Lang Resources & Evaluation (2010) 44:7–21
DOI 10.1007/s10579-009-9093-0

Annotation of multiword expressions in the Prague dependency treebank

Eduard Bejček · Pavel Straňák

Published online: 11 August 2009

Abstract We describe annotation of multiword expressions (MWEs) in the Prague dependency treebank, using several automatic pre-annotation steps. We use subtrees of the tectogrammatical tree structures of the Prague dependency treebank to store representations of the MWEs in the dictionary and pre-annotate following occurrences automatically. We also show a way to measure reliability of this type of annotation.

Keywords Multiword expressions · Treebanks · Annotation · Inter-annotator agreement · Named entities

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 (MWEs) 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.

E. Bejček (✉) · P. Straňák (✉)
Institute of Formal and Applied Linguistics, Charles University in Prague, Prague, Czech Republic
e-mail: bejcek@ufal.mff.cuni.cz

P. Straňák
e-mail: stranak@ufal.mff.cuni.cz

There are various projects for identification of named entities (for an overview see Ševčíková et al. 2007). We explain below (mainly in Sect. 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 addition to a surface syntactic layer. This deeper layer is generally agreed to concern lexical meaning. To our best knowledge, the lexico-semantic annotations still deal with separate words, phrases are split and their parts are connected with some kind of dependency. Furthermore, only words with valency are involved in projects like NomBank (Meyers et al. 2004), PropBank (Palmer et al. 2005) or PDT.

1.1 Prague dependency treebank

We work with the Prague dependency treebank (PDT; see Hajič 2005), which is a large corpus with rich annotation on three layers: it has in addition to the morphological and the surface syntactic layers also the tectogrammatical layer. (In fact, there is also one non-annotation layer, representing the "raw-text" segmented into documents, paragraphs, and tokens.) Annotation of a sentence on the morphological layer consists of attaching several attributes to the tokens of the w-layer, the most important of which are morphological lemma and tag. A sentence at the analytical layer is represented as a rooted ordered tree with labeled nodes. The dependency relation between two nodes is captured by an edge with a functional label. The tectogrammatical layer has been construed as the layer of the (literal) meaning of the sentence and thus should be composed of monosemic lexemes 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 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 (see Sect. 3).

Figure 1 shows the relations between the neighboring layers of PDT.

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 monosemic lexeme, i.e. that

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

[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 and are combined into dependency trees, based on syntactic valency of the t-nodes.

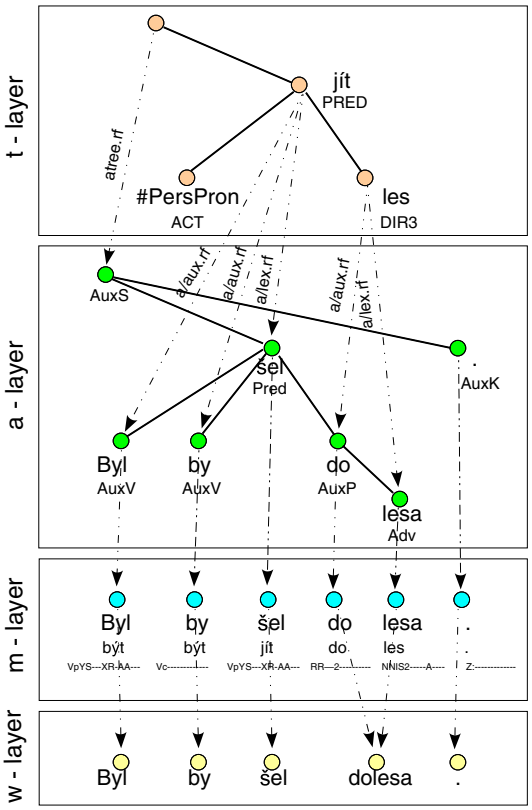

Fig. 1 The rendered Czech sentence *Byl by šel dolesa*. (lit.: He-was would went toforest.) contains past conditional of the verb "jít" (to go) and a typo "toforest" repaired on m-layer

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 MWEs and create a lexicon of the corresponding monosemic lexemes*. A simple view of the result of our annotations is given in the Fig. 2, some technical details are in Sect. 4.2.

