Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0

Eduard BEJČEK
Jarmila PANEVOVÁ   Jan POPELKA
Pavel STRAŇÁK   Magda ŠEVČÍKOVÁ
Jan ŠTĚPÁNEK   Zdeněk ŽABOKRTSKÝ
Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Malostranské náměstí 25, Praha, Czech Republic
`{bejcek,panevova,popelka,stranak,sevcikova,stepanek,zabokrtsky}@ufal.mff.cuni.cz`

Abstract

We present the Prague Dependency Treebank 2.5, the newest version of PDT and the first to be released under a free license. We show the benefits of PDT 2.5 in comparison to other state-of-the-art treebanks. We present the new features of the 2.5 release, how they were obtained and how reliably they are annotated. We also show how they can be used in queries and how they are visualised with tools released alongside the treebank.

Title and Abstract in Czech

Pražský závislostní korpus 2.5 – rozšířená verze PDT 2.0

Představujeme nejnovější verzi Pražského závislostního treebanku PDT 2.5, který bude poprvé vydán pod veřejnou licencí. Výhody PDT 2.5 ukážeme na srovnání s nejmodernějšími treebanky. Představíme nové vlastnosti verze 2.5, popíšeme, jak byly anotovány i jak spolehlivá tato anotace je. Ukážeme, jakými dotazy lze nové jevy hledat a jak se zobrazují v nástrojích dodávaných spolu s treebankem.

Keywords: treebank, linguistic theory, PDT, annotation, syntax, semantics, multiword expressions, pair/group meaning, clauses.

Czech Keywords: treebank, lingvistická teorie, PDT, anotace, syntax, sémantika, víceslovné výrazy, souborovost, klauze.

1 Introduction

The task of grammatical theories is to explicitly describe phenomena of language with the purpose of creating a description that analyses and/or generates language as natural as possible. The creation of a treebank then serves as an ultimate test of a linguistic theory.

Treebanks that have not been based on an elaborate theory which takes into account most phenomena of natural language start with a simple design. If they become popular, various additional, more-or-less ad hoc linked projects end up being piled upon the original simple design.

On the other hand, there are very complex theories that have been being developed for decades to take into account all of the possible phenomena of natural language but have not yet undergone the ultimate test of large-scale treebanking (e.g. Mel'čuk and Polguère, 1987).

In this paper we introduce the Prague Dependency Treebank 2.5 (PDT 2.5), the latest instance of a large treebank whose design is based on the Functional Generative Description. It is a step from PDT 2.0 towards PDT 3.0 (coming in 2013 with additional large-scale annotation of discourse, anaphora and more) which brings annotation of three new features finished so far and a number of corrections.

The rest of the paper is organised as follows: we introduce some basic facts about PDT 2.5 and the previous version of the treebank in Section 2. In Section 3, we provide some background on treebanking projects in general and compare PDT 2.5 with other popular treebanks. Next we present the new features of PDT 2.5: annotation of multiword expressions in Section 4, new semantic distinction of pair/group meaning of nouns in Section 5, and identification of clauses in Section 6. We summarise the state of the treebank and its general ecosystem in Section 7.

2 Prague Dependency Treebank 2.5 and the previous version

Functional Generative Description (FGD) is a relatively complex linguistic theory and as such it has provided many fundamental ideas that are directly reflected in the PDT design, e.g. multiple layers of linguistic description and the dependency approach to syntax based on the theory of valency (Sgall et al., 1986; Sgall, 1967). Nonetheless, FGD does not encompass all phenomena either, not even in the language core.

In this paper we focus on PDT 2.5, which is an updated release of PDT 2.0. For this new release, the data of PDT 2.0 have been enriched with annotation of three new phenomena (see Sections 4 to 6). Furthermore, some of the errors in the PDT 2.0 data have been corrected in the new release (they mostly involved morphological tags and lemmas[1] for personal names and abbreviations, yet some of these changes were also reflected on the higher layers). However, the design of the PDT 2.5 annotation as well as the size of the data are identical with PDT 2.0.

PDT 2.0 (Hajič et al., 2006) is a collection of Czech newspaper texts from 1990s with annotation added on four layers: on the word layer (w-layer), the source texts have been tokenized and segmented. The morphological layer (m-layer) provides a lemma and a positional tag for each token (word form or punctuation mark). On the analytical layer

[1] About 10 thousand morphological nodes were fixed.

(a-layer), a sentence is represented as a dependency tree with labelled nodes and edges, which correspond to surface-syntactic relations (such as subject, object etc.); one analytical node corresponds to exactly one morphological token. On the tectogrammatical layer (t-layer), the meaning of the sentence is represented as a dependency tree structure with additional features and constraints.

Tectogrammatical nodes (t-nodes) represent content words (including pronouns and numerals), whereas functional words such as prepositions have no separate node in the tree.[2] All t-nodes are labelled with t-lemmas, dependency relations (functors, such as an actor ACT, addressee ADDR or location specification LOC) and grammatemes (see Section 5). Furthermore, annotation of valency, coreference, and topic-focus articulation are all available in tectogrammatical trees as well (Mikulová et al., 2006).

