Finalising Multiword Annotations in PDT

Eduard Bejček and Pavel Straňák and Jan Hajič

Charles University in Prague, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
E-mail: {bejcek,stranak,hajic}@ufal.mff.cuni.cz

Abstract

We describe the annotation of multiword expressions and multiword named entities in the Prague Dependency Treebank. This paper includes some statistics of data and inter-annotator agreement. We also present an easy way to search and view the annotation, even if it is closely connected with deep syntactic treebank.

1 Introduction

The units of tectogrammatical layer of the Prague Dependency treebank should not be just words but lexemes. Essential for this improvement is annotation of multiword expressions. Its goal is to identify the multiword lexemes that should become single tectogrammatical nodes in future.

As there was a lack of such annotated data, the project Bejček, Straňák, and Schlesinger (2008) started four years ago. In this project (usually) two annotators have been reading the newspaper texts from PDT 2.0 (see Hajič et al., 2006), searching for multiword expressions in it, and annotating them. They have concentrated on both multiword named entities (NEs) and multiword lexemes. The aim of the project was to develop reliable training data for further research and to improve the state of "t-nodes" in the trees of PDT (see below in Section 1.1 about PDT). As a side effect, the lexicon of multiword lexemes was created, entries have been inserted into it and existing ones have been corrected throughout the whole annotation process.

Now the project is near the end. What was our approach, what are the results, are there any interesting outputs?

The paper starts with a few words about the corpus we used and the existing annotation we both use and enhance. The second section is about our annotation and the approach we chose. Section 3 brings overview of the annotation done, various types of inter-annotator agreement etc. At the end we report on the technical aspects of releasing the data, and how it can be used, which is work in progress.

1.1 PDT

The Prague Dependency Treebank 2.0 (Hajič, 2005) includes rich annotation on a deep syntactic level of almost 50,000 sentences (for details on this "tectogrammatical layer" see Mikulová et al., 2006). On this layer, each node should correspond to one lexeme, but this is not the case now: multiword expressions (MWEs) are still represented by several nodes each.

As mentioned above, our goal is to integrate MWEs like "New York City", "computational linguistics" or "kick the bucket" each into one joint node in a syntactic tree (with a meaning "NE: place", "lexicon entry: 39485 [gloss: science branch]" and "lexicon entry: 13985 [synonym: die]", respectively).

The advantage of the existing syntactic annotation of our source text was the possibility of preannotation and consistency checking. That could be done because all instances of the same MWE should have the same tree structure.[1]

2 The Way of Annotation

In this section, we introduce very briefly our project. Much more information can be found in Bejček and Straňák (2009).

Both multiword NEs and lexemes – if multiword – are called MWEs in this paper. Assigning the type of the MWE itself (such as a particular lexicon entry) is of little importance comparing to the fact, that it is found and its boundaries are marked. To create a typology of NEs (or phrasemes) is not our aim. Our effort leads to simple annotation guidelines for annotators; concrete labels are just an aid for further classification. Thus we adopt nine main types of NEs from Ševčíková et al. (2007)[2] (such as person name, name of a place, address etc.) and use more than 5,000 lexicon entries for other lexemes (such as phrasemes, non-compositional or non-substitutable collocations).[3] These lexicon entries have been collected from three lexicons and the set has been extended by annotators. Thereby the lexicon called SemLex was developed.

For the majority of time, we had two annotators, who annotated the same texts in parallel. (Overall, we had five annotators during the time as it can be seen in Table 2, but that is not crucial.) However, when we had enough data for inter-annotator agreement evaluation, we stopped parallel annotations (only with an occasional testing parallel document). The amount of data annotated in parallel can be seen in Table 1. By now, 85 % of the whole data is annotated.

[1] There are some marginal cases where the structure is not exactly the same. These differences will hopefully disappear in future, perhaps in PDT 3.0.

[2] This typology (with embedded types) was used for manual annotation of a corpus (Kravalová et al., 2009).

[3] The main simple criterion was "principle of compositionality"—whether it could be disassembled to parts that compose the meaning; if not, it should be in the lexicon. The second one is "substitutability of a part"—i.e. the possibility to substitute its component words with synonyms. Then we have annotation guidelines and meetings, where problems are solved.

amount of parallel annotations	in nodes	in % of PDT
three annotators	464	14.7 %
two annotators	1201	38.1 %
one annotator	1044	33.1 %
total	2709	85.8 %
total by at least two annotators	1665	52.8 %
PDT t-layer	3156	100 %

Table 1: The amount of single, double and triple parallel annotations.