2 Introduction

In our project we annotate all occurrences of MWEs (including named entities, see below) in PDT 2.0. When we speak of *MWEs* we mean "idiosyncratic interpretations that cross word boundaries" (Sag et al. 2002). We do not inspect various types of MWEs, because we are not concerned in their grammatical attributes. We only want to identify them. Once there will be a lexicon with them and their occurrences annotated in corpora, the description and sorting of MWEs will take place. We hope that annotation of a treebank will help—MWEs with fixed syntactic form will be easily distinguished from the others that can be modified by added words.

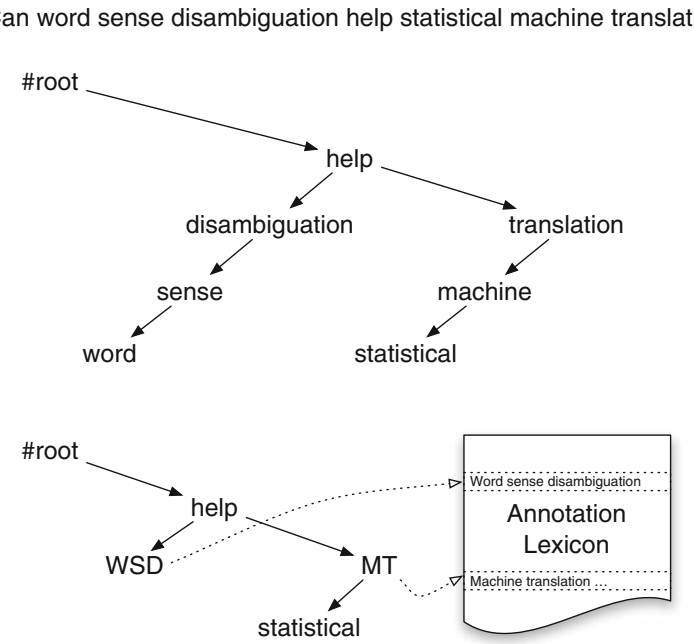

Fig. 2 Schema of the changes in t-trees after integration of our annotations; every MWEs forms a single node and has its lexicon entry

We distinguish a special type of MWEs, for which we are mainly interested in its type, 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 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 MWEs that was annotated. We explain this in detail in Sect. 4.1.

3 Current state of MWEs in PDT 2.0

During the annotation of valency, which is a part of the tectogrammatical layer of PDT 2.0, the t-lemmas, have been basically identified for all the verbs and some nouns and adjectives. The resulting valency lexicon is called PDT-VALLEX,

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

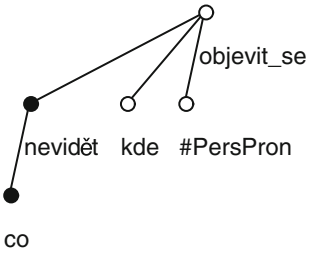

Fig. 3 Idiom *Co neviděť* meaning "in a blink (of an eye)", (literally: what not-see)

Hajič et al. (2003) and we can see it as a repository of lexemes based on verbs, adjectives and nouns in PDT that have valency.[4]

This is a starting point for having t-nodes corresponding to lexemes. 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 formally dependent t-nodes in a tectogrammatical tree (e.g. nodes with t-lemmas "co" in Fig. 3 or "k_dispozici" in Fig. 5). Phrasemes consisting of copulla "být" (to be) and a noun or adjective are also split into two nodes, where the nominal part is governed by the verb. Idioms that do not contain any morphological verb have either been annotated and assigned their own valency frames just like the above described verbal idioms (in case of idioms containing nouns derived from verbs by suffixes -ní or -tí), or (in case of the idioms consisting of only one t-node) have not been annotated at all in the current PDT. For detailed description see Sect. 6.8. of Mikulová et al. 2006).