The PDT data consist of 7,110 manually annotated textual documents containing 115,844 sentences with 1,957,247 tokens (word forms and punctuation marks). All these documents were annotated on the m-layer, 75 % of them were annotated on the a-layer (5,330 documents, 87,913 sentences, 1,503,739 tokens). 59 % of the a-layer data were annotated also on the t-layer (i.e. 45 % of the m-layer data; 3,165 documents, 49,431 sentences, 833,195 tokens).

As we are improving the PDT by providing more explicit and consistent annotation guidelines, we are also improving the theoretical framework of FGD. The same is true when we add analysis of phenomena not tackled by the original theory. The theoretical studies preceded the annotation stage.

2.1 Prague Markup Language

PDT uses the PML format (Pajas and Štěpánek, 2006) based on XML. Each token and node has been assigned a unique identifier; any layer built atop of another uses the identifiers from the lower layer as reference targets, effectively creating inter-layer links (of various types). Each node can be assigned an attribute-value structure, an *attribute* in short, that represents various grammatical categories.

Another advantage of the PML format is the availability of the framework surrounding it. The tools provided include the tree editor `TrEd` (Pajas and Štěpánek, 2008), the query language and engine PML-TQ (Pajas and Štěpánek, 2009, see also Figure 2) and a highly modular NLP system Treex (Popel and Žabokrtský, 2010).

3 Related work

During the past decade, plenty of treebanks have been published. New treebanks keep appearing at least bi-monthly.[3] There are some features, though, that set PDT 2.5 apart from most of them.

The most popular treebank of all times is the Penn Treebank (Marcus et al., 1993). It has been since extended by several projects: PropBank (Palmer et al., 2005), NomBank (Meyers et al., 2004), BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005), and a few more.

[2] There are several exceptions of a technical nature. For instance, counterparts of coordinating conjunctions are included in the tree because they are used for the representation of coordinating constructions.

[3] LDC has published 5 new treebanks so far in 2012: `http://www.ldc.upenn.edu/Catalog/ByYear.jsp`

The pitfall of this process can be demonstrated on the Chinese Treebank (Xue et al., 2010), whose development followed the Penn Treebank pattern. The additional layers of annotation follow the stand-off principle, linking them to the original data. What remains problematic, though, is the format of these links: for example, both the proposition and word sense annotations use a "token number" to refer to a particular token in a sentence, but the latter only counts terminals with a surface form, while the former includes various added nodes as well (e.g. traces, pro's). Similarly, the coreference annotation (and named entity annotation, too) operates directly on the text, not taking the underlying phrase structure into account. Therefore, units that enter the coreference relations sometimes do not correspond to a continuous subtree of a tree.

As a result, it is substantially non-trivial to search for phenomena that involve several such layers (e.g. "list all the verbs at which a given named entity or a pronoun corefering to it can appear as Arg0").

The PML format (see Section 2.1), on the other hand, results in unambiguous interconnection of the annotation layers.

There are other projects aiming at standardisation of the solutions and conversion of old formats to new ones, cf. (Ide and Suderman, 2009). The solution used in the PDT is comparable to these efforts and standards (such as the LAF or TEI and its variants), but it has the added advantage of being supported by a complete suite of tools for annotation, search and processing mentioned earlier.

Finally, not all treebanks are freely available. Various license restrictions (and usage fees) exist. PDT 2.5 is now being distributed under the standard Creative Commons license (3.0-BY-NC-SA) allowing free access and distribution of additions and modifications.

4 Multiword expressions

Multiword expressions (MWEs) such as idioms, phrasemes, and multiword named entities are an important and sometimes overlooked part of natural languages. Usually they form a significant portion of the vocabulary, particularly in special domains where terminology is in play, but not only there: 16.3% of content words in the PDT are part of a MWE.

Multiword expressions are a boundary phenomenon on the interface of grammar and lexicon. We understand them, in accordance with Sag et al. (2002), Baldwin et al. (2003), Pecina (2009), and other authors, as phrases that contain some *idiosyncratic element* that differentiates them from normal expressions. This idiosyncratic element can be morphological, syntactic, or semantic. Although the annotation belongs to semantic layer, we have means for annotation on the other layers as well.

As a practical guideline for how idiosyncratic the expression must be to constitute a MWE, the most important criteria are the absence of *compositionality of meaning* and *word-for--word translation*. Neither of these criteria is absolute either by itself or together, but they are strong indicators, nonetheless. If these or any other secondary criteria compel a native speaker to add the expression to a dictionary because it requires an explanation, we consider it a MWE. For examples, see page 6.

Although some grammatical theories have accounted for MWEs decades ago (see e.g. Mel'čuk and Polguère, 1987), the annotation of MWEs is one of the least developed phenomena in treebanks. There were some MWEs annotated in PDT 2.0 (such as personal names or foreign phrases) and there are other treebanks that include named entities and/or MWEs to some extent, e.g. the German TüBa-D/Z or the Bulgarian BulTreeBank.[4]

In PDT 2.5, we annotate all occurrences of MWEs, including named entities (see below). We do not inspect various linguistic subtypes of MWEs in the treebank because we believe it is not the right place to analyse their grammatical attributes—only their instances should be identified in the treebank. Once that is done, a lexicon linked to their occurrences in corpora can be compiled and the MWEs (MWE types) can be analysed, taking into account all the information that can be acquired from the annotated occurrences in the data, as well as other resources. We have compiled a preliminary version of such a lexicon. It is complete with regard to lexemes occurring in the PDT 2.5 data and it is freely available.[5] The elaborate lexicographic analysis of all its entries, however, has yet to be performed. That is why the dictionary is not part of the PDT 2.5 release.

