Documentation

Prague Dependency Treebank







Documentation on this page applies only to the innovations since PDT 2.0. Its comprehensive documentation can be found here. The new features of PDT 2.5 are:

- Annotation of Multiword Expressions
- Pair/Group Meaning
- Clause Segmentation
- Corrections of the Original Data

Annotation of Multiword Expressions

All the multiword expressions in a given sentence are stored in an attribute `mwes` of a root node of the tectogrammatical tree. The attribute `mwes` is a lists, 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.

A MWE can be either a multiword lexeme (phraseme, a light verb construction, etc.), or a type of a named entity. For named entities we specify its kind). The MWE `type` can thus have following values:

- *"lexeme"* – a multiword lexeme
- *"person"* – a name of a person or an animal
- *"institution"* – an institution name
- *"location"* – a geographical location
- *"object"* – a name of a book, a unit of measurement, a biological name of a plant or an animal
- *"address"*
- *"time"* – date and time expressions
- *"biblio"* – a bibliographic entry
- *"foreign"* – a foreign expression
- *"number"* – a numerical value, usually a range

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, or they can be collapsed into a single node. When collapsed, children of the members of a MWE become children of the MWE node itself. In the "node group" mode the groups are drawn in different colour, representing different types of MWEs.

Annotation Procedure

We annotated all occurrences of MWEs (including named entities, see below) in the tectogrammatical layer of PDT 2.0. A large part of 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 PDT (t-data).

Annotated data					
parallel annot.	**1**	**2**	**3**	**PDT**	**2+3/PDT**
t-files	1,288	1,412	465	3,165	59%
t-nodes	248,448	343,834	82,683	674,965	63%

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

- In case a MWE was recognised by only one annotator, we kept it, since test had shown that it was much more common for an annotator to miss a MWE, then 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 other's larger MWE, smaller ones were kept.
- The cases when the annotators created intersecting MWEs were judged by a third annotator.

- 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, were also judged manually by a third annotator.

Examples

Prezident Havel by měl 15. července* na Pražském hradě jmenovat třináct soudců Ústavního soudu***.**

* – "15. 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)

Funkce ústavního soudce* je neslučitelná s členstvím v politických stranách.**

* – "[of] constitutional judge" (genitive) – *lexeme*, *basic_form*"ústavní soudce" (nominative)

** – "in political parties" (instrumental, plural) – *lexeme*, *basic_form* "politická strana" (nominative, singular)

Pair/Group Meaning

By the values of the grammateme *typgroup*, the semantic opposition of the pair/group meaning vs. meaning of single entities is represented (values *group* vs. *single*, respectively; the third value *nr* was used for ambiguous cases). In Czech, nouns such as *ruce* 'hands, arms', *boty* 'shoes' or *klíče* 'keys' refer with their plural forms rather to a pair or to a typical group even more often than to a larger amount of single entities; cf. the plural form *ruce* 'hands, arms' denotes a pair or several pairs of arms rather than several upper limbs, the form *boty* 'shoes' usually denotes a pair or several pairs of shoes, the form *klíče* 'keys' means a bundle or more bundles of keys. Since pairs/groups can be referred to with most Czech concrete nouns and since it manifests in some peculiarities as to the compatibility of these nouns with numerals (if expressing pairs/groups, the noun is compatible with set numerals only, whereas when referring to single entities, a cardinal numeral is used; cf. *dvoje boty* 'two-pairs-of shoes' vs. *dvě boty* 'two shoes'), the pair/group meaning is considered as a grammaticalized meaning of nouns in Czech.