In Figs. 3, 4, and 5 we give several examples of t-trees in PDT 2.0, that include idioms, light verb constructions and named entities:

4 Methodology

4.1 Building SemLex

Each entry we add into SemLex is considered to be a monosemic MWEs. We have also added nine special entries to identify NE types, so we do not need to add all the expressions themselves. These types are derived from the 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 their NE type), because this allows for their following occurrences to be pre-annotated automatically (see Sect. 5). For others, like addresses or bibliographic 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 MWEs. Its base has been composed of MWEs extracted from Czech

[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.

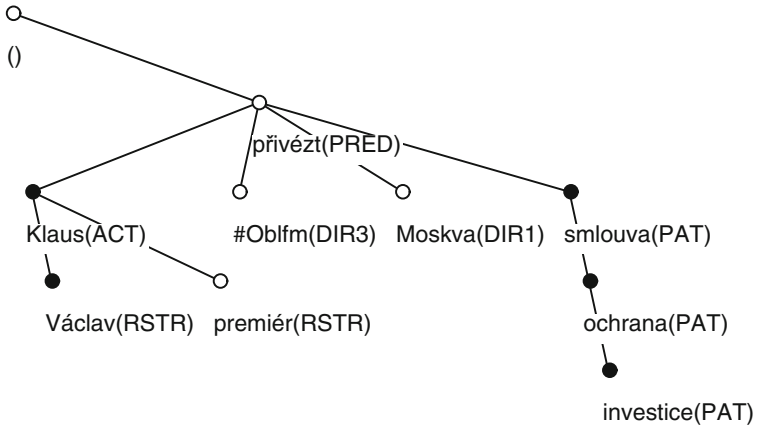

Fig. 4 A sentence featuring a personal name and a name of a bilateral treaty (which is not the exact official name, however, thus it is not capitalised)

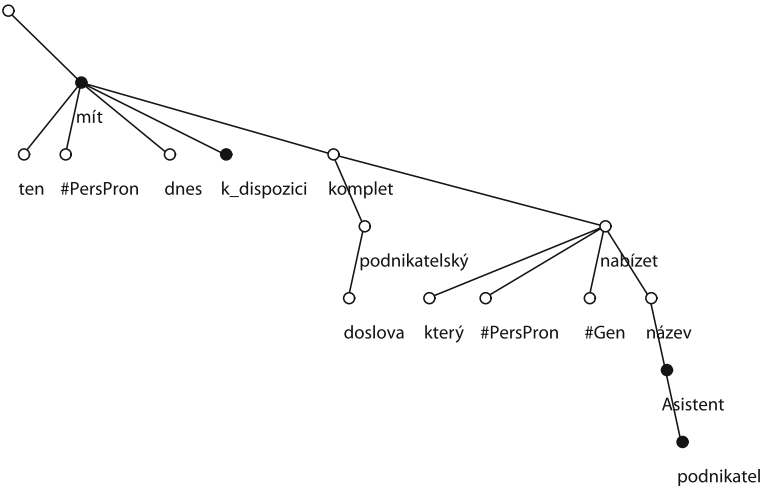

Fig. 5 A t-tree of a sentence featuring a light verb construction *mít k dispozici* (lit.: to have at [one's] disposal) and a named entity (a product name) *Asistent podnikatele* (lit.: assistant of-businessman) that looks like a common phrase, except for the capital 'A'

WordNet (Smrž 2003), Eurovoc (Eurovoc 2007) and Dictionary of Czech Phraseology and Idiomatics (Čermák et al. 1994). Currently there are over 30,000 MWEs in SemLex and more are being added during annotations.