We distinguish a special type of MWEs: **named entities (NE)**.[6] In their case we are interested mainly in their *type* (see the list below) and their *basic form*. Since Czech is an inflectional language, a *basic form* of a MWE often differs from the form used in the sentence, but also from the sequence of basic forms (lemmas) of the individual words. This is illustrated by Examples 1 and 2. In the current release of PDT 2.5, basic forms have been manually added to some types of MWEs (see the full list below): lexemes, persons, locations, objects and some institutions. 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 to be able to display each NE as a single node, just like we do it with MWEs in general. See details in Section 4.1.

Tectogrammatical layer is the layer of linguistic meaning, so the MWE annotation belongs there. MWEs can be more easily captured on the t-layer, because: (i) there are added nodes for words not present in the surface sentence (ellipses), (ii) each MWE constitutes a **continuous** subtree on the t-layer; such subtree is consequently collapsible and it can be represented as a single t-node, and (iii) the t-layer also does not feature nodes for auxiliary words,[7] which significantly simplifies the annotation process.

All the multiword expressions in a given sentence are stored in the attribute mwes of the root node of the tectogrammatical tree. The attribute mwes is a list, whose members represent MWEs in the tree. Each MWE contains an ID, a basic-form, a type and a list of identifiers of t-nodes that are a part of the MWE. The type of a MWE can have one of the following values:

- lexeme,[8]
- person (a name of a person or an animal),
- institution,
- object (e.g. a name of a book, a unit of measurement, a biological name of a plant or an animal),
- location,
- address,
- time,
- biblio (a bibliographic entry),
- foreign (a foreign expression),
- number (esp. a numerical range).

[4] See http://arbuckle.sfs.uni-tuebingen.de/en_tuebadz.shtml and (Osenova and Kolkovska, 2002).

[5] http://ufal.mff.cuni.cz/lexemann/mwe/

[6] An easier annotation of *single-word named entities* is left for future versions of PDT.

[7] Auxiliary words are instead accessible through attributes and links.

[8] "Conjunction of the lexical form and the individual meaning" (Filipec, 1994). Compare also "lexical unit"

Examples:

(1) *Prezident Havel by měl 15. července*[*] *na Pražském hradě*[†] *jmenovat třináct soudců Ústavního soudu*[‡] .
transl.: *President Havel is expected to appoint thirteen judges of the Constitutional Court on 15th of July at the Prague Castle.*

[*] *'on 15th of July'* – date, basic-form: "15. červenec" (nominative case)
[†] *'at Prague Castle'* (locative case) – location, basic-form: "Pražský hrad" (nominative case)
[‡] *'[of] Constitutional Court'* (genitive) – institution, basic-form: "Ústavní soud" (nominative)

(2) *Funkce ústavního soudce*[*] *je neslučitelná s členstvím v politických stranách*[†] .
transl.: *The role of a constitutional judge is incompatible with political party membership.*

[*] *'[of a] constitutional judge'* (genitive) – lexeme, basic-form: "ústavní soudce" (nominative)
[†] *'in political parties'* (locative, plural) – lexeme, basic-form: "politická strana" (nominative, singular)

4.1 MWE display and search

There are two modes of viewing the MWEs in TrEd: they can be seen either as coloured groups of t-nodes in a tectogrammatical tree (see Figure 3C), or they can be collapsed into a single node (see Figure 3B). When collapsed, children of the members of a MWE become children of the MWE node itself as we can see with *deficit* and its parent *miliarda* in Figure 3. In the "node group" mode the groups are drawn in different colours representing different types of MWEs. In Figure 3 (B) and (C) there is a subtree (*'a 33 billion budget deficit'*) with 2 MWEs (NE *'33 billion'* and a lexeme *'budget deficit'*)[9] in a compact collapsed form (B) and in a coloured group-view (C). Orange colour represents a multiword number and pink represents a lexeme in (C).

4.2 Annotation procedure

We annotated all occurrences of MWEs (including named entities, see below) at the tectogrammatical layer of PDT. A large part of the data was annotated in parallel. A table below shows how much data was annotated by 1, 2, or 3 annotators in parallel, compared to the size of the t-layer.

	number of annotators			$\sum$ annotated	$\sum$ parallel
	one	two	three	(100% of PDT t-layer)	(in % of PDT t-layer)
t-files	1,288	1,412	465	3,165	59
t-nodes	248,448	343,834	82,683	674,965	63

Table 1: Parallel annotation of data

The data produced by individual annotators is not part of PDT 2.5, but it is freely available at the project web page.[10] For the present release it was used to produce gold standard MWE annotation in the following manner: if the annotators agreed, the MWE was kept as gold. Disagreement was decided as follows:

of Cruse (1986).

[9] Multiword numeric entity is always annotated. The reason for annotation of "budget deficit" (translatable, as you can see) is non-compositionality: it is different than, e.g., "oxygen deficit", because there is no budget shortage, but shortage of money in the budget (or even an income shortage comparing to costs).

