Documentation

This document summarizes changes in the annotation of PDT after the publication of version 2.0. For further details on the individual versions, please refer to their respective documentations:

- PDT 2.0 documentation (http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch05.html)
- PDT 2.5 documentation (http://ufal.mff.cuni.cz/pdt2.5/en/documentation.html)
- PDiT 1.0 documentation (http://ufal.mff.cuni.cz/pdit/documentation)

This document is also available as a technical report in the PDF format (in Czech (/pdt3.0/doc/tr53.pdf) and in English (/pdt3.0/doc/tr54.pdf)).

From PDT 2.0 to PDT 3.0

Content

Introduction
A. Modifications and complements on tectogrammatical layer
1 Grammatemes and *sentmod* attribute
1.1 Grammateme *typgroup*
1.2 Grammateme *factmod*
1.3 Grammateme *diatgram*
1.4 The *sentmod* attribute
2 Modification of the annotation of sentences with lemma *#Benef*
3 Coreference and bridging relations
3.1 Extended textual coreference
3.2 Bridging relations
3.3 Coreference and bridging for the 1st and 2nd person pronouns
4 Discourse relations
5 Genre specification
6 Multiword expressions
7 Valency lexicon PDT-Vallex 3.0
B. Modifications and complements on analytical layer
8 Clause segmentation

Introduction

The second version of the Prague Dependency Treebank 2.0 (PDT 2.0 in further; [4]) was published in 2006. The journalistic texts selected from The Czech National Corpus were included in this version. These texts were annotated on three layers: 2 millions of word tokens were annotated on the morphological layer, the part of them (1.5 million tokens corresponding to 88 thousand sentences) was annotated on the analytical layer (level of surface syntax), and 0.8 millions of word tokens (corresponding to 49 thousand sentences) was annotated on the tectogrammatical layer (deep syntax level). These text annotations enriched by the detailed grammatical data fill first of all two aims:

- In the domain of computational linguistics, the data from PDT are used as the tools for natural language processing, namely for the machine translation. The annotated data are also used for machine learning procedures based on the natural language.
- The other aim of PDT is an exploitation of the treebank data for linguistic studies of contemporary Czech. A great number of scientific articles, books, conference contributions have been published; all submitted results (including diploma and doctoral thesis) were based on empirical data from PDT or they were used as the training data for statistical models of language.

The published data from PDT 2.0 are used first of all by the researchers from UFAL and by their undergraduate and postgraduate students in the branch of informatics as well as of humanistic studies. PDT 2.0 is very well known and it is highly evaluated not only in the Czech Republic, but even abroad. The scenario of PDT as well as annotated data are used at international scientific communities (for Slovak, Slovene, Greek, Danish, Arabic languages). PDT 2.0 is evaluated as one of the richest annotated natural language.

However, it was known from the very beginning of annotation procedure that some rules and instructions applied for the annotation are only preliminary because of the complications of natural language, namely in the domain of the deep syntax. It was also clear (though the data passed through multiple checking) that some errors of human annotators are present.

For the future development of the PDT, the decision that the improvement of the PDT will be reached rather by the deeper and theoretically more adequate analysis of data than by their extension was accepted. The extended annotation of discourse relations was introduced into the new annotation scenario (within the Prague Discourse Treebank project; see [6] and [31]).

The PDT 2.5 (see [2] and [3] was published as a middle step between the development of the PDT 2.0 to the PDT 3.0 version in 2012. It continues the same input data as PDT 2.0, but their annotation was extended in the following way:

- the new grammateme of noun (*typegroup*) was introduced and systematically annotated through all tectogrammatical data
- the dictionary of multiword expressions was applied
- the algorithm for the segmentation of the sentence into clauses was developed and applied for the analytical layer
- the individual errors were corrected on all 3 layers

Information about PDT 2.5 (including detailed annotation as well as the documentation for the annotated data) is available at http://ufal.mff.cuni.cz/pdt2.5/en/ (http://ufal.mff.cuni.cz/pdt2.5/en/) (due to the care of Jan Štěpánek).

The more extended improvement of the older scenario was applied in PDT 3.0. It is presented in this technical report. The description of changes and additions for the annotation procedure, the motivation for them, the procedure of their annotation and exemplification are given in the first part of this report. The extensions of the textual coreference, inter-sentential relations and genre classification of the texts included in PDT 3.0 are described in the other part of the report.

A. Modifications and complements on tectogrammatical layer

1 Grammatemes and *sentmod* attribute

Description

Substantive grammatemes

The new substantive grammateme *typgroup* was *introduced reflecting* the semantic opposition of the pair/group meaning vs. meaning of single entities. The values of this new grammateme, examples and the procedure of their annotating in PDT 3.0 are given below (see 1.1). The other substantive grammatemes stay without changes.

Verbal grammatemes

The changes in the scheme of verbal grammatemes (compared with the scheme applied for PDT 2.0) were influenced by the deeper theoretical studies concerning the meanings of verbal categories. The new grammateme “diathesis” (*diatgram)* was introduced reflecting the distinctions partially covered by the category of verbal voice in the traditional handbooks of Czech grammar. The values of this new grammateme, examples and the procedure of their annotating in PDT 3.0 are given below (see 1.3).
The new scenario of verbal grammatemes (comparing with the PDT 2.0 and PDT 2.5):
The following grammatemes **were canceled**:

- the grammateme *dispmod*,
- the grammateme *resultative*.

The grammateme *verbmod* **was changed** into grammateme *factmod* (see 1.2)
The following grammatemes **stay without changes**:

- the grammateme *deontmod* (deontic modality),
- the grammateme of *aspect,*
- the grammateme of *tense,*
- the grammateme of *iterativness.*

The following grammateme **was introduced**:

- the grammateme of diathesis (*diatgram*).

1.1 Grammateme *typgroup*

Description

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).

For the related literature, see the publications [7], [8], [9] and [10] in the list of references at the end of this document.

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.)