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

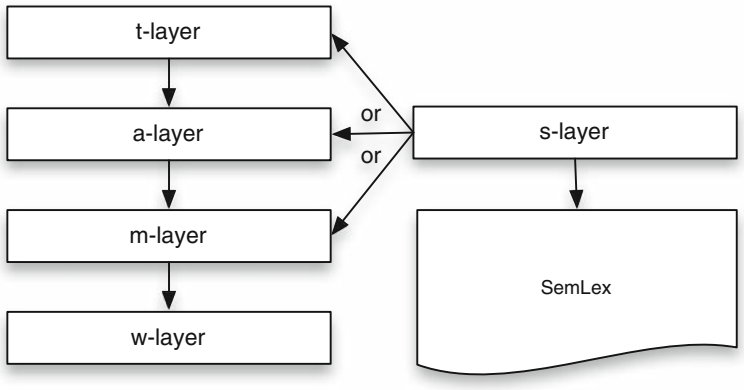

Fig. 6 Relation of s-layer to PDT and SemLex

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 corresponds to one file of PDT and consists of a list of s-nodes. Each s-node corresponds to an occurrence of a MWEs and is composed of a link to an entry in SemLex and a list of identifiers of t-nodes that correspond to this s-node. Figure 6 shows a relation of s-layer to PDT layers and SemLex.[5]

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 speed improvement and for improving the consistency of annotations. On the assumption that *all occurrences of a MWEs 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:

[5] Although we have created the PML schema of s-layer primarily for annotations of MWEs, we made it quite generic. It can be utilised for any treebank annotations that use a large lexicon. For instance one s-file can contain multiple annotations of valency referencing to different valency dictionaries. This generic nature of s-layer is the reason why it allows references to morphological, analytical or tectogrammatical layer of PDT, even though in our current project we only need the references to t-layer.

(A) External pre-annotation provided by our colleague (see Hnátková 2002). With each MWEs a set of rules is associated that limits possible forms and surface word order of parts of a MWEs. This approach was devised for corpora that are not syntactically annotated and is very time consuming.
(B) Our one-time pre-annotation with those MWEs from SemLex that have been previously 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 (while annotating previous files).
(D) When an annotator tags an occurrence of a MWEs in the text, other occurrences of this MWEs in the article are identified automatically.[6]

Pre-annotation (A) was executed once for all of the PDT. (B) is performed each time we merge MWEs added by annotators into the main SemLex. We carry out this annotation in one batch for all PDT files remaining to annotate. (C) is done for each file while it is being opened in the annotation environment. (D) happens each time the annotator adds a new MWEs into SemLex and uses it to annotate an occurrence in the text. In subsequent files instances of this MWEs 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 MWEs happened to have two different tree representations. Below we analyse the most important sources of these inconsistent t-trees and possible improvements:

– *Occasional lemmatisation errors*. They are not very frequent, but there is no efficient way to find and correct them before the annotations. There is not much we can do about them, our annotations can however serve as a source for automatic corrections.
  - e.g. wrongly lemmatised *jižní Korea* vs. correct *jižní Korea* (southern vs. South Korea)
– *Annotator's mistake (not marking correct words)*. When an annotator makes an error while marking a first occurrence of a MWEs, the tree representation that gets stored in SemLex is incorrect. As a result, pre-annotation gives false positives or fails to work.
  It is therefore necessary to allow annotators to correct the tree structure of a SemLex entry, i.e. extend functionality of the annotation tool. Once all the types of pre-annotation are employed, this error can happen only once, because all the following occurrences of a MWEs are pre-annotated automatically. We are currently working on these improvements.
– *Gender opposites, diminutives and augmentatives*. These are currently represented by different t-lemmas. We believe that they should be represented by attributes of t-nodes that could be roughly equivalent to some of the lexical

[6] This is exactly what happens: (1) Tree structure of the selected MWEs 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 MWEs using its tree structure (via TrEd), and (4) Other occurrences returned by TrEd are tagged with this MWEs' 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.

functions in the Meaning-text theory (see Mel'čuk 1996). This should be tackled in some future version of PDT. Once resolved it would allow us to identify following (and many similar) cases automatically.