The pair/group meaning is expressed by formally unmarked plural forms of nouns. Since the plural form is disambiguated either by the numeral, which however co-occurs rather rarely in the data, or on the basis of context or knowledge of the world, most of plural forms of nouns were candidates for the manual disambiguation. Nevertheless, since a rather low frequency of the pair/group meaning was expected on the background of a pilot annotation experiment, only plural forms of those nouns were manually annotated for which the pair/group meaning was considered as prototypical, in order to make the annotation as efficient as possible. The following groups of nouns were expected to be prototypical pair/group nouns:

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

Annotation Procedure

In the PDT 2.5, the grammateme *typgroup* was assigned semi-automatically with all denominating semantic nouns (nodes with *sempos=n.denot|n.denot.neg*). First of all, occurrences for manual assignment were selected on the basis of a list of tectogrammatical lemmas (t-lemmas). In the list of prototypical pair/group nouns to be annotated, nouns were involved which co-occur with a set numeral in the PDT 2.0 and in the SYN2005 data, the list was further enriched using grammar books and theoretical studies on number in Czech as well as linguistic introspection. For the t-lemmas from the resulting list, more than 600 instances of plural forms were found in the PDT 2.5 data (most of the instances belong to the following t-lemmas: *oko* 'eye', *rodič* 'parent', *ruka* 'hand, arm', *bota* 'shoe').

Manual annotation of these instances was carried out by two annotators in parallel, with an inter-annotator agreement of 75.1% of the annotated instances (Cohen's kappa score 0.67). After the manual 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 pair/group meaning is closely connected with the grammatical category of number of nouns; the category of number is constituted with the opposition of singular and plural in Czech. In connection with the manual annotation of the pair/group meaning, the values of the grammateme *number* (values *sg*, *pl*, and *nr*) were changed in comparison to the original (PDT 2.0) annotation in the following way: if a plural form of a noun was identified as expressing a single pair/group (*typgroup=group*), the value of the grammateme *number* was set to *sg*; if more pairs/groups were denoted (*typgroup=group*), the value of the grammateme *number* did not change (remained *pl*); if the annotators cannot decide between a single pair/group and several of them (*typgroup=group*), the value *nr* was filled in the grammateme *number*.

With denominating semantic nouns that were not involved in the manual annotation, the grammateme *typgroup* was assigned automatically. A simple, two-step "algorithm" was provided for the automatic annotation: in the first step, nouns accompanied 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.

In the data, the following combinations of the values of the grammatemes *number* and *typgroup* occur:

- *sg.group* – the meaning of one pair/group, expressed by a plural form of nouns

- *pl.group* – the meaning of more than one pair/group, expressed by a plural form of nouns
- *nr.group* – one or more pairs/groups are referred to, this meaning is expressed by a plural form of nouns
- *sg.single* – the meaning of one entity, expressed by a singular form of nouns
- *pl.single* – the meaning of more than one single entities, expressed by a plural form of nouns
- *nr.single* – nodes with which the number was not recognized (*number=nr*) were assigned the value *single* of the grammateme *typgroup* by default
- *nr.nr* – ambiguous occurrences were assigned this combination: neither the combination *sg.group*, nor *pl.group*, nor *pl.single* could be excluded (the combination *sg.single* is not to be considered under this combination!)

Examples

The values of the grammatemes *number* and *typgroup* are given in italics for each denominating semantic noun, nouns that were assigned the *typgroup* value manually are marked in bold:

1. Navlékla bych si dvoje **ponožky**.*pl.group* a hrála bych naboso, dokud by mi někdo nesehnal nějaké **boty**.*sg.group*.
   'I would put on two-pairs-of **socks**.*pl.group* and would play barefooted until somebody would get some **shoes**.*sg.group* for me.'
2. Pro něho připravila firma.*sg.single* Lotto.*sg.single* speciální **kopačky**.*nr.group*.
   'The Lotto.*sg.single* company.*sg.single* developed special **football boots**.*nr.group* for him.'
3. Sečíst pouhým okem.*sg.single* stranickou příslušnost.*sg.single* zvednutých **rukou**.*pl.single* bylo ve dvousetčlenné Poslanecké sněmovně.*sg.single* nemožné.
   'It was impossible to count up with the naked eye.*sg.single* the party affiliation.*sg.single* of the risen **hands**.*pl.single* in the two-hundred-member Chamber.*sg.single* of Deputies.'
4. ... je to také odpověď.*sg.single* na vzdělávací požadavky.*pl.single* **rodičů**.*nr.nr*, žáků.*pl.single*, ale i měnícího se trhu.*sg.single* práce.*sg.single*.
   '... it is an answer.*sg.single* to educational requirements.*pl.single* of the **parents**.*nr.nr*, pupils.*pl.single*, but of the changing job.*sg.single* market.*sg.single* as well.'
5. Obsah PCB.*nr.single* ve vepřovém a drůbežím mase je již minimální.
   'Content of PCB.*nr.single* in pork and poultry meat is already minimal.'