1.2 Grammateme *factmod*

Description

The *factmod* grammateme captures the difference whether an event is presented by the speaker as given or hypothetical (so-called factual modality). These modal meanings are expressed by the morphological category of verbal mood in the surface structure of the sentence. Events are presented as given by the indicative form of a verb. Two types of hypothetical events are distinguished according to the structure of the category of verbal mood in Czech: events that could happen (potential events) are expressed by the present conditional, events that cannot happen (irreal events) are unambiguously expressed by the past conditional (which is, however, frequently or even predominantly substituted by the present conditional in Czech in the last decades). Although the imperative mood has been considered as a means for expressing communicative function rather than factual modality in theoretical studies on Czech mood, events expressed by (both synthetic and analytical) imperative forms are captured as the fourth modal meaning of the *factmod* grammateme (events presented by the speaker as required) in PDT 3.0, in order to cover all meanings expressed by the category of verbal mood by a single means (grammateme) in the annotation.
The following four values of the *factmod* grammateme have been defined:

- *asserted* - events presented as given (asserted events)
- *potential* - events presented as potential
- *irreal* - events presented as irreal
- *appeal* - events presented as required

Since the modal meanings captured by the *factmod* grammateme are expressed only by verb forms specified for verbal mood (i.e. finite verb forms), infinite verb forms (infinitives, participles and transgressives) were assigned another (fifth) value *nil*.
The *factmod* grammateme substitutes the *verbmod* grammateme, which was assigned to the data of PDT 2.0. The main difference between these grammatemes concerns both types of hypothetical events. In PDT 2.0, potential and irreal events were both assigned the same *verbmod* value (*cdn*) and discerned by the value of the *tense* grammateme; this annotation contradicted the theoretically well-described fact that forms of the conditional mood lack the temporal meaning in Czech. In the PDT 3.0 data, the values *potential* vs. *irreal* of the *factmod* grammateme enable to reflect the difference between the past and present conditional, which consists in the feasibility of the respective event, in a theoretically adequate way. In connection with this modification, the *tense* grammateme had to be changed to *nil* with nodes assigned the *potential* or *irreal* value of the *factmod* grammateme in PDT 3.0.
For the related literature, see the publications [12], [13], [14] and [15] in the list of references at the end of this document.

Annotation procedure

The *factmod* grammateme has been assigned to nodes that represent finite verb forms by a semi-automatic procedure. Information from the morphological layer has been extensively used during the automatic part of the procedure. Subsequently, lists of assigned occurrences have been checked manually in order to improve the automatic assignment and to identify exceptions in specific contexts that had to be handled individually.

Examples

Verbs in examples (1) to (5) are assigned the basic values of the *factmod* grammateme; an infinitive with the value *nil* can be found in ex. (1). Sentences with synthetic imperative forms are considered as expressing imperative sentence modality in the PDT (ex. (4)), whereas analytical imperative forms are usually part of desiderative sentences (ex. (5); see the *sentmod* grammateme capturing the sentence modality). In ex. (6), an irreal event is expressed by the present conditional instead of the past conditional; this substitution has not been marked in the annotation, the grammateme value *potential* was chosen on the basis of the formal features. The value of the *factmod* grammateme is displayed with the respective verb form (marked in bold).

1. Pokud **dojde**.*asserted* k omylu, **lze**.*asserted* zpětně **požádat**.*nil* nového majitele, **aby poukázal**.*potential* peníze správnému majiteli cenných papírů. (= When a mistake **occurs**.*asserted*, it **is**.*asserted* possible **to ask**.*nil* the new owner that he **would remit**.*potential* money to the right owner of securities.)
2. Uhrát tu remízu **by bylo**.*potential* úspěchem. (= To draw the game **would be**.*potential* a success.)
3. Většina bangladéšského muslimského obyvatelstva **by** za normálních okolností inkriminované interview samozřejmě vůbec **bývala nezaznamenala**.*irreal*. (= Of course, the majority of Bangladesh Muslim inhabitants **would not have noticed**.*irreal* the interview in question under common circumstances at all.)
4. **Zvedněte**.*appeal* telefon a **zavolejte**.*appeal*. (= **Take**.*appeal* the phone and **call**.*appeal* (us).)
5. **Ať si** provincie konečně **oddychne**.*appeal*. (= **Let** the province finally **relax**.*appeal*.)
6. Svatý pijan Joseph Roth **by** dnes **oslavil**.*potential* rovnou stovku. (= The saint drunkard Joseph Roth **would celebrate**.*potential* his 100th birthday today.)

1.3 Grammateme *diatgram*

Description

The intention to combine the morphological meanings of active and passive voices, resultative and recipient diathesis, dispositional diathesis and reflexive deagentive constructions under the same category called diathesis was the primary aim for introducing the grammateme *diatgram*. These values are understood as meanings of single verbal category; in the case of dispositional diathesis and reflexive deagentive construction the special syntactic requirements must be filled. Some of these diathesis are more productive (passive, deagentive), some are less productive, nevertheless grammaticalized enough to be considered morphological categories belonged to the verbal paradigm.
For any finite form of the verb one of the following values must be applied:

(a) ***act*** Karlovu univerzitu **založil**.*act* Karel IV.
(b) ***pas*** Karlova univerzita **byla založena**.*pas* Karlem IV.
(c) ***res1*** Obchod **je otevřen**.*res1* denně mimo neděli.
(d) ***res2.1*** Obchod **má otevřeno**.*res2.1*.
res2.2 Firma už **má** smlouvu **podepsánu**.*res2.2.*
(e) ***recip*** Horníci **dostanou** v lednu **přidáno**.*recip.*
(f) ***disp*** Tento produkt **se** dobře **prodává**.*disp.*
(g) ***deagent*** **Čeká se**.*deagent* krutá zima.
Knihy **se** dnes **vydávají**.*deagent* i v elektronické podobě.

For the values (a), (d), (e), (f), (g) the annotation procedure is based on the formal exponents (the presence of auxiliaries *mít, dostat* for (d) and (e), reflexive particle co-occurred with adverb of evaluation for (f), reflexive form co-occurred with the general actor (*#Gen*.ACT) in one clause). The most difficult part for the right assignment of verbal grammatemes is to describe the difference between (b) and (c), because their forms are formally identical, though one of them expresses an action (*pas*) and the other (*res1*) describes a state.
For the related literature, see the publications [16], [17] and [18] in the list of references at the end of this document.

Annotation procedure

In PDT 3.0, this set of grammatemes was annotated semiautomatically according to the new scheme of grammatemes.
(d) Possesive resultative (*res2.1* and *res2.2*) was searched in the PDT 2.0 by the script based on co-occurrence of the verb *mít* and *–n/-t p*articiple. The syntactic structure of these examples was checked and corrected: for the cases, where the ACT is in the position of subject, the grammateme *res2.1* was assigned (see ex. (1)) without changes in grammatical structure; for the cases where the ADDR is used as surface subject of the clause, the grammateme *res2.2* is assigned (see ex. (2)) and the syntactic structure is changed (the ADDR is in a position of subject and usually the node for *#Gen*.ACT is added).
(e) The examples of the **recipient diathesis** *(recip*) were searched automatically, the used script was based on the occurrence of the auxiliary *dostat* and *–n/-t* participle (see ex. (3)).
(f) Dispositional constructions were annotated in PDT 2.0 in the grammateme *dispmod.* Its value *dispgram=1* was shifted to the position of the grammateme *diatgram* with the value *disp* (see ex. (4)).
(b) Reflexive deagentive was searched automatically by the co-occurence of reflexive form and the node for *#Gen*.ACT. In PDT 2.0 1 973 examples were found (see Table 1.1). Their value *deagent* vas filled as a value of the grammateme *diatgram* (see ex. (5) and (6) with transitive and intransitive verbs, respectively).
(b), **(c)** For the constructions with the forms of the verb *být* and *–n/-t* participle there were several steps, how to determine the difference between **passive** (voice/diathesis) and **simple resultative** (*res1*). They combined manual and automatic procedures:

- an ACT(or) is present as a child of this form → *pas*
- the form of *–n/-t* participle was *neutrum sg*, *#Gen*.ACT is present in the clause → *res1*
- analytical structure consist of verb and *AuxV* as its child → *pas*
- analytical structure consist of the predicate *být* and *PNom* as its child → *res1*

The sample of 750 examples were checked manually. The differences between the results of the script and the manual annotating procedure were checked once more (227 examples and 108 examples where from the various reasons the script did not fit), see ex. (7) and (8).
(a) The rest of examples are annotated by the grammateme *act* as an unmarked member of the diathesis category.
The total numbers of the *diatgram* values in the PDT 3.0 are in Table 1.1.

diatgram		
(a)	*act*	81,257
(b)	*pas*	3,743
(c)	*res1*	967
(d)	*res2.1*	55
	res2.2	28
(e)	*recip*	0

(f)	*disp*	9
(g)	*deagent*	1,973

Table 1.1: Values of grammateme *diatgram* in PDT 3.0

Examples

1. Já.ACT **nemám** vše **domyšleno**.*res2.2*, nejsem si jist, jestli…
2. Klub.ADDR **má** na letošní rok financování **zajištěno**.*res2.1* ze státního rozpočtu.
3. Výrobci **nedostanou zaplaceno**.*recip* dříve než v březnu.
4. **Hrálo se**.*disp* mi.ACT výborně, vůbec se mi nechtělo střídat.
5. Doplatili na to, že **se potvrdil**.*deagent* jejich optimistický odhad inflace.
6. Na bezpečnost práce **se** mnoho **nehledí**.*deagent*.
7. Od té doby **byl** černý trh tímto opiátem.ACT **přehlcován**.*pas*.
8. Z vyšší daňové sazby **je vyňato**.*res1* ubytování a stravování při dětských rekreacích a táborech.

1.4 The *sentmod* attribute

Description

The *sentmod* attribute captures the modality of the sentence, i.e. whether the sentence expresses an assertion, a question, a demand etc. In written texts, sentence modality is expressed by a combination of formal means in the surface structure of the sentence, namely by the mood of the verb form, by the final punctuation mark, by the word order, and by modal particles *ať, kéž, nechť*.

The *sentmod* attribute was already available in the data of PDT 2.0. However, since the *sentmod* assignment in PDT 2.0 was simplified in that only one *sentmod* value was determined for the whole coordination structure (i.e. the fact that coordinated clauses can have different sentence modalities was intentionally omitted), the annotation had to be revised and reimplemented in the data of PDT 3.0.

The values of the *sentmod* attribute used in PDT 3.0 are the same as in PTD 2.0:

- *enunc* - declarative modality (assertions)
- *excl* - exclamative modality (exclamations)
- *desid* - desiderative modality (wishes)
- *imper* - imperative modality (requests/orders)
- *inter* - interrogative modality (questions)

The principle that the values of the *sentmod* attribute are assigned on the basis of its position in the tectogrammatical tree remained in the PDT 3.0 data the same as in PDT 2.0.

The difference between the *sentmod* assignment in PDT 2.0 and 3.0 concerns the set of nodes to which a *sentmod* value is assigned. In PDT 2.0 a *sentmod* value was assigned to nodes listed under (a) to (c) bellow. The main motivation for revision of the *sentmod* annotation has been the elimination of the above mentioned simplification with regard to the treatment of coordination structures. However, at the very beginning of the revision, subtrees representing title structures (identified with the *ID* functor) have been recognized as another type of embedded structures (in addition to direct speech in (b) and parentheses in (c)) which express their 'own' sentence modality, which might differ from the modality of the sentence that the title is embedded in (cf. (d)).

Since the decision to specify a *sentmod* value for each clause in coordination in PDT 3.0 affected all the original subgroups (a) to (c) (as well as the subgroup (d)) and, moreover, errors of several types had been corrected during a systematic revision of the PDT 2.0 data carried out in the recent two years, the *sentmod* values available in the PDT 2.0 data were canceled and the set of nodes to be assigned a *sentmod* value has been newly delimited in the PDT 3.0 data. For the delimitation of the candidate nodes, the steps (a) to (c) have been completed with the steps (d) and (e) and all the steps have been applied to the data from the scratch:

a. child nodes of the technical root node, i.e. nodes representing the main verb or noun and the coordination roots (root nodes of coordination structures),
b. root nodes of subtrees representing direct speech (identified on the basis of the attribute *is_dsp_root*),
c. root nodes of subtrees representing a (syntactically independent) parenthesis, the effective root of which was assigned the *PAR* functor,
d. root nodes of title subtrees (labeled with the functor *ID*),
e. from all these candidates, coordination roots were extracted and handled separately.

Annotation procedure

The non-coordination nodes, which remained after application of the step (e), were assigned a *sentmod* value semi-automatically according to the following procedure, taking advantage of the links between the tectogrammatical, analytical and morphological annotation:

i. if the node represented a synthetic imperative verb form (i.e., technically, if one of the morphological tokens which the node was interlinked with was assigned the tag *Vi.** (imperative verb form)), the node was assigned the *sentmod* value *imper*;
ii. if the syntactic structure to which the node belonged ended with a question mark (technically, if the node corresponded to an analytical node that had a question mark among its child nodes), the *sentmod* value *inter* was filled in;
iii. from the rest of the nodes, nodes that were a part of a sentence introduced by the particles *ať, kéž, nechť* and/or ended with an exclamation mark were identified and assigned manually one of the *sentmod* values *desid*, *excl* or *imper*;
iv. the remaining nodes were assigned the *sentmod* value *enunc*.

Coordinations were handled as a homogeneous group, regardless which of the subgroups (a) to (d) they belonged to. On the basis of the extracted list of coordination roots, the set of root nodes of coordinated clauses which were to be assigned a *sentmod* value was delimited.

For the sake of specification of the *sentmod* value for the root of each coordinated clause, the step (i) of the annotation procedure could be applied “locally”, i.e. just for the particular clause of the coordination structure, not for all the clauses in a coordination: root nodes of the individual coordinated clauses that represent an imperative form were assigned the value *imper*.