[10] http://ufal.mff.cuni.cz/lexemann/mwe/

In case a MWE was recognised by only one annotator, we kept it, since a test had shown that it was much more common for an annotator to miss a MWE than to annotate a false MWE. In case one annotator annotated a subset of the other's MWE, we kept the larger MWE. On the other hand, when one annotator chose several small MWEs covering the other's larger MWE, the smaller ones were kept.[11] The cases when the annotators created intersecting MWEs were judged by a third annotator, as were the cases when one annotator identified several subsets of the other's MWE but the subsets didn't cover the full extent of the large MWE.

5 The grammateme typgroup representing the pair/group meaning

In Czech, nouns typically have two sets of forms according to the grammatical category of number: singular forms and plural forms. Forms of the former set are used to denote a single entity (singularity meaning), plural forms express, in general, more than one entity (plurality meaning). Within the theoretical linguistic framework of FGD as well as in the annotation scenarios of PDT 2.0 and PDT 2.5, the semantic opposition of singularity and plurality is represented by the values sg vs. pl of the grammateme number; grammatemes are attributes of nodes of the tectogrammatical tree, which capture the semantically relevant morphological categories.

In addition to the existence of nouns accompanied in the lexicon with the feature "singulare tantum", which blocks the semantic opposition of sg vs. pl (e.g. *kamení 'stones'*) and "plurale tantum", where sg and pl are expressed by the same form (e.g. dveře 'door/doors'), there are nouns in Czech that have both singular and plural forms but their plural forms are used to refer to a pair or to a typical group of entities rather than to a plurality of them. For instance, the plural *ruce 'arms'* denotes a pair or several pairs of arms rather than several upper limbs, the form *boty 'shoes'* denotes a pair or several pairs of shoes, the form *klíče 'keys'* means a bundle or more bundles of keys. The meaning is referred to as the "pair/group meaning" in the present paper.

As the pair/group meaning is compatible with most Czech concrete nouns and it manifests in some peculiarities as to the compatibility of the particular nouns with numerals,[12] we proposed to treat the pair/group meaning as a grammaticalized meaning constituting a new grammatical category of Czech nouns (Panevová and Ševčíková, 2011).

5.1 Grammateme typgroup

For the purpose of including the pair/group meaning into the tectogrammatical annotation of PDT 2.5, a new grammateme typgroup was added to the existing set of 15 grammatemes used in PDT 2.0. For the typgroup grammateme, three values were defined: group for the pair/group meaning, single for the meaning of single entities, and nr ("not recognised") for unresolvable cases.

The pair/group meaning is closely related to the meanings of the number category. In connection with the annotation of the grammateme typgroup, values of the grammateme

[11] Because it is typically a case like the composer and a symphony annotated together as a concert performance.

[12] The counting of pair/group nouns requires using a set numeral instead of a cardinal one. This is a strong argument in favor of considering the pair/group meaning a grammatical category. Cf. the set numeral *dvoje 'two sets'* in the example *Máme dvoje sklenice – na bílé a červené víno. 'We have two-sets of glasses – for the white and for the red wine'* vs. the cardinal numeral *dvě 'two'* if counting single entities in the sentence *Postavil na stůl dvě sklenice. 'He put two glasses on the table'*.

number as implemented in the PDT 2.0 data had to be changed in some cases. For instance, if the plural form was identified as denoting a pair or group, the value pl (assigned to the node representing this form in PDT 2.0) was changed to sg in the PDT 2.5 data.

5.2 Manual annotation of the typgroup grammateme with selected nouns

In the PDT 2.5, the grammateme typgroup was assigned semi-automatically with all nouns; the manual annotation concerned the nouns for which a higher frequency of the pair/group meaning was expected, the rest of the nouns was assigned a value of the typgroup grammateme automatically.

Occurrences for manual assignment were selected on the basis of a list of tectogrammatical lemmas (t-lemmas) of prototypical pair/group nouns. Nouns which co-occur with a set numeral in the PDT 2.0 and in the SYN2005 (ÚČNK, 2005) corpus data were analyzed as good candidates for this list. The list was further enriched using grammar books and theoretical studies on number in Czech as well as linguistic introspection. In the resulting list, 141 Czech nouns were involved, only 67 of them with 618 instances of plural forms were found in the PDT 2.5 data. Most of the nouns belong to one of the following groups:

- nouns denoting body parts occurring in pairs or groups (*uši 'ears'*, *prsty 'fingers'*, *vlasy 'hair'*),
- nouns denoting clothes and accessories for these body parts (*náušnice 'earrings'*, *rukavice 'gloves'*),
- nouns denoting family members such as *rodiče 'parents'*, *sourozenci 'siblings'*, and
- nouns denoting objects of everyday use and foods sold or used in typical amounts (*klíče 'keys'*, *sirky 'matches'*, *cigarety 'cigarettes'*, *sušenky 'biscuits'*).