- *obchodní´ředitel* vs. *obchodní´ředitelka*
  (lit.: managing director-man vs. m. director-woman)
- *rodinný dům* vs. *rodinný domek*
  (lit.: family house vs. family little-house; but the diminutive *domek* means basically "family house", as opposed to "an apartment building")

So as a step towards this goal we annotate these cases as occurrences of the same MWEs. A drawback of this solution is that automatic pre-annotation (types B–D) can't identify the instances with the derived variants of t-lemma (like *ředitelka* or *domek*), since these pre-annotations rely on t-lemmas. Thus these variants of MWEs must be identified manually for now.

- *Newly established t-nodes corresponding to elided parts of MWEs in coordinations*. Since t-layer contains many newly established t-nodes, many of whom cannot be lexicalised, our original decision was to hide all of these nodes from annotators and generate for them pure surface sentence. This decision resulted however in the current situation, when some MWEs in coordinations cannot be correctly annotated. For instance *První´a druhá světová válka* (First and Second World War) is a coordination of two multiword lexemes. A tectogrammatical tree that includes it does have newly established t-nodes for "world" and "war" of the first lexeme but they are (and in fact they have to be) elided in the surface sentence.

  After analysing annotated examples like the one above we have decided to generate surface words from some of the newly established t-nodes in order to allow correct annotation of all the MWEs. These "added" words will be displayed in grey and while some morphological forms of these words may be incorrect, we believe they will serve their purpose.

Up to now we have not found any MWEs such that its structure cannot be represented by a single tectogrammatical tree. 1.1% of all occurrences were not connected graphs, but this happened due to errors in data and to our incorrect handling of coordinations with newly established t-nodes (see above). This corroborates our assumption that (disregarding errors) all occurrences of a MWEs 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-annotations (A) and (D).

6 Analysis of annotations

Two annotators have 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

Table 1 Annotated instances of significant types of MWEs

Type of MWEs	A	B
SemLex entries	8,447	8,312
Different items	3,844	4,089
Named entities	8,435	8,903
Person/animal	2,797	2,811
Institution	1,702	2,047
Number	1,343	1,053
Object	1,129	888

tags in the form of different colour labels and they can delete the pre-annotated tags. In this experiment the data consists of approx. 310,000 tokens, which correspond to 250,000 t-nodes. Both annotators have marked about 37,000 t-nodes ($\approx$15%) as parts of MWEs and grouped them into 17,000 MWEs. So the average length of a MWEs is 2.2 t-nodes.

The ratio of general named entities versus SemLex entries was 50:50 for annotator *A* and 52:48 in the case of annotator *B*. Annotator *A* used SemLex more frequently (than she used named entities and also than annotator *B* used SemLex), but did not utilise as many lexicon items as annotator *B*. This and some other comparison is given in Table 1.

Both annotators also needed to add missing entries to the originally compiled SemLex or to edit existing entries. Annotator *A* added 1,361 entries while annotator *B* added 2,302. They modified 1,307 and 2,127 existing entries, respectively.

6.1 Measuring inter-annotator agreement

In this section our primary goal is to assess whether with our current methodology we produce a reliable annotation of MWEs. To that end we measure the amount of inter-annotator agreement that is above chance. Our attempt exploits *weighted kappa measure* κ_w (Cohen 1968).

The reason for using a weighted measure is essential: we do not know which parts of sentences are MWEs and which are not. Therefore annotators work with all words and even if they do not agree on the type of a particular MWEs, it is still an agreement on the fact that this t-node is a part of some MWEs and thus should be tagged. This means we have to allow for partial agreement on a tag.

There are, however, a few sources of complications in measuring agreement of our task even by κ_w:

- Each tag of a MWEs 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). Instead, standard measures like κ assumes fixed, bounded items, which are assigned some categories.
- There is no clear upper bound as to how many (and how long) MWEs there are in texts. Cohen's κ_w counts only agreement on known items and these are the

same for both annotators. We, on the other hand, want to count also agreement on the fact that a given word is not a part of MWEs.
– There is not a clear and simple way to estimate the amount of 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 issues above, we have decided to start from Cohen's κ_w [quoted from (Artstein and Poesio 2007)]:

$$\kappa_w = 1 - \frac{D_o}{D_e} = \frac{A_o - A_e}{1 - A_e} \tag{1}$$

(further explained in Eq. 3) and to make a few adjustments to allow for an agreement on non-annotation and an estimated upper bound. We explain these adjustments in following paragraphs.