Clause Segmentation

Analytical trees in PDT 2.5 are 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). Annotation of clauses can be used for training clause boundary identifiers, which are supposed to be helpful in a number of NLP tasks such as parsing, information extraction, machine translation, and speech applications.

It was hoped that clause boundaries can be identified automatically with very high reliability if gold-standard morphological and especially analytical representations of a sentence are already available. Therefore clause boundaries were annotated manually only in a limited portion of the PDT data. Then the manual annotation was used for developing a rule-based clause-identification procedure, whose f-measure reaches 97.51%. To make the annotation consistent across all the data, all the clause annotation distributed in PDT 2.5 was generated by this procedure; the original manually annotated samples are not shipped with PDT 2.5.

Annotation Scheme

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 (tagged as `Z:`) or coordinating conjunctions (tagged as `J^`). Note that subordinating conjunctions (tagged as `J,`) 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 could become ungrammatical.

Visualization

Clause segmentation can be comfortably visualized in TrEd (see Figure 1). The new extension for viewing PDT 2.5 data offers two additional macros related to clause segmentation:

- **Toggle clause folding (f)** – When clause folding is switched on the analytical tree of a sentence displays its structure on the level of clauses. All nodes forming a single clause are collapsed into one node and the dependency relations between clauses become apparent.
- **Toggle clause coloring (c)** – When clause coloring is switched on the sentence string displayed above the analytical or tectogrammatical tree is rendered with each clause colored in a different color (actually there are only ten colors being reused in the rare cases where the clause count exceeds ten). When an uncollapsed analytical tree is displayed the same coloring is applied also to the nodes and edges of the tree.

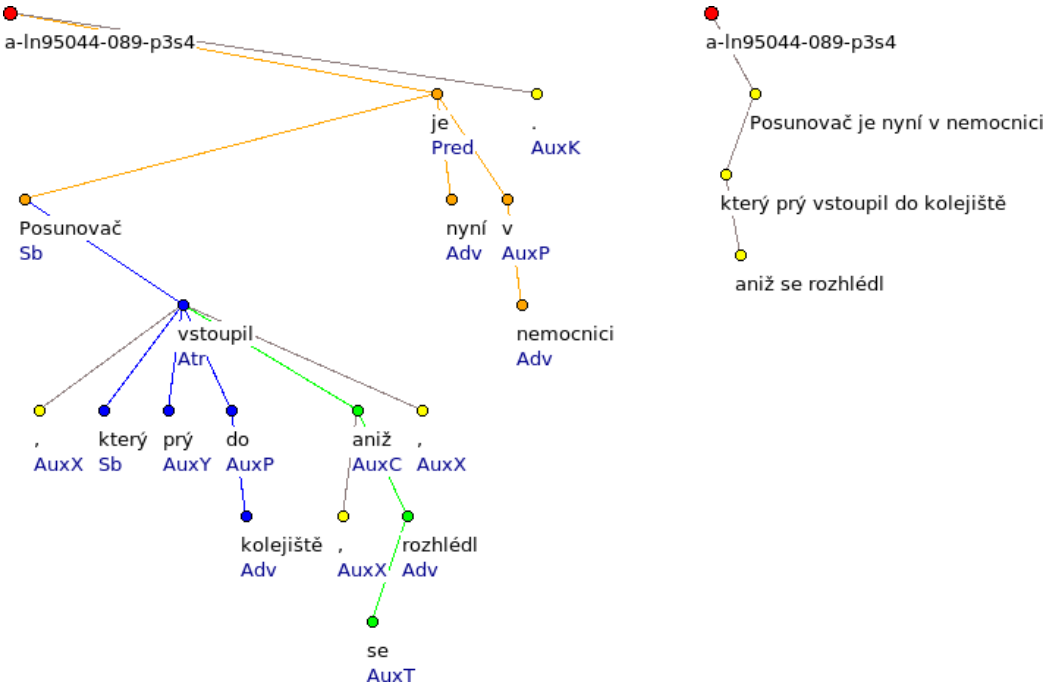

Figure 1: Sentence '*Posunovač, který prý vstoupil do kolejiště, aniž se rozhlédl, je nyní v nemocnici.*' represented by two trees: full (and colored) on the left side and with collapsed clauses on the right side.