For preannotation of the text we use Czech_geo_named_ent_recognizer and Czech_named_ent_SVM_recognizer from the Tecto-MT framework (Žabokrtský et al., 2008). These find some NEs. We also use external preannotation of phrasemes provided by our colleague (see Hnátková (2002)).

2.1 GUI

We developed a tool for our task, a GUI for annotators. Although we actually annotate the nodes in the trees in the background, we need to show only plain sentence to annotators. For each syntactic tree, the surface sentence is generated and every annotator's operation on it is converted back into the nodes and saved. In addition, as the subtree forming just annotated expression is identified, all the other occurrences of that subtree can be found in the neighbourhood automatically. That assists annotators with their manual work. Annotators are also allowed to view, modify and extend the SemLex in the same GUI.

This tool can be used for any annotations of treebank, where the annotated trees are better viewed as plain text.

2.2 Merging SemLexes

An annotator works off-line. Also their SemLex is modified off-line. Therefore the longer the annotation proceeds the more their SemLexes differ. Then we need to merge SemLexes and return the new one back to them for further annotation. The process described bellow is for two annotators.

First, we merge all entries that could be merged automatically. That means either entries that were the same in both SemLexes, or entries that were inserted into just one of the SemLexes.

Second, conflicting entries are delegated to third annotator who decides the correct forms. For this task we use a modal editor.[4] That means an editor with one mode for typing a text (which we disable) and other for macro executing. Macro invocation could be very simple—we use only one key for each operation. These are the reasons why we use a modal editor for manual merging.

[4] Vim in our case, http://www.vim.org/.

We put all conflicting lexicon entries into a file, each as a simple list of its values, and provided a syntax highlighting for the editor. Where there are more values (i.e. a conflict), we show all of them in a warning colour. We prepare some simple macros for the editor, such as "go to next conflict", "choose first value", "insert a comment" etc. and disable the option of typing in the text. When the conflict is resolved, the warning colour disappears. It confirmed that it is very fast to create and very simple and safe to use.

Third, the decisions from the third annotator were imported back into the SemLex and added to merged entries from the first step. After that, the annotated data are modified to correspond with the new SemLex.

3 Statistics

All statistics presented in this section are calculated for all our users. They didn't annotate the same data, though, as can be seen in Table 2. This table also shows the ratio of PDT annotated by each of them.

annotator\part	PDT	amount
#1	● ●	2.7 %
#2	●●●●●●●● ●●●●●●●●	55.0 %
#3	●●●●●●●●●●●●●●● ●●● ● ●●	67.2 %
#4	●● ● ●●● ● ●●	21.2 %
#5	●●●● ●●●	13.4 %

Table 2: Annotated parts and the ratio to the whole PDT per each annotator.

The annotated parts of PDT slightly differ,[5] but the overall characteristic stays. There is very similar usage of NEs across all annotators[6] in the Table 3.

Besides nine types of NEs, the annotators use approximately 8,000 of SemLex entries; some of them 100×,[7] third of them only once. Since there is no straight borderline stating whether an occurrence is a NE or shether it should be marked as a SemLex entry, the agreement has to be evaluated together for NEs and SemLex entries.

[5] For example after a changeover from one newspaper to another.

[6] Only annotator #1 evidently differs, but this one annotated only less than 3 % of PDT.

[7] Foremost lexicon entries are "state budget", "annual meeting", "environment", "join stock company" etc.

annotator	#1	#2	#3	#4	#5
address	-	0.4 %	0.1 %	0.6 %	0.7 %
biblio	-	0.1 %	0.2 %	-	0.0 %
foreign	0.2 %	0.7 %	0.5 %	1.0 %	0.6 %
institution	9.7 %	22.6 %	19.4 %	24.1 %	21.7 %
location	6.2 %	6.1 %	8.4 %	8.5 %	12.3 %
object	30.6 %	10.3 %	14.2 %	16.1 %	14.7 %
other	3.2 %	13.1 %	16.3 %	10.9 %	15.6 %
person	38.2 %	30.9 %	32.0 %	30.9 %	26.5 %
time	12.1 %	15.9 %	8.8 %	7.8 %	7.8 %
All NEs	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %

Table 3: Usage of named entities by particular annotators

3.1 Weighted Kappa Agreement

Because of the complicated character of our annotations,[8] we use weighted kappa for inter-annotator agreement in Table 4.

annotators	#1	#2	#3	#4	#5
#1	*	0.61	-	-	-
#2	0.61	*	0.56	0.21	-
#3	-	0.56	*	0.55	0.73
#4	-	0.21	0.55	*	0.70
#5	-	-	0.73	0.70	*

Table 4: Pairwise weighted kappa.

