.215 - 0.047} = 0.6760$$

When we analyse the cases of disagreement and partial agreement we find that most of it has to do with SemLex lexias rather than NEs. This is mostly due to imperfectness of the dictionary and its size (annotators could not explore each of almost 30,000

[7] If we did not do this, there would be no difference between t-nodes, that were not tagged (annotators agreed they are not a part of a MWE) and the t-nodes that one annotator tagged and the other did not (i.e. they disagreed).

of SemLex entries). Our current methodology, which relies too much on searching the SemLex, is also to blame. This should, however, improve by employing pre-annotation (B) and (C).

One more reason for disagreement consists in the fact that there are cases, for which non-trivial knowledge of the world is needed: “Jang Di Pertuan Agong Sultan Azlan Šáh, the sultan of the state of Perak, [. . .] flew back to Perak.” Is “Sultan Azlan Šáh” still a part of the name or is it (or a part of it) a title?

The last important reason of disagreement is simple: both annotators identify *the same* part of text as MWE instances, but while searching the SemLex they choose different lexias as the tags. This can be rectified by:

- Removing duplicate entries from SemLex (currently there are many close identical entries originating from Eurovoc and Czech WordNet).
- Imploring improved pre-annotation B and C, as mentioned above.

7 Conclusion

We have annotated multi-word lexias and named entities in a part of PDT 2.0. We use tectogrammatical tree structures of MWEs for the automatic pre-annotation. In the analysis of inter-annotator agreement we show that a weighted measure that accounts for partial agreement as well as the estimation of maximal agreement is needed.

The resulting $\pi_w = 0.6760$ is statistically significant and should gradually improve as we clean up the annotation lexicon, more entries can be pre-annotated automatically, and further types of pre-annotation are employed.

8 Acknowledgement

This work has been supported by grants 1ET201120505 of Grant Agency of the Academy of Science of the Czech Republic, projects MSM0021620838 and LC536 of the Ministry of Education and 201/05/H014 of the Czech Science Foundation.

References

Ron Artstein and Massimo Poesio. 2007. Inter-coder agreement for computational linguistics. *Submitted to Computational Linguistics*.

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

D.A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Eurovoc. 2007. http://europa.eu/eurovoc/.

Josef Filipec and František Čermák. 1986. *Česká lexikologie*. Academia.

Josef Filipec. 1994. Lexicology and lexicography: Development and state of the research. In P. A. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*, pages 163–183, Amsterdam/Philadelphia. J. Benjamins.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Jan Hajič, Martin Holub, Marie Hučínová, Martin Pavlík, Pavel Pecina, Pavel Straňák, and Pavel Martin Šidák. 2004. Validating and improving the Czech WordNet via lexico-semantic annotation of the Prague Dependency Treebank. In *LREC 2004*, Lisbon.

Jan Hajič, 2005. *Insight into Slovak and Czech Corpus Linguistics*, chapter Complex Corpus Annotation: The Prague Dependency Treebank, pages 54–73. Veda Bratislava, Slovakia.

Eva Hajičová, Barbara H. Partee, and Petr Sgall. 1998. *Topic-focus articulation, tripartite structures, and semantic content*, volume 71 of *Studies in Linguistics and Philosophy*. Kluwer, Dordrecht.

Milena Hnátková. 2002. Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*.

A. Kilgarriff. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proc. LREC*, pages 581–588, Granada.

Rada Mihalcea. 1998. Semcor semantically tagged corpus.

Petr Pajas and Jan Štěpánek. 2005. A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague DependencyTreebank 2.0. Technical Report TR-2005-29, ÚFAL MFF UK, Prague, Czech Rep.

Petr Pajas. 2007. TrEd. http://ufal.mff.cuni.cz/∼pajas/tred/index.html.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Third International Conference, CICLing*.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Zpracování pojmenovaných entit v českých textech (treatment of named entities in czech texts). Technical Report TR-2007-36, ÚFAL MFF UK, Prague, Czech Republic.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/Reidel Publ. Comp., Praha/Dordrecht.

Pavel Smrž. 2003. Quality control for wordnet development. In Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 206–212. Masaryk University Brno, Czech Republic.