Those non-imperative clauses which were coordinated with the imperative ones were extracted to be assigned a *sentmod* value manually. The second portion for manual annotation were roots of coordinated clauses that were part of a coordination structure ending with a question mark. Our assumption that the question mark occurring as the final punctuation mark of the whole coordination structure is to be interpreted as a signal of the sentence modality just for the final clause of the coordination structure (i.e. it does not mirror the sentence modality of the non-final clauses) proved to be true during the annotation. Roots of coordinated clauses which were part of a coordination structure ending with an exclamation mark and/or involving the particles *ať*, *kéž* and *nechť* were the third portion for manual annotation. The manual annotation was carried out by two annotators in parallel, with the inter-annotator agreement of 93.7% (Cohen's Kappa 0.89).

All the remaining coordination structures ended with a period (or without punctuation etc.) and involved only clauses with an indicative or conditional verb form. As in 100 coordination structures randomly selected from this group, only coordinated clauses with declarative modality were found, clauses in these coordination structures were automatically assigned the *sentmod* value *enunc*.

The resulting assignment of the *sentmod* values to the PDT 3.0 data is contrasted to the annotation available in the PDT 2.0 data before the revision in Figure 1.1 and 1.2.

Visualisation

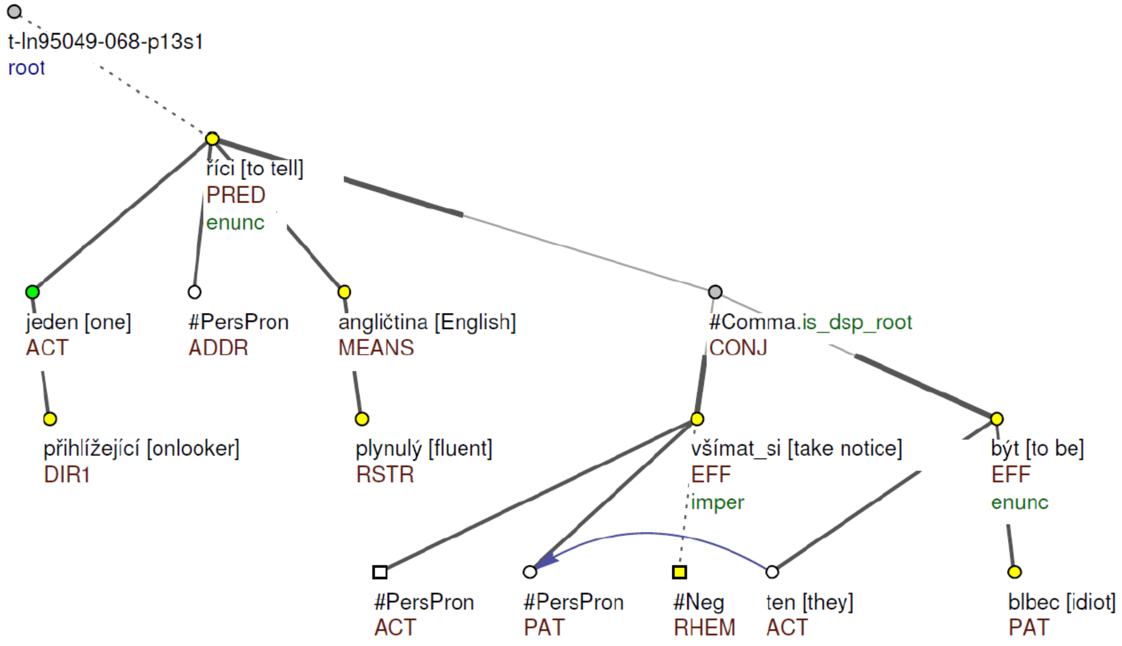

Figure 1.1: Values of the *sentmod* attribute in PDT 3.0: tectogrammatical tree of the sentence *“Nevšímejte si jich, jsou to blbci,” řekl mi plynulou angličtinou jeden z přihlížejících*. (“Do not take notice of them, they are idiots,” told me one of the onlookers in fluent English.), in which two clauses with different sentence modalities are coordinated within a direct speech (the coordination root is assigned the functor *CONJ* and the attribute *is_dsp_root*). In PDT 3.0, an individual *sentmod* value is specified for each clause of the direct speech (values *imper* and *enunc*) as well as for the matrix clause (*enunc*).

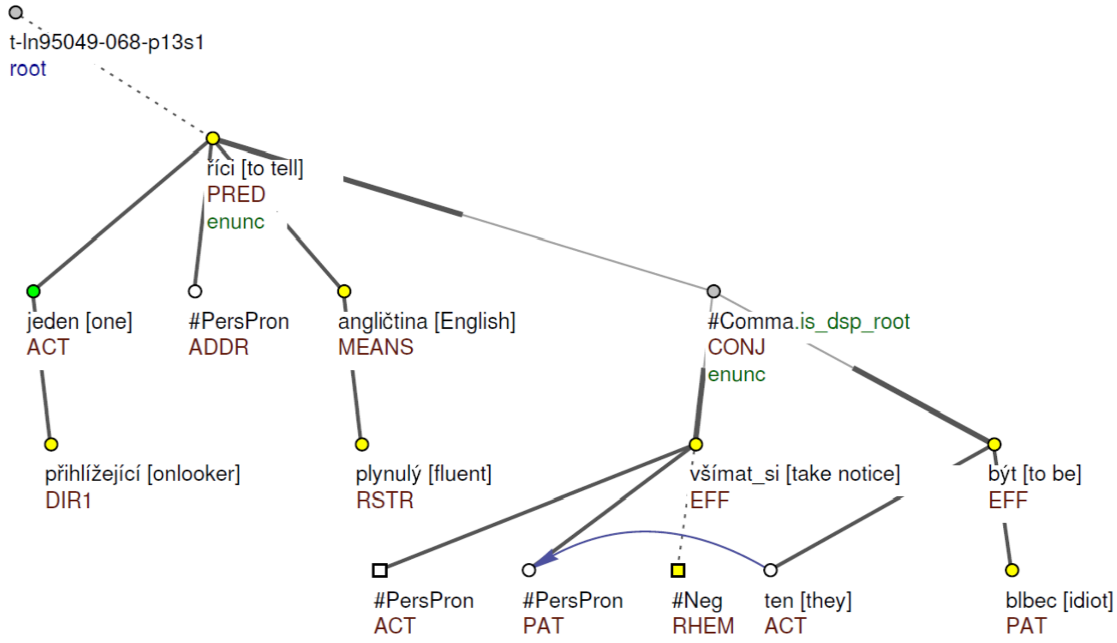

Figure 1.2: Values of the *sentmod* attribute in PDT 2.0: tectogrammatical tree of the sentence described in Figure 1.1. Within the original PDT 2.0 annotation, the *CONJ* node was assigned the *enunc* value, the imperative modality of the first clause of the direct speech was omitted.