The plural forms to be annotated were extracted from the data together with a short, both preceding and following context. In order to make the annotation task as simple as possible, the annotators did not specify the values of both the typgroup and number grammatemes, but they were asked to choose one of the annotation choices 1 to 6; the correspondences between the annotation choices and the grammateme values are described in Table 2. All 31 files were annotated manually by two annotators (native Czech speakers) in parallel during four months, the annotation was preceded by a short training period. The language intuition of native speakers played a crucial role in the annotation process. The annotators agreed on 464 (75.1%) out of 618 instances annotated, with a Cohen's Kappa score of 0.67. After the manual parallel annotation, instances of disagreement were adjudicated by a third annotator and the instances on which annotators agreed were revised in order to check the correctness and consistency of the annotation. The frequency of the choices in the revised annotation is summarized in the last column of Table 2.

5.3 Automatic assignment of the typgroup grammateme to the remaining nouns

Nouns which were not in the list (and consequently in the manual annotation) were assigned a value of the typgroup grammateme fully automatically. A simple, two-step procedure was provided for the automatic annotation: in the first step, nouns accompanied

Annotation choice	Grammateme values typgroup	number	# of instances (percentage)	
1 - plurality	single	pl	133	(21.5%)
2 - one pair/group	group	**sg**	230	(37.2%)
3 - several pairs/groups	group	pl	30	(4.9%)
4 - one pair/group or several pairs/groups	group	**nr**	154	(24.9%)
5 - cannot be resolved	nr	**nr**	70	(11.3%)
6 - —	*to indicate a mistake*		1	(0.2%)

Table 2: Manual annotation: annotation choices and corresponding combinations of the values of the grammatemes typgroup and number and their frequency in the manually annotated data. The number values marked in bold were changed from the pl value (as available in the PDT 2.0 annotation) to the marked value, influenced by the annotation of the pair/group meaning.

Grammateme values typgroup	number	# of instances
single	sg	185086
single	pl	59912
single	nr	10232
group	sg	237
group	pl	35
group	nr	153
nr	nr	66

Table 3: Combinations of values of the grammatemes typgroup and number and their frequency in the PDT 2.5 data.

with a set numeral *jedny 'one-pair/group'* (except for pluralia tantum) were assigned the value group of the grammateme typgroup and the value of the grammateme number was changed to sg in this connection; if the noun collocated with a set numeral of a higher numeric value (*dvoje 'two-pairs/groups-of'*, *troje 'three-pairs/groups-of'* etc.), the value group was filled in the grammateme typgroup whereas the grammateme number remained unchanged (i.e. pl). Secondly, all the other nouns were assigned the value single in the grammateme typgroup, the value of the grammateme number was not changed in these cases, compared to the original (PDT 2.0) annotation.

Combinations of the values of the grammatemes typgroup and number in the full PDT 2.5 data and their frequency is displayed in Table 3.

6 Automatic annotation of clause segmentation

Analytical trees in PDT 2.5 have been enriched with annotation of clause segmentation. Clauses are grammatical units out of which complex sentences are built. A clause typically corresponds to a single proposition expressed by a finite verb and all its arguments and modifiers (unless they constitute clauses of their own).[13] Annotation of clauses can be

[13] Given that Czech is a pro-drop language (pronouns in subject positions are often dropped, since their gender, number and person values are already expressed by verb inflection), this definition based on finite verbs matches

used for training clause boundary identifiers, which are generally supposed to be helpful in a number of NLP tasks such as parsing (since most dependencies do not cross clause boundaries), text summarisation (for instance, relative clauses might contain information of lesser importance and thus are more likely to be removable), machine translation (most reordering patterns are to be applied inside clauses), and speech applications (clause boundaries often imply prosodic boundaries).

We believed that clause boundaries could be identified automatically with very high reliability, since gold-standard morphological and more importantly, analytical annotation had already been available. Therefore clause boundaries were annotated manually only in a small portion of the PDT data and this annotation was used for developing a rule-based clause-identification procedure. To make the annotation consistent across the whole data, all the clause annotation distributed in PDT 2.5 was generated by this procedure; the original manually annotated samples are not included in PDT 2.5.

6.1 Basic conventions for clause representation

Technically, clause boundaries are represented by the dedicated attribute `clause_number` added to analytical nodes. If two analytical nodes in a tree share the same non-zero value of this attribute, then they belong to the same clause. Zero value of this attribute is reserved for boundary tokens, i.e. tokens that are located on the boundary of two clauses and cannot be unequivocally assigned to either of these clauses. Boundary tokens are typically various types of punctuation marks or coordinating conjunctions.[14]

Coindexing by the dedicated attribute is rendered by colours in the PDT 2.5 clause segmentation samples below:

(3) *U sochy básníka seděl vlasatý mladík a*⋆ *hrál Vysockého písně.*†
transl.: *There was a hairy guy sitting at a statue of a poet playing Vysockij's songs.*
⋆ Clause boundary is formed by the coordinating conjunction between the two clauses.
† Sentence boundary is manifested by the final punctuation.

(4) *Pokud jde o kupní smlouvu a*⋆ *všechny náležitosti s ní spojené,*† *musí si to zařídit a*⋆ *zaplatit strany samy.*
transl.: *Considering the buying contract and all related requirements, it has to be set and paid by contracting parties themselves.*
⋆ The coordinating conjunction that joins *sentence constituents* belongs to the clause.
† Clause boundary manifested by the punctuation symbol.