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 just the t-nodes that "should be annotated". This also means that the theoretical maximal agreement (upper bound) U cannot be 1 (as in the denominator of Eq. 1). If it were 1, it would be saying that all nodes are part of MWEs.

Since we know that $U < 1$ but we do not know its exact value, i.e. we do not know the 'correct' ratio of MWEs and NEs in a text, we use the *estimated upper bound* $\widehat{U}$ (see Eq. 2). 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 on a t-node not being a part of a MWEs, i.e. not tagged at all. This allows us to positively discriminate the cases where annotators agree that a t-node is not a part of a MWEs from the cases where one annotator annotates a t-node and the other one does not, which is evidently worse.

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:

$$\widehat{U} = \frac{n_{A\cup B}}{N} + 0.051 \times \frac{N - n_{A\cup B}}{N} = 0.213. \tag{2}$$

The weight 0.051 used for scoring the t-nodes that were not annotated is explained below (class $c = 4$). Because there is some disagreement of the annotators and we count all these nodes as annotated ($n_{A\cup B}$) for calculation of $\widehat{U}$ we believe that the real upper bound U lies somewhat below it and the agreement value 0.213 is not something that should (or could) be achieved. It is however important to note that this is based on the assumption that the data we have not yet seen have similar proportion of MWEs as the data we have used for the upper bound estimate. Since the PDT is composed of only news articles, the assumption seems reasonable.

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

$c = 1$ If the annotators agree on the exact tag from SemLex, we get maximum information: $w_1 = 1$.

$c = 2$ If they agree that the t-node is a part of a NE or they agree that it is a part of some entry from SemLex, but they do not agree which NE or which entry, we estimate we get about a half of the information compared to when $c = 1$: $w_2 = 0.5$.

$c = 3$ If they agree that the t-node is a part of a MWEs, but disagree whether a NE or an entry from SemLex, it is again half the information compared to when $c = 2$, so $w_3 = 0.25$.

$c = 4$ If they agree that the t-node is not a part of a MWEs, $w_4 = 0.051$. This low value of w accounts for frequency of t-nodes that are not a part of a MWEs, 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:

$$w_4 = w_3 \times \frac{\sum annotated}{\sum not\ annotated} = 0.25 \times \frac{42779}{208437} \approx 0.051.$$

We can see that even two ideal annotators who agree on all their assignments could not reach agreement $U = 1$, since they naturally leave some t-nodes without annotation and even if they are the same t-nodes for both of them, this agreement is weighted by w_4. Now we can look back at Eq. 2 and see that $\widehat{U}$ is exactly the agreement which two ideal annotators reach.

$c = 5$ If the annotators do not agree whether to annotate a t-node or not, $w_5 = 0$.

It should be explained why the upper bound does not need to be corrected in other weighted measures like κ_w. There are weights for some types of disagreement in κ_w to distinguish "better" disagreement from "worse" one, too. But it is still a disagreement and annotators could agree completely. In our task, on the contrary, this class $c = 4$ represents an agreement of its kind. The reason, why we do not count it as an agreement, is the biased resulting measure, if we do so: The less they would annotate the higher the agreement would be (if they annotated nothing, κ_w would equal 1).

We have also measured standard κ without weights. All partial (dis)-agreements had to be treated as full disagreements, because of lack of a weight function. In κ_1 we counted every non-annotated t-node as a disagreement, too; in κ_2 we think of non-annotation as a new category, so it is counted as an agreement. The difference is quite clear ($\kappa_1 = 0.04$ and $\kappa_2 = 0.68$). κ_2 might seem as a usable measure, even though over-generalising. However we can see that agreement on not-annotating is again counted as equally valuable as full agreement on a MWEs. Thus it has the same problem as κ_w without the class $c = 4$, as explained above.