Examples

U sochy básníka seděl vlasatý mladík a* hrál Vysockého písně.**

* – clause boundary, coordinating conjunction joining two clauses

** – final punctuation, sentence boundary

Pokud jde o kupní smlouvu a* všechny náležitosti s ní spojené, musí si to zařídit a* zaplatit strany samy.**

* – coordinating conjunctions joining sentence members within the scope of a single clause

** – clause boundary, punctuation

Lidé na nás tehdy chodili, aby* se odreagovali od přítomného režimu.

* – subordinating conjunction

Posunovač, který prý vstoupil do kolejiště, aniž se rozhlédl, je nyní v nemocnici*.

* – main clause split into two parts by an embedded relative clause (which is further modified by a dependent clause)

Annotation Procedure

The automatic clause-identification procedure can be outlined as follows:

- Clause seeds are identified. Every occurrence of a finite verb form is marked as a distinct clause seed.
- Seeds forming a compound verb are joined together. Seeds with the analytical function of an auxiliary verb (AuxV) cannot constitute a clause on its own.
- 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.
- Clause completion step. The tree is recursively traversed (pre-order) and each node is processed along with its children. Typically the children that do not yet belong to any clause are just added to the clause of the parent node. Coordinations however require a special handling. The undecided children are processed in the linear order and appended to the clause of the nearest left or right sibling that already constitutes a clause. The decision is based on the linear order of the parent node and the children. The clause membership of the parent node can also be adjusted in this step.
- 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.

Manually Annotated Data

We follow the concepts thoroughly formulated in Lopatková et al. (2011) and used in the pilot project of manual annotation of sentence structure. The project provided us with a valuable collection of 2505 sentences manually annotated with respect to the sentence structure. We use these gold-standard sentences for automatic evaluation of our automatic clause-identification procedure. Despite being a subset of PDT data, the manually annotated sentences are not shipped with PDT 2.5 and all the data is consistently annotated automatically.

Mostly because of the different scope of the project, we have adopted slightly different annotation guidelines. Let us briefly summarize the original concepts and emphasize the differences.

The theory behind the pilot project is centered on the so called *segmentation charts*. Prior to manual annotation, tokenized 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 conjunction. The task of the annotators was to identify individual clauses, i.e. to group the segments forming a single clause, and to assign an 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.

As opposed to the manual annotation the automatic clause-identification procedure does not rely on the boundary segments and extensively uses the analytical tree of the sentence. There are three key differences in the annotation rules:

- The automatic procedure does not attempt to assign levels of embedding. The inter-clausal relations are explicitly captured in the analytical tree.
- 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.
- Parenthetical expression is not considered a separate clause unless it contains a finite verb form.

Especially the last-named rule raised 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. The following tables present basic statistics of the original and post-processed manually annotated data.

	Original	Post-processed
Sentence count	2,505	2,505
Clause count	5,311	4,948

Post-processed data statistics	
Number of clauses	Number of sentences
1	1,000
2	877
3	416
4	146
5	44
6	16
7	3
8	2
9	1

Evaluation

For the purpose of evaluation we use the post-processed manually annotated data. The evaluation is performed on the clause basis using standard precision and recall metrics. Each automatically recognized clause is evaluated as either *correct* or *incorrect*. Clause is considered correct if and only if there is a clause in the manually annotated data consisting of the very same set of nodes, otherwise it is considered incorrect. Based on the number of correct clauses **C**, number of incorrect clauses **I** and number of clauses in the gold-standard data **G**, we define:

- precision **P = C/(C+I)**
- recall **R = C/G**
- f-measure **F = 2*P*R/(P+R)**

Given the statistics of the manually annotated data we can compute baseline for our automatic procedure. If a baseline procedure did not attempt to recognize clauses in complex sentences, all sentences would be annotated as consisting of just one clause. In such case the total number of annotated clauses would equal to the number of sentences in the training data, and the amount of correctly recognized clauses would equal to the number of simple sentences among them. The following table gives comparison of our results and the hypothetical baseline.