(5) *Lidé na nás tehdy chodili, aby*⋆ *se odreagovali od přítomného režimu.*
transl.: *People in those days used to attend our sessions, so that they could lay off the present government.*
⋆ The subordinating conjunction belongs to the subordinate clause.

(6) *Posunovač, který prý vstoupil do kolejiště, aniž se rozhlédl, je nyní v nemocnici*⋆ .
transl.: *The switchman that is said to enter the railyard without even looking around is in the hospital.*
⋆ The matrix clause is split into two parts by the embedded relative clause, which is further modified by the dependent clause.

(for Czech) the traditional notion of a clause as a group of words having a subject and a predicate.

[14] Note that subordinating conjunctions are systematically annotated as part of the respective dependent clause. The reason for this decision lies in their linguistic properties: subordinating conjunctions in Czech make an integral part of the dependent clause and, if omitted, the clause might become ungrammatical.

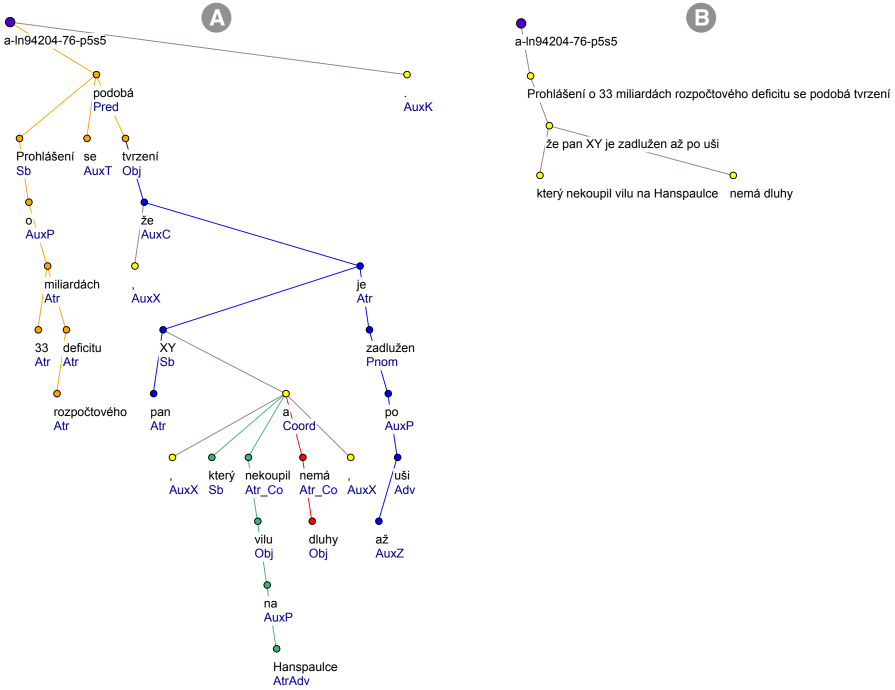

Figure 1: Two ways to visualise sentence segmentation in TrEd. Translation in Figure 3.

Clause segmentation can be comfortably visualised in the tree editor TrEd (see Figure 1) in two styles: either in full (unfolded) or in folded trees. In the former case, the tree topology is displayed as usual and clause segmentation is signalled by node and edge colours (see Figure 1A). In the latter case, the set of nodes belonging to one clause collapses to a single node which represents the whole clause (see Figure 1B).

6.2 Annotation procedure for clause segmentation

The automatic clause identification procedure can be outlined as follows:

1. Clause seeds are identified. Every occurrence of a finite verb form (the POS tag identifies finiteness reliably) is marked as a distinct clause seed.
2. Seeds forming a compound verb are joined together. Seeds with the analytical function of an auxiliary verb (AuxV) cannot constitute a clause on their own.
3. The tree is recursively traversed (post-order) and each coordination head is temporarily added to the clause of its rightmost member that already belongs to a clause.
4. Clause completion step. The tree is traversed recursively and the children that do not yet belong to any clause are typically added to the clause of the parent node (special handling of coordinations is needed here), or to their nearest left or right sibling that already constitutes a clause.
5. All potential boundary nodes are excluded from the clauses and their clause membership is re-estimated. The criteria is based mostly on the linear order of tokens but attention is also paid to the tree structure.

```
t-root [
    atree.rf $aroot,
    2+x member mwes [  ],
    descendant t-node [
        gram/typgroup = "group"
] ];

a-root $aroot := [
    1+x descendant a-node [
        clause_number = 3
] ];
```

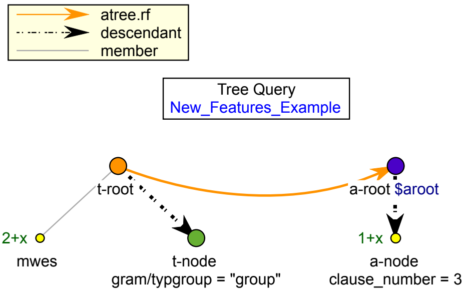

Figure 2: The PML-TQ query used to obtain Figure 3 in a textual and in a graphical form. It searches for a sentence with at least three clauses, two MWEs, and one word with the pair/group meaning.