3.2 Agreement with "`is_name_of_person`" in PDT

There is an attribute called `is_name_of_person` in PDT, which is though assigned automatically. Figure 5 shows the agreement with our annotation. The corresponding value should be "named entity: person" and it is in 86 % of nodes.

There are three reasons, why the value does not correspond:

- The name of the person is not multiword. The connection of two `is_name_of_persons` was not marked in PDT, therefore the dependency edge between two such nodes may or may not create a MWEs; they could be names

[8] Firstly, the annotations are assigned to nodes, which are in m:n mapping to words. Secondly, two annotators happen to annotate differently although their annotations have non-empty intersection. Thirdly, some MWEs fall into more than one category, like phraseme and a name of an institution at the same time.

[9] Such node was annotated, but not as "person". The name of person could be part of a name of institution or even part of some multiword lexeme.

Annotators	PDT	Ratio
t-node annotated as		
NE “person”	nothing	4.5 %
other[9]	`is_name_of_person`	2.3 %
nothing	`is_name_of_person`	6.9 %
NE “person”	`is_name_of_person`	86.3 %

Table 5: The agreement between the PDT attribute `is_name_of_person` and named entity “person”.

of two people as well. One such example is “Pucciniho Turandot” (Puccini’s Turandot), which is the name of the composer and the name of the title character of his opera.

- The name of the person is a part of another MWEs. In that case, only the larger one should be annotated. For example, there is a name of person in “Pěvecký recitál Petera Dvorského” (Peter Dvorský’s Choral Recital), but it is an object as a whole.
- There is a mistake in our annotation (or in PDT, theoretically).

There are some others annotations in PDT (namely FPHR, DPHR, IDPH, and CPHR[10]), which are significant for us. These disagreements will be checked and mistakes will be repaired—either automatically, or manually.

4 Publication of the Data

As PDT is stored in the PML format (Pajas and Štěpánek, 2005), we also use PML for our annotations. That allows us to store the data as a stand-off annotation in so called “s-files” separately from the rest of PDT annotations. The MWEs found in each document are saved in a file linked with other files in PDT containing all the other annotations of this document as well as the surface sentences. (See Figure 1.)

This format allows us to show annotations in editor and viewer TrEd (Pajas and Štěpánek, 2008). It can present any part of the annotation in easily comprehensible, uncluttered way. User may choose to show or hide many detailed information about every word in the sentence (or node in the tree). There is also PML-TQ extension in TrEd, which allows a user to ask the queries in an easy user-friendly way. The queries are translated into SQL and evaluated by the database server containing treebanks. The resulting trees can be displayed either in TrEd, or on the web page using SVG. An example of a query is in Figure 2. Another query (in Figure 3) combines more layers of annotation.

Our complete data will be released under PDT licence within several months.

[10] These are “foreign phrase”, “dependent part of a phraseme”, “identification structure”, and “copula verbonominal predicate”, respectively.

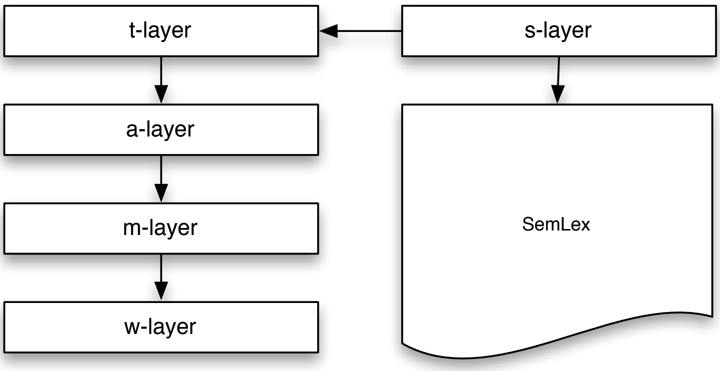
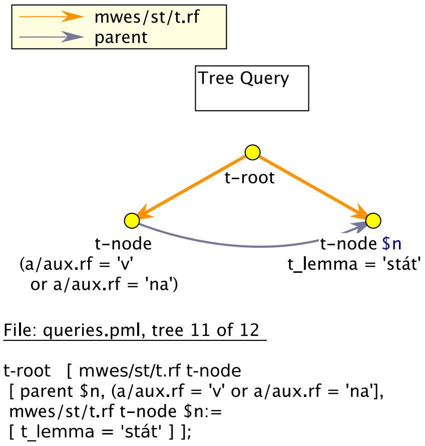