	Baseline	Results
Precision	39.92%	97.25%
Recall	20.21%	97.78%
F-measure	26.83%	97.51%

PDT 2.5 Annotation

The automatic clause-identification procedure was used to annotate all the sentences provided with gold-standard analytical trees, which amounts to 87,913 sentences. Several new phenomena not seen in the sample data were encountered during this annotation that led to further improvements of the automatic procedure. When looking for possible annotation errors the following checks have proved useful.

- Any place in the data where transition between two clauses happens without an intermediate boundary token is suspicious.
- A boundary token appearing inside a single clause is an error.
- A boundary token with morphological tag different than `Z:` or `J^` is suspicious.

Statistics

The PDT 2.5 data provides clause segmentation for 87,913 sentences formed by a total number of 153,434 clauses. We estimated relative sentence counts of two kinds: see Figure 2 for clause counts per sentence and Figure 3 for the most common sentence structure *patterns*.

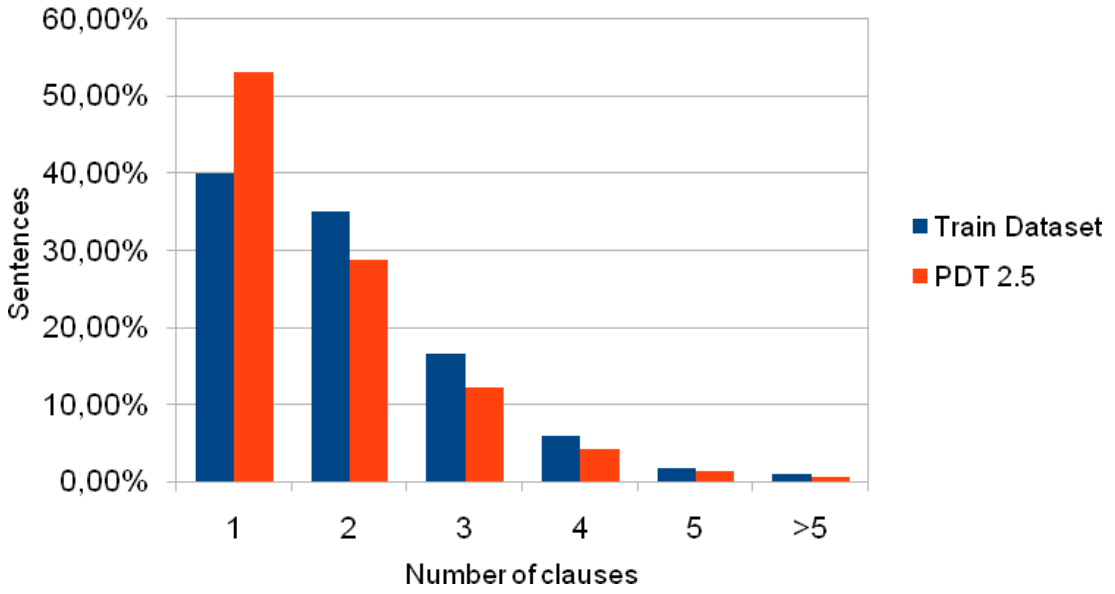

Figure 2

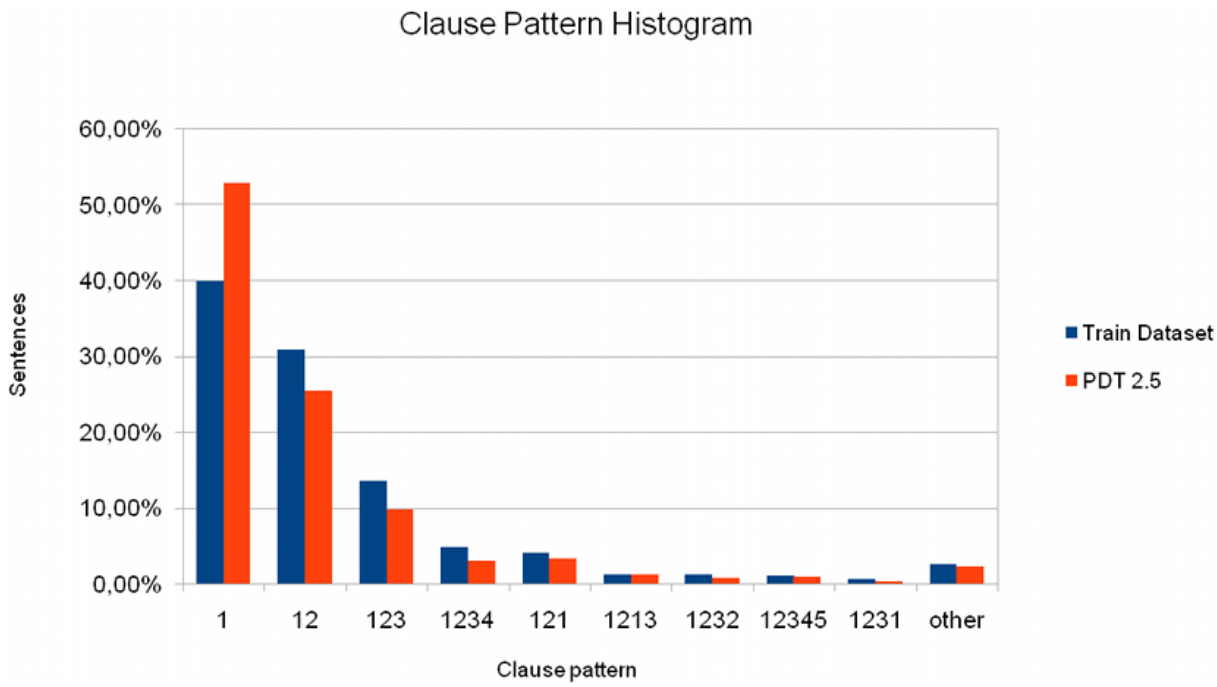

Figure 3: For the sake of brevity, clauses are numbered by single digits. For example, the pattern "12" stands for a complex sentence formed by two clauses, the pattern "121" also represents a two-clause sentence but with the second clause embedded, etc.

Corrections of the Original Data