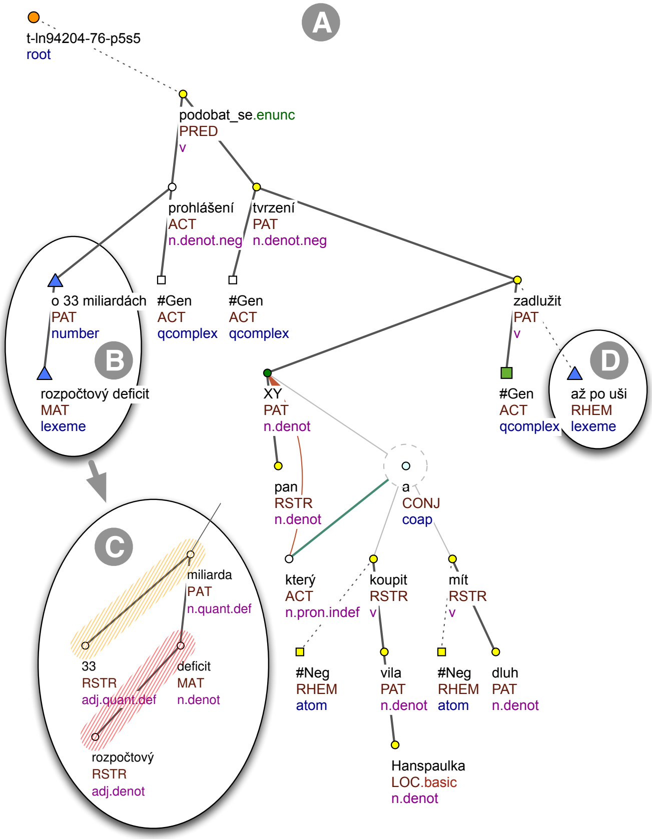

Figure 3: An example tree showing all the new features of PDT 2.5 as shown in TrEd: *Prohlášení o 33 miliardách rozpočtového deficitu se podobá tvrzení*, *že pan XY*, *který nekoupil vilu na Hanspaulce* a *nemá dluhy*, *je zadlužen až po uši*⋆.
transl.: *The statement on 33 billion [of] budget deficit is similar to the statement that Mr. XY*, *who didn't buy a luxury villa* and *doesn't owe to anyone*, *is up to his ears*⋆ *in debt*.

⋆ pair meaning

6.3 Evaluation and application on PDT 2.5 data

For the purpose of evaluation, we obtained data from a pilot project on annotating sentence structure, whose methodology is thoroughly formulated in (Lopatková et al., 2012). The project provided us with a valuable collection of 2505 manually annotated sentences. We use these gold-standard sentences for evaluation of our automatic clause-identification procedure. Despite being a subset of the PDT data, the manually annotated sentences are not shipped with PDT 2.5 due to reasons explained below; for clause boundaries, all the PDT 2.5 data are annotated automatically.

Mostly because of the different scope of the project, we have adopted a slightly different annotation scheme. Let us summarise the original concepts and emphasise the differences.

The theory behind the pilot project (Lopatková et al., 2012) is centred on *segmentation charts*. Prior to manual annotation, tokenised and morphologically annotated sentences are automatically split into individual *segments*. All punctuation marks and coordinating conjunctions serve as *segment boundaries*. A single clause then consists of one or more segments. This scheme is viable given the very strict rules for punctuation in Czech – there must be some kind of a boundary between two finite verb forms, be it a sentence boundary, punctuation or a conjunction. The task of the annotators was to identify individual clauses, i.e. to group the segments forming a single clause, and to assign the appropriate level of embedding, thus allowing the distinction between coordination and super- or subordination. The usage of analytical layer during the annotation was intentionally quite limited. Only the analytical functions of tokens were used to help the annotators decide on the correct level of embedding and to disambiguate if more readings of a particular sentence were possible.

Unlike the manual annotation, the automatic clause-identification procedure does not rely on the boundary segments and extensively uses analytical trees. There are three key differences in the annotation rules:

(a) The automatic procedure does not attempt to assign levels of embedding, as the interclausal relations are explicitly captured in the analytical tree.
(b) Segment boundaries delimiting segments within the scope of a single clause are annotated as part of the clause, so that the distinction between coordination of sentence members and coordination of clauses is made obvious.
(c) A parenthetical expression is not considered a separate clause unless it contains a finite verb form.

Especially the last rule created the need of further post-processing of the gold-standard data, to make automatic evaluation possible. During the post-processing, parenthetical expressions were automatically merged with their surrounding clauses. In the original manually annotated data there are 2,505 sentences divided into 5,311 clauses. After post-processing, the number of clauses drops to 4,948.

The evaluation was performed on the basis of clauses using standard precision, recall, and f-measure metrics, reaching values 0.973, 0.978, and 0.975 respectively.[15] This confirmed the initial hypothesis that a highly reliable segmentation can be induced from the already available dependency annotations. As for the few remaining wrongly recognized

[15] Each automatically recognised clause was considered correct if and only if there is a clause in the manually annotated data spanning the very same set of nodes.

boundaries, the error analysis has shown that they have no single dominating cause that could be easily fixed. Such sentences are often difficult to annotate even for a human, for instance because of elipsis or intricate interplay of hypotactic and paratactic structures.