Examples

Examples (1) to (5) illustrate the five values of the *sentmod* attribute, respectively. The example (6) demonstrates that a sentence that involves just one main (syntactically independent) clause expresses (as a whole) a single *sentmod* value. On the contrary, in a coordination structure each of the syntactically independent clauses can have a different modality (ex. (7)). Similarly, embedded structures (direct speech, parenthesis, and title structures) that express their 'own' sentence modality are described in ex. (8) to (10), respectively. If there was a conflict between the mood of the verb (synthetic imperative form in the main clause in ex. (11)) and the final punctuation mark (a question mark in (11); cf. steps (i) and (ii) above), the *sentmod* value was assigned according to the final punctuation mark (*inter* in (11)). The value of the *sentmod* attribute is displayed with the head of the respective clause (marked in bold).

1. Ekonomika **jde**.*enunc* do vzestupu už letos. (= The economy **rises**.*enunc* already this year.)
2. Jaká **je**.*inter* nezaměstnanost v této zemi? (= How big **is**.*inter* the unemployment in this country?)
3. **Podívej se**.*imper* na mě! (= **Look**.*imper* at me!)
4. Ať **si** provincie konečně **oddychne**.*desid*. (= **Let** the province finally **relax**.*desid*.)
5. To **nejsou**.*excl* špatně rozdané karty! (= The cards **have been dealt**.*excl* not at all badly!)
6. **Neptejte se**.*imper* mě, proč jsem přijel do Prahy. (= **Do not ask**.*imper* me why I came to Prague.)
7. Poprvé **jste nastoupil**.*enunc* v závěru zápasu v Benešově, jaké to **bylo**.*inter*? (= For the first time you **entered**.*enunc* the game before the end of the match in Benešov, what **was**.*inter* it like?)
8. Kam **se poděla**.*inter* má bojovnost? **ptala**.*enunc* se sama sebe po utkání Martinezová. (= Where **did** my fighting spirit **disappear**.*inter*? Martinezová **asked**.*enunc* herself after the match.)
9. Pane kolego, **věřte**.*imper* **nevěřte**.*imper*, počítač **nelže**.*enunc*. (= Mr. colleague, **believe**.*imper* it or **do not believe**.*imper* it, a computer **does not lie**.*enunc*.)
10. Zítra **bude** u příležitosti III. výročí české a slovenské edice Playboy **otevřena**.*enunc* výstava **Pohlaďte si**.*imper* králíčka sestavená z ilustrací pro časopis Playboy. (= An exhibition **Stroke**.*imper* a bunny rabbit consisting of illustrations for the magazine Playboy **will be opened**.*enunc* tomorrow on the occasion of the 3rd anniversary of the Czech and Slovak editions of Playboy.)
11. **Hádejte**.*inter*, kde se toto menu ve Windows najde? (= **Guess**.*inter* where this menu is to be found in Windows?)

For the related literature, see the publication [19] in the list of references at the end of this document.

2 Modification of the annotation of sentences with t-lemma *#Benef*

Description

The substitutional t-lemma *#Benef* belonged to a newly created node (*is_generated=1*) in PDT 2.0 data – the node represented unexpressed free modification with the meaning of 'beneficient' (*functor = BEN*) on the surface form of the sentence. The node was complemented into the position of the controlling member (controller) in the following types of structures with control in which the infinitive has a role of subject or attribute:

- Construction *být* (*to be*) + predicative substantive: the infinitive has a role of subject, e.g. *Transformovat bezpečnostní složky je hračkou* [for anyone] (= literally: To transform the security forces is a child's play [for anyone], It is a child's play to transform the security forces [for anyone]); *Je nutností* [for someone] *pořídit vybavení* (= It is a necessity [for anyone] to purchase equipment.); *Je radost* [for anyone] *dostávat dárky.* (= It is a pleasure [for anyone] to receive gifts.) (in the Manual [5] Chapters 8.2.4.4.4.2, 8.2.4.4.4.3 and 8.2.4.4.4.5);
- Construction *být* (*to be*) + predicative adjective: the infinitive has a role of subject, e.g. *Je nutné* [for anyone] *přijít.* (= It is necessary [for anyone] to come.); *Je trapné* [anyeone] *přijít pozdě.* (= It is embarrassing [for someone] to be late.) (in the Manual [5] Chapters 8.2.4.4.4.4 and 8.2.4.4.4.5);
- Construction *být* (*to be*) + predicative adverb, e.g. *Je škoda* [for anyone] *se ochudit o tolik vzácných látek.* (= It is a pity [for anyone] to impoverish yourself of so many rare substances.) (In Czech, the lexeme škoda (‚pity') is an adverb; in the Manual [5] Chapter 8.2.4.4.4.6);
- The infinitive depends on the predicate *lze* ('it is possible, can'), e.g. Lze [for anyone] *tam přijít kdykoli.* (= It is possible [for anyone] to come there any time.) (in the Manual [5] Chapter 8.2.4.4.5);
- The control of the type *Je vidět Sněžku.* (= literally: (It) is seen Sněžka.Accusative; Sněžka is seen.) (in the Manual [5] Chapter 8.2.4.4.5);
- Constructions derived from the above mentioned (in the Manual [5] Chapters 8.2.4.5.1 and 8.2.4.7.1).

In the PDT 3.0 data, the node with the t-lemma *#Benef* was replaced:

- By a node with t-lemma *#Gen* (*functor*=BEN; *is_generated=1*) in the case of general benefactor.
  E.g. *Je dobré chodit brzo spát.* (= It's good to go to bed early.) = that is good for anyone
- By a node with t-lemma *#PersPron* (*functor=BEN; is_generated=1*) in case of textual ellipsis. In these cases, also the corresponding textual coreference was annotated and the appropriate grammatemes of the node were filled in.
  E.g. *Pavel přišel včera pozdě. Bylo by dobré jít dnes brzo spát.* (= Paul came late yesterday. It would be good to go to bed early today.) = it would be good for Paul to go to bed early.

Motivation for the change

The empirical research demonstrated that the constructions with the control contain also such infinitive constructions in which their subject is controlled by a free verbal modification expressing benefit (BEN – beneficient). The free verbal modification is either explicitly expressed (*Povinnost starat se o zámek plyne* ***pro majitele*** *ze zákona.* (= The obligation to take care of the castle follows **for the owner** from the act.)) or there is a contextual ellipsis (*Čím větší odchylka, tím víc čeká firmu práce navíc, protože je třeba* [**pro firmu**.BEN] *výpadek kompenzovat jiným zbožím.* (= The greater the deviation is, the more the company expects some extra work because it is necessary [**for the company**.BEN] to compensate the failure for other goods.)) or the verbal modification is generalized (*Je dobré chodit brzo spát.* (= It is good to go to bed early.)). The interim solution applied in PDT 2.0 and 2.5 where the artificial t-lemma *#Benef* was added was canceled. The annotation nowtook similar shape like in other structures with the expressed controller or with the t-lemma for generalization (*# Gen*) with the BEN functor.
For the related literature, see the publications [20], [21], [22] and [23] in the list of references at the end of this document.