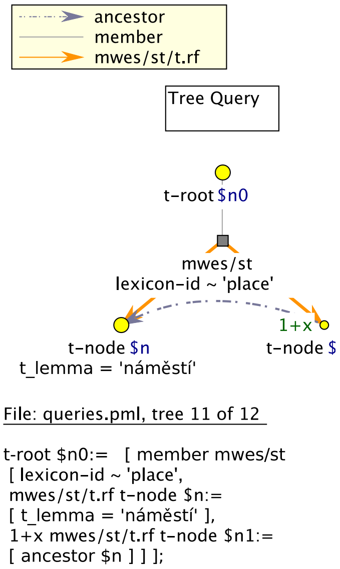
Figure 1: One document is stored in five interlinked files with links also to SemLex.

We will release the s-files themselves, which are very simple XML files, as well as scripts to merge the s-files with PDT to enable searching via PML-TQ and displaying the trees in TrEd. Export to CoNLL format (Hajič et al., 2009) will be also provided. CoNLL format is a simple table, which enables the data to be directly processed by many statistical tools. The GUI for annotators will be released as well.



```
t-root  [ mwes/st/t.rf t-node
 [ parent $n, (a/aux.rf = 'v' or a/aux.rf = 'na'],
 mwes/st/t.rf t-node $n:=
 [ t_lemma = 'stát' ] ];
```

Figure 2: This query searches for all annotated MWEs, such that it consists of the phrase "stát v ..." or "stát na ..." (meaning "stand in/at/on" as well as "hold ground", "keep sentinel", "tiptoe" etc.).



```
t-root $n0:=  [ member mwes/st
 [ lexicon-id ~ 'place',
 mwes/st/t.rf t-node $n:=
 [ t_lemma = 'náměstí' ],
 1+x mwes/st/t.rf t-node $n1:=
 [ ancestor $n ] ] ];
```

Figure 3: A query searching for a NE of a place containing the word "náměstí" (square) and at least one other word.

5 Conclusion

Identification of MWEs moves the Prague Dependency Treebank towards better separation of tectogrammatical lemmas from the morphological lemmas and thus closer to the Functional Generative Description (Sgall et al., 1986), i.e. the theoretical framework the PDT was built upon. In future, this work will help to produce such PDT t-layer, where all units will correspond to whole NEs or lexemes (and so some of them will be multiword).

We employed several different methods to optimise the annotation both in terms of speed and precision. We will continue further refinements of tectogrammatical lemmas before the next release of the treebank.

Acknowledgement

This work has been supported by grant 1ET201120505 of Grant Agency of the Academy of Science of the Czech Republic, and grants GAUK 4307/2009 and GAUK 4200/2009 of the Grant Agency of Charles University in Prague.

References

Eduard Bejček, Pavel Straňák, and Pavel Schlesinger. Annotation of multiword expressions in the prague dependency treebank. In *IJCNLP 2008 Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 793–798, 2008. 1

Eduard Bejček and Pavel Straňák. Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation*, 43(3), 2009. 2

Jan Hajič. *Insight into Slovak and Czech Corpus Linguistics*, chapter Complex Corpus Annotation: The Prague Dependency Treebank, pages 54–73. Veda Bratislava, Slovakia, 2005. ISBN 80-224-0880-8. 1.1

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA, 2009. 4

Jan Hajič, Jarmila Panevová, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. Prague Dependency Treebank 2.0, 2006. Published by Linguistic Data Consortium, Philadelphia, PA, USA. 1

Milena Hnátková. Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*, 2002. 2

Jana Kravalová, Magda Ševčíkov á, and Zdeněk Žabokrtský. Czech named entity corpus 1.0, 2009. 2

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., 2006. 1.1

Petr Pajas and Jan Štěpánek. 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., 2005. 4

Petr Pajas and Jan Štěpánek. Recent advances in a feature-rich framework for treebank annotation. In Donia Scott and Hans Uszkoreit, editors, *The 22nd International Conference on Computational Linguistics - Proceedings of the Conference*, volume 2, pages 673–680, Manchester, UK, 2008. The Coling 2008 Organizing Committee. ISBN 978-1-905593-45-3. 4

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. Zpracování pojmenovaných entit v českých textech (treatment of named entities in czech texts). Technical Report TR-2007-36, ÚFAL MFF UK, Prague, Czech Republic, 2007. 2

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/Reidel Publ. Comp., Praha/Dordrecht, 1986. 5

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, OH, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-09-1. 2