The numbers of t-nodes n_c and weights w per class c are given in Table 2.

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

$$\kappa_w^U = \frac{A_o - A_e}{\widehat{U} - A_e} = \frac{D_e - D_o}{\widehat{U} - 1 + D_e}. \tag{3}$$

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

Table 2 The agreement per class and the associated weights

	Agreement				Disagreement
	Annotated			Not annot.	
	Agr. on NE / SL entry				
	Full agr.	Disagr.			
Class, c	1	2	3	4	5
# of t-nodes, n	24,386	6,355	1,399	208,437	10,639
Weight, w	1	0.5	0.25	0.051	0
$w_c n_c$	24,386	3,178	350	10,695	0

agreement above chance and maximum agreement above chance. In equivalent (and often used) definition, D_o and D_e are observed and expected disagreements.

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_c (see Table 2), summing these five weighted sums and dividing this sum of all the observed agreement in the data by the total number of t-nodes:

$$A_o = \frac{1}{N}\sum_{c=1}^{5} w_c n_c = \frac{1}{251216}(24386 + 3178 + 350 + 10695 + 0) \doteq 0.154.$$

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. (Let us say a combination of tags i and j has a probability p_{ij} and is weighted by w_{ij}.)

We estimated these probabilities from annotated data

$$A_e = \sum_{i}^{SemLex}\sum_{j}^{SemLex} \frac{n_{q_iA}}{N_A}\frac{n_{q_jB}}{N_B} w_{ij} \approx 0.046,$$

where n_{q_iA} is the number of lexicon entry q_i in annotated data from annotator A and N_A is the amount of t-nodes given to annotator A. Here, the non-annotation is treated like any other label assigned to a t-node.

The resulting κ_w^U is then

$$\kappa_w^U = \frac{A_o - A_e}{\widehat{U} - A_e} = \frac{0.154 - 0.046}{0.213 - 0.046} \doteq 0.644.$$

6.2 Interpretation of inter-annotator agreement

When we analyse disagreement and partial agreement we find that mostly it has to do with SemLex entries rather than NEs. This is due to the problems in the

dictionary and its size (annotators cannot explore each of almost 30,000 SemLex entries), but also our current methodology that relies too much on searching the SemLex. This should, however, improve by employing pre-annotations (B) and (C).

One more source of disagreement are the sentences for which non-trivial knowledge of the world is needed: "Jang Di Pertuan Agong Sultan Azlan Shah, the sultan of the state of Perak, [...] flew back to Perak." Is "Sultan Azlan Shah" still a part of the name or is it (or is a part of it) a title?

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

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

We introduced generalised κ_w^U measure, which is Cohen's weighted kappa with the upper bound $U \leq 1$, and we argue why such generalisation is essential for annotation project of this kind.[7] We also explain why and how the estimation of the upper bound of annotations should account for a difference between (agreement on) not annotating a unit (a t-node) and disagreement on annotation.

The main problem with interpretation of our results is that we don't know of any direct comparison. We were not able to find any published results of inter-annotator agreement on a task like ours, i.e. task with no exact upper bound on a number of tags, and a possibility of partial agreement on the size and type of tags. Until our results are compared to other such projects, the informative value of our numbers is limited.

7 Conclusion

We have annotated multi-word lexemes and named entities on a part of PDT 2.0. We use tectogrammatical tree structures of MWEs for automatic pre-annotation. In Sect. 5 we show that the richer the tectogrammatical annotation the better the possibilities for automatic pre-annotation that minimises human errors. In the analysis of inter-annotator agreement we show that a weighted measure that accounts for partial agreement as well as estimation of maximal agreement is needed.

The resulting $\kappa_w^U = 0.644$ should gradually improve as we clean up the annotation lexicon, more entries are pre-annotated automatically, and further types of pre-annotation are employed.

Acknowledgements This work has been supported by grants 1ET201120505 and 1ET100300517 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 and a grant GAUK 4307/2009 of the Grant Agency of Charles University in Prague.