Anotation procedure

The PDT 2.0 data contained 1,394 nodes with t-lemma *#Benef* (*functor=BEN; is_generated=1*). 100 occurrences were replaced manually. The remaining 1,294 nodes were automatically transferred to the node with the t-lemma *#Gen* (*functor = BEN; is_generated = 1*). This was done because the incomplete annotation of valency of nouns and adjectives (see the Manual [5] Chapter 5.2.4) does not enable, in most cases, to annotate cases of contextual ellipsis correctly (there is no place where to lead the coreferential arrow from supplemented node with the t-lemma *#PersPron*).

Examples

Examples with the general benefactor in the position of the controlling member

1. Je-li vypovídání smluv legální, je nutné **[*#Gen*.BEN]** novelizovat zákony. (=If the canceling contracts is legal, it is necessary [**# Gen**.BEN] to amend the laws. (See Figure 2.1)
2. Česká republika, která je toho času nestálým členem Rady bezpečnosti, má možnost zaujmout ke vzniklé realitě jednoznačné stanovisko, neboť je třeba **[*#Gen*.BEN]** podívat se pravdě do očí. (=The Czech Republic which is an unstable member of the Security Council at the moment has the opportunity to take a clear stand on arising reality because it is necessary [**# Gen**.BEN] to face the truth.)

Examples of textual ellipses of benefactor in the position of controlling member

1. Rady **diků m**

   Znovu je tady čas, kdy je třeba **[*#PersPron*.BEN]** se rozhodnout.
   Na majitele kuponových knížek dotírají otázky - kam vložit své body, jaký obor si vybrat, raději investovat do velkého podniku, nebo do neznámého podničku?
   Podobných otázek, na něž samotní dikové, bez patřičných informací jen těžko hledají odpověď, je daleko víc.

(=Pieces of advice to **"DIKs"** ('holders of investment coupons')
Again, it's the time when it is necessary [**# PersPron**.BEN] to decide.

   The owners of coupon books are snowed under with questions – where to put their points, what discipline to choose, is it better to invest in large company or in an unknown small company?
   There are far more similar questions which the DIKs without adequate information can answer very difficultly.)

2. Čím větší odchylka, tím víc čeká **firmu** práce navíc, protože je třeba **[#*PersPron*.BEN]** výpadek kompenzovat jiným zbožím. (=The greater the deviation is, the more the **company** expects some extra work because it is necessary [#**PersPron**.BEN] to compensate the failure for other goods.) (See Figure 2.2)
3. **Hráč** musí sám vědět, co to znamená **[#*PersPron*.BEN]** být profesionálem. (=The **player** himself must know what it means [# **PersPron**.BEN] to be professional.) (See Figure 2.3)

Visualisation

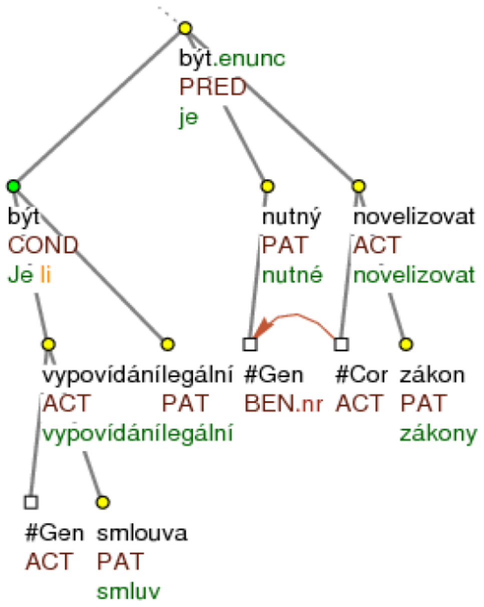

Figure 2.1: *Je-li vypovídání smluv legální, je nutné novelizovat zákony.*

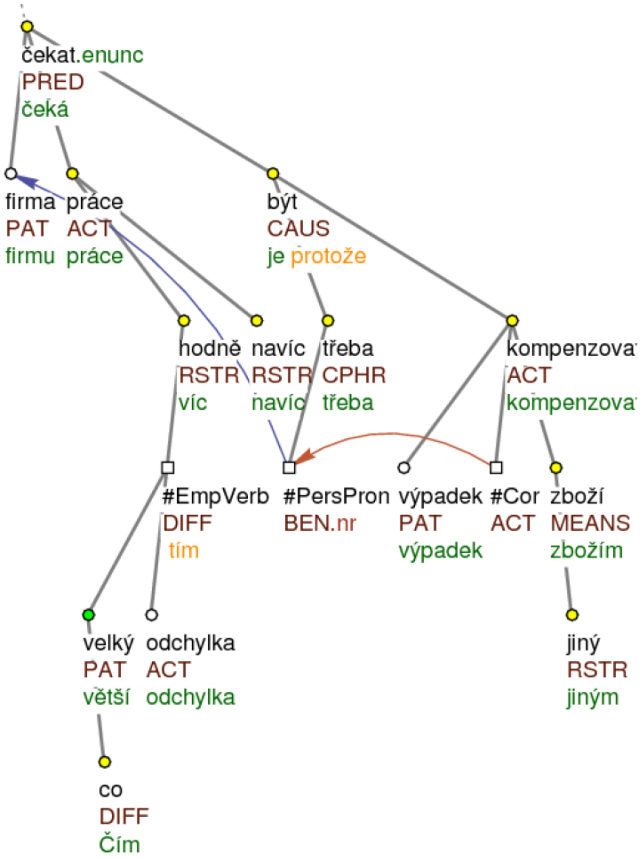

Figure 2.2: *Čím větší odchylka, tím víc čeká firmu práce navíc, protože je třeba výpadek kompenzovat jiným zbožím.*

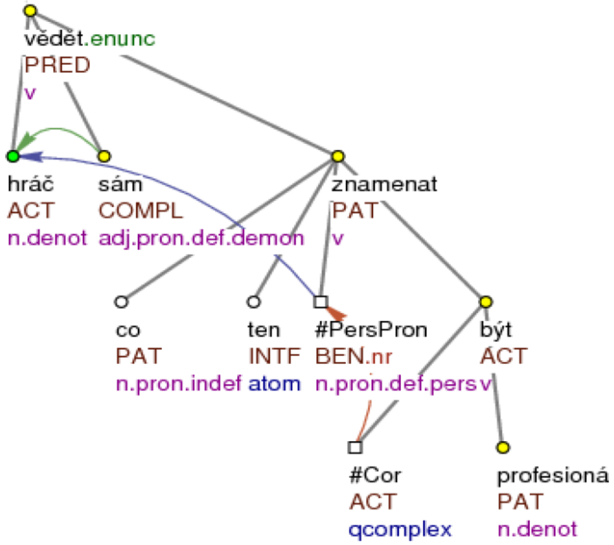

Figure 2.3: *Hráč musí sám vědět, co to znamená být profesionálem.*

3 Coreference and bridging relations

Description

In PDT 2.0, the tectogrammatical level includes the manual annotation of coreference links of two types: grammatical coreference (in which it is possible to pinpoint the coreferring
expression according to grammatical rules; in the Manual [5] Chapter 8.2) and textual coreference (where reference is not only expressed by grammatical means, but also via context; in the Manual [5] Chapter 8.3). Textual coreference is annotated for 3rd person personal and possessive pronouns, the demonstrative pronouns *ten, ta, to,* and textual ellipsis.
In PDT 3.0, the annotation of this phenomenon was enriched with the following:

- coreference annotation was extended to other types of coreferring expressions (see 3.1);
- annotation of some types of bridging relations was manually provided (see 3.2);
- coreference and bridging relations were also annotated with the first and second person pronouns (see 3.3).

By annotating coreference and bridging relations, the principle of maximum size of an anaphoric expression was applied. It is always the whole subtree of the antecedent/anaphor which is subject to annotation. Technically, coreference arrows go from/to the governing nodes of the coreferring expressions.
Annotation of textual coreference is based on the chain principle, the anaphoric entity always referring to the last preceding coreferential antecedent. In case of bridging anaphora, the chain principle is not preserved.
Exactly speaking, coreference and bridging relations are part of discourse layer and that portrays linguistic phenomena from the perspective of the discourse structure and coherence. However, technically the annotation of extended nominal coreference and bridging relations is based on the tectogrammatical level. This methodological approach allows us to include the relevant syntactic phenomena annotated previously (functors, node types, grammatemes etc.), and to take advantage of the syntactic structure in itself (the resolution of elliptical structures, parentheses, predicative relations, appositions, etc.).
For the related literature, see the publications [24], [25], [26], [27] and [28] in the list of references at the end of this document.

Annotation procedure

Annotating extended textual coreference and bridging anaphora consists of the following actions:

- automatic pre-annotating (e.g. linking some named entities),
- automatic useful tools which help annotators find the correct antecedents (highlighting already linked items in the trees, underlining the same lemmas, etc.),
- manual annotating,
- automatic check of some aspects of coreference links (finding the nearest antecedent, preserving coreferential chains, bridging long coreferential chains)

Visualisation

The Figure 3.1 shows the basic features of the coreference and bridging annotation. Coreference/bridging relations between subtrees are marked by arrows of different colors (dark-red arrows for grammatical coreference, dark-blue arrows for textual coreference and light-blue arrows for bridging reference), the arrow pointing from an anaphor to an antecedent. If an antecedent is found in one of the preceding sentences, its lemma is written in dark-blue next to its anaphor.

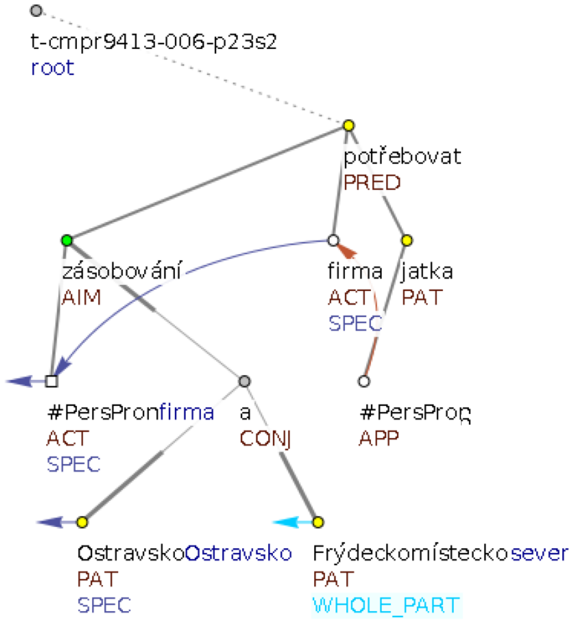

Figure 3.1: *Pro zásobování Ostravska a Frýdeckomístecka potřebuje firma svá jatka.*

3.1 Extended textual coreference

Description

In PDT 3.0, coreference relations were also manually annotated for noun phrases, adjectives derived from named entities (*pražský* (= adj. derived from Prague)*, český* (=Czech)), and for pronominal adverbs (*tak* (=so), *tam* (=there), *tehdy* (=then) etc.) which have explicit antecedents in previous (ev. subsequent) context.
The textual coreference now consists of pronominal and zero coreference (completed for PDT 2.0) and extended nominal coreference. The coreference annotation is captured in a structured attribute coref_text at the start node of the relation, containing the identifier of its antecedent and the type.
In PDT 3.0, coreference relations of the following two types have been annotated:

- ***SPEC*** - coreference of noun phrases with specific reference (*Germany – the state*); ex. (1).
- ***GEN*** - coreference of noun phrases with generic reference; ex. (2)

All coreferring nodes are classified according to their specificity/genericity. By default, pronouns and zero anaphoric nouns get the SPEC type. If needed, this can be changed manually. In ambiguous cases with specific nouns, the coreference SPEC type is preferred.
The annotation concentrates on marking the equivalence of referents of the antecedent and anaphoric expressions, not only anaphoric relations in a restricted case are annotated.
The textual coreference is marked within the length of up to 20 sentences. Annotating coreference for a greater length is only possible in the case of automatic pre-annotating named entities coreference.
Only one textual coreference arrow can start from or end in one tectogrammatical node.
Otherwise, two special cases of (co)reference were annotated in PDT 2.0 within the textual coreference group: references to situations or reality external to the text (coref_special, exoph type) and references to a discourse segment consisting of more than one sentence (coref_special, segm type). In PDT 3.0, they were enriched with the cases where the referring expressions were neither zero nor pronouns:

- Exophoric relations (exoph type) are marked in case of time and local deixis, deixis with pronominal adverbs and by exophoric reference to the whole text; ex (3).
- Reference to a segment (segm type) takes place when either a noun phrase refers to a substantial section of a text consisting of more than one sentence, or a noun phrase refers to a tree segment which cannot be technically extracted as the antecedent.

Examples

1. Jeho dojetí znásobila při vyhlašování přítomnost [...] pořadatelů **soutěže**. Na letošním ročníku **soutěže**.*SPEC* se spolupodílí i Profit. (= He was strongly impressed by the presence […] of the organizers of **the competition** during the announcement .The Profit magazine is also taking part in this year's **competition**.*SPEC*)

2. **Droga** je tak účinná, že ten, kdo **ji**.*GEN* užívá, se snadno dostane do „pohody“ kouřením nebo šňupáním. (= The **drug** is so effective that one can easily achieve the state of “coolness” by smoking or snorting **it**.*GEN*)

3. Dokončeny by měly být ... **v těchto dnech**.*exoph* (= It should have been finished **in these days**.*exoph* [meaning, in the recent days]).