The following changes were made *manually* to the original PDT 2.0 data:

w-layer

- Whitespace around left parenthesis fixed.

m-layer

- The morphological tags of the following words were fixed to agree with their analytical functions: až, co, dál, dále, daleko.
- Tag for the word form "budoucnu" was fixed.
- Many plural female first names were in fact annotation errors.
- Some surnames homonymous to other words were not recognised by the original morphological analysis: "Čermák", "Pešek", "Homolka", "Hromada", "Zeman".
- The two possible analyses of the word "druhý" were merged.
- "ES" is not a form of the word "eso", but an abbreviation of "European Unions".
- The analysis of the word "International" was unified.
- Annotation of abbreviated or short name parts fixed (like "J." or "O'"), including agreement in gender.
- Term types fixed for firstname and surname lemmas (`_Y` and `_S`).
- Annotation of "P. O. Box" fixed.
- Errors in encoding fixed ("ˇ" instead of double quotes, a dash, or "§").
- Annotation of "s. r. o." and "a. s." unified.
- The town name part "Králové" fixed.

a-layer

- Constructions using "co" fixed to accord with the m-layer.

t-layer

- All changes to lemmata on the m-layer are reflected on the t-layer.
- `is_name_of_person` fixed to agree with morphological annotation.
- Grammatemes of firstnames and surnames fixed to agree (in gender and number).

Plus various single fixes over all the layers.