[7] In our previous work we used a weighted variant of π (which does not reflect individual coders' distributions) with the same result as $\kappa_w^U : \pi_w = 0.644$). See Bejček et al. (2008).

References

Artstein, R., & Poesio, M. (2007). Inter-coder agreement for computational linguistics. *Submitted to Computational Linguistics*.

Bejček, E., Straňák, P., & Schlesinger, P. (2008). Annotation of multiword expressions in the Prague dependency treebank. In *IJCNLP 2008 Proceedings of the third international joint conference on natural language processing* (pp. 793–798).

Čermák, F., Červená, V., Churavý, M., & Machač, J. (1994). *Slovní k české frazeologie a idiomatiky*. Praha: Academia.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin 70*(4), 213–220.

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

Hajič, J. (2005). Complex corpus annotation: The Prague dependency treebank, Chap. In *Insight into Slovak and Czech corpus linguistics* (pp. 54–73). Bratislava, Slovakia: Veda.

Hajič, J., Holub, M., Hučinová, M., Pavlík, M., Pecina, P., Straňák, P., et al. (2004). Validating and improving the Czech WordNet via lexico-semantic annotation of the Prague dependency treebank. In *LREC 2004*, Lisbon.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., & Pajas, P. (2003). PDT-VALLEX. In J. Nivre & E. Hinrichs (Eds.), *Proceedings of the second workshop on treebanks and linguistic theories*, Vol. 9 of *Mathematical modeling in physics, engineering and cognitive sciences* (pp. 57–68). Vaxjo, Sweden: Vaxjo University Press.

Hajičová, E., Partee, B. H., & Sgall, P. (1998). *Topic-focus articulation, tripartite structures, and semantic content*, Vol. 71 of *Studies in linguistics and philosophy*. Dordrecht: Kluwer.

Hnátková, M. (2002). Značkováni frazémů a idiomů v Českém národním korpusu s pomoci Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*.

Kilgarriff, A. (1998). SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of LREC* (pp. 581–588). Granada

Krenn, B. & Erbach, G. (1993). *Idioms and Support Verb Constructions in HPSG*. Technical report, Universität des Saarlandes, Saarbrücken.

Mel'čuk, I. (1996). Lexical functions: A tool for the description of lexical relations in a lexicon. In L. Wanner (Ed.) *Lexical functions in lexicography and natural language processing*, Vol. 31 of *Studies in language companion series* (pp. 37–102). Amsterdam: John Benjamins.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., et al. (2004). The NomBank project: An interim report. In A. Meyers (Ed.), *HLT-NAACL 2004 workshop: Frontiers in corpus annotation* (pp. 24–31). Boston, MA, USA : Association for Computational Linguistics.

Mihalcea, R. (1998) SEMCOR Semantically tagged corpus.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., et al. (2006). *Annotation on the Tectogrammatical Level in the Prague Dependency Treebank Annotation manual*. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep.

Pajas, P. (2007). TrEd. http://ufal.mff.cuni.cz/~pajas/tred/index.html.

Pajas, P., & Štěpánek, J. (2005). *A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague Dependency Treebank 2.0*. Technical Report TR-2005-29, ÚFAL MFF UK, Prague, Czech Rep.

Palmer, M., Gildea, D., & Kingsbury, P. (2005) The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal 31*(1).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Third international conference, CICLing*.

Ševčiková, M., Žabokrtský, Z., & Krůza, O. (2007). *Zpracováni pojmenovaných entit v českých textech (Treatment of Named Entities in Czech Texts)*. Technical Report TR-2007-36, ÚFAL MFF UK, Prague, Czech Republic.

Sgall, P., Hajičová, E., & Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Praha/Dordrecht: Academia/Reidel Publishing Company.

Smrž, P. (2003). Quality control for wordnet development. In P. Sojka, K. Pala, P. Smrž, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the second international WordNet conference—GWC 2004* (pp. 206–212). Masaryk University Brno: Brno, Czech Republic.