The automatic clause-identification procedure was used to annotate all sentences provided with gold-standard analytical trees in PDT 2.5, which amounts to 87,913 sentences. The procedure identified 153,434 clauses.

7 Conclusion

The Prague Dependency Treebank has been used as a model for several other treebanks, showing that both the general linguistic model of FGD and the technical realisation of PDT using PML are flexible and generic. They are not limited to a particular language, or a language family. By now there are at least five treebanks[16] annotated in the "PDT style".[17]

We have shown that PDT is exceptional in the richness of the information it provides. PML fits this richness well and thus all the PML-based tools such as the `TrEd` editor and the PML-TQ tree-query language (Pajas and Štěpánek, 2009) can be seamlessly used with PDT. PDT 2.5 is the most complex release of PDT to date. It is an intermediate step on the way to PDT 3.0, which will add even more annotation (discourse and extended anaphora, for example). It also contains corrections of more than 10,000 technical and annotation errors found in the previous release.

In Sections 4 through 6 we have presented major new features of PDT 2.5: what they are, how they were obtained, and what is the resulting quality and reliability. In Figures 2 and 3 there is an example of a complex query involving all of these features using the PML-TQ search tool and a result found together with its visualizations.

PDT 2.5 and all of the tools mentioned above are freely available (not just) for research purposes under standard, permissive licenses at `http://ufal.mff.cuni.cz/pdt2.5` and in the LINDAT-Clarin repository at `http://lindat.cz`. The Prague Dependency Treebank 2.5 itself has a citable persistent identifier `http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8`.

Acknowledgement

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). The work on this project was supported by the grants P406/2010/0875, P406/12/P175, and P406/10/P193 of the Grant Agency of the Czech Republic.

[16] Prague Dependency Treebank, Prague Arabic Dependency Treebank, Prague Czech-English Dependency Treebank, Index Thomisticus Treebank, Slovak Treebank (analytic layer).

[17] The markup language PML is even more generic. In fact, it has no direct connection to FGD and can be used to represent treebanks in any linguistic formalism (phrase-based or dependency, or any other variety based on trees in the basic graph-theory sense). Currently almost thirty treebanks are available in PML and can be queried online using our PML-TQ server. We have yet to meet a treebank that can't be converted to PML.

References

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 89–96, Morristown, NJ, USA. Association for Computational Linguistics.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

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

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and Ševčíková Razímová, M. (2006). Prague Dependency Treebank 2.0. CD-ROM. Linguistic Data Consortium.

Ide, N. and Suderman, K. (2009). Bridging the gaps: Interoperability for GrAF, GATE, and UIMA. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP ’09, pages 27–34, Suntec, Singapore. Association for Computational Linguistics.

Lopatková, M., Homola, P., and Klyueva, N. (2012). Annotation of sentence structure: Capturing the relationship between clauses in Czech sentences. *Language Resources and Evaluation*, 46(1):25–36.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mel’čuk, I. A. and Polguère, A. (1987). A formal lexicon in The Meaning-Text Theory (or how to do lexica with words). *Computational Linguistics*, 13(3-4):261–275.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The NomBank Project: An Interim Report. In Meyers, A., editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA. Association for Computational Linguistics.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., and Žabokrtský, Z. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep.

Osenova, P. and Kolkovska, S. (2002). Combining the named-entity recognition task and NP chunking strategy for robust pre-processing. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*.

Pajas, P. and Štěpánek, J. (2006). XML-based representation of multi-layered annotation in the PDT 2.0. In Hinrichs, R. E., Ide, N., Palmer, M., and Pustejovsky, J., editors, *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47, Genova, Italy.

Pajas, P. and Štěpánek, J. (2008). Recent Advances in a Feature-Rich Framework for Treebank Annotation. In Scott, D. and Uszkoreit, H., editors, *The 22nd International Conference on Computational Linguistics - Proceedings of the Conference*, volume 2, pages 673–680, Manchester, UK. The Coling 2008 Organizing Committee.

Pajas, P. and Štěpánek, J. (2009). System for querying syntactically annotated corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, Singapore. Association for Computational Linguistics.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105.

Panevová, J. and Ševčíková, M. (2011). Jak se počítají substantiva v češtině: poznámky ke kategorii čísla. *Slovo a slovesnost*, 72:163–176.

Pecina, P. (2009). *Lexical Association Measures: Collocation Extraction*, volume 4 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Popel, M. and Žabokrtský, Z. (2010). TectoMT: Modular NLP framework. In Loftsson, H., Rögnvaldsson, E., and Helgadottir, S., editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Third International Conference, CICLing*.

Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Praha.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/Reidel Publ. Comp., Praha/Dordrecht.

ÚČNK – Institute of the Czech National Corpus (2005). Czech National Corpus - SYN2005. `http://www.korpus.cz`. Prague.

Weischedel, R. and Brunstein, A. (2005). BBN Pronoun Coreference and Entity Type Corpus. CD-ROM. Linguistic Data Consortium.

Xue, N., Jiang, Z., Zhong, X., Palmer, M., Xia, F., Chiou, F.-D., and Chang, M. (2010). Chinese treebank 7.0. CD-ROM. Linguistic Data Consortium.