3.2 Bridging relations

Description

Apart from extended textual coreference, non-coreferential association relations are annotated as bridging relations if they are related in one of specific types of semantic, lexical or conceptual ways to their antecedents. The bridging annotation is captured in a structured atribute bridging at the start node of the relation, containing the identifier of its antecedent and the type.
In PDT 3.0, bridging relations of the following types have been annotated:

- ***PART_WHOLE*** and ***WHOLE_PART*** - metonymical relation between a part and a whole; ex. *room - ceiling*, *Germany – Bavaria - Munich*;

- ***SET_SUB*** and ***SUB_SET*** - the relation between a set and its subsets/elements; ex. *students – some students – a student;*
- ***P_FUNCT*** and ***FUNCT_P*** - the relation between an entity and a singular function on this entity; ex. *prime-minister – government, trainer – football team*
- ***CONTRAST*** - the relation between coherence-relevant discourse opposites; ex. (1),
- ***ANAF*** - non-coreferring explicit anaphoric relation; ex. (2),
- ***REST*** - further underspecified group: family (*grandfather - grandson*), place – inhabitant, author – work, the same denomination to support the cohesion of the text (*a chance helped – another chance entered the game*) and an event – a participant of the event (*enterprise - entrepreneur*).

The types *PART*, *SUBSET* and *FUNCT* are further underspecified according to the linear order of the antecedent and the anaphor in the text, e.g. the *PART_WHOLE* is used for cases where the antecedent of the anaphoric NP corresponds to the whole of which the anaphor is a part, and *WHOLE_PART* - for the opposite.
Unlike PDT 2.0, the reference of a pronoun to more than one tectogrammatical node is now marked as a bridging relation, *SUB_SET* type. (Cf. *na ně* (=for them) referring to both *Marie* and *Vlasta* in ***Marie***_*vzala* ***Vlastu***_*do divadla, kde* ***na ně*** *čekal Marek*. (=**Marie** took **Vlasta** to the theatre, where Marek was waiting **for them**.)).

Examples

1. Dnes, po rozdělení ČSFR, je jasné, že **osud ČR** bude stále více spojený s Německem a přes něj s Evropskou unií a **osud Slovenska**.*CONTRAST* s Ruskem. (= Nowadays, after the split of Czechoslovakia, it is clear that the **future of the Czech Republic** will become more associated with Germany, further with the European Union, while the **future of Slovakia**.*CONTRAS* will be more associated with Russia.)

2. A přesvědčen jsem ještě o jednom – je třeba mít **vysoké cíle** a s **malými [cíli**.*ANAF***]** se nespokojit. (= And I am sure about one thing: it is necessary to have **lofty aims** and not to be satisfied with **small [ones**.*ANAF***]**.)

3.3 Coreference and bridging for the 1st and 2nd person pronouns

Description

Annotation of textual coreference and bridging relations for 1st and 2nd person pronouns was provided additionally on the whole PDT using the annotation guidelines for extended textual coreference and bridging relations.
All cases of first and second person coreference, regardless of whether they can be considered as anaphoric or not, are subject to annotation (ex. (1)).
The generic use of first and second person pronouns is quite frequent, often causing low inter-annotator agreement. Generic pronouns in the first and second person are frequently used for the speakers' companies, states, teams, interest-groups, etc. In clear cases, the coreference relation to their non-pronominal antecedents is annotated (ex. (2)).
The "cataphoric" use of first and second person pronouns is annotated as exophoric reference (*coref_special*, *exoph* type), further coreferential noun phrases (either pronominal or not) referring to them anaphorically (ex. (3)).

Examples

1. Potřebu dalších investic **[*#PersPron*.ACT]** odhaduji do roku dva tisíce na více jak dvě miliardy korun, říká ředitel **Nováček**. (= **I** estimate the need for further investment in the year two thousand to more than two billion, says the director **Nováček**.)
2. **Slévárně Škoda** v Českých Budějovicích dluží plzeňská Škoda 61 miliónů Kčs. **[*#PersPron*.ACT]** Potřebujeme je hned a na stůl. Situace je vážná a z **naší** strany téměř neřešitelná. Bez finančních prostředků se už **[*#PersPron*.ACT]** neobejdeme," řekl včera Milan Fučík. (= The Škoda's branch in Pilsen owes **the foundry Skoda** in České Budejovice 61 million crowns. **We** need them now, and on the table. The situation is serious and almost unsolvable from **our** side. **We** will not manage [to resolve] it without funds," Milan Fucik said yesterday.)

3. Ačkoliv **naše** produkty se běžně prodávají v různých evropských zemích, **Česká republika** ještě není plnoprávným partnerem na evropském trhu. (= Although **our** products are widely sold in various European countries, **Czech Republic** is not yet a full partner in the European market.)

4 Discourse relations

Description

Annotation of discourse relations in the PDT is inspired by the Philadelphia annotation project Penn Discourse Treebank 2.0 [32] and it also partly uses the scenario of the PDT tectogrammatical representation. Czech data with discourse annotation have been first released as a part of the Prague Discourse Treebank 1.0 (PDiT 1.0; [6], [31]) in 2012. The PDT 3.0 includes an update of this annotation, enriched with several newly annotated discourse-related phenomena: genre specification of the corpus texts, annotation of some type of rhematizers (or focusing particles) as discourse connectives, annotation of second relations (discourse relations with more than one semantic type), and the introduction of a new attribute *discourse_special*.

Discourse connectives, discourse units

Discourse annotation in PDT 3.0 is focused on analysis of discourse connectives (DCs), the text units (or arguments) they connect and the semantic relation expressed between these two units. As basic discourse unit entering a discourse-semantic relation is understood an utterance containing a finite verb form (a finite clause). A discourse connective is defined as a predicate of a binary relation – it takes two text spans (mainly clauses or sentences) as its arguments. It connects these units to larger ones while signaling a semantic relation between them at the same time. DCs are morphologically inflexible and they never act as grammatical constituents of a sentence. Like modality markers, they are "above" or "outside" of the proposition. They are represented by coordinating conjunctions (e.g. *and*, *but*), some subordinating conjunctions (e.g. *because*, *if*, *while*), some particles (e.g. *also*, *only*) and sentence adverbials (e.g. *afterwards*), and marginally also by some other parts-of-speech – mainly in case of fixed compound connectives like *in other words* or *on the contrary*. In the PDT 3.0 release, like in the PDiT 1.0, the annotation only focused on discourse relations indicated by overly present (explicit) discourse connectives – the relations not indicated by a discourse connective were not annotated in this stage of the project.

Apart from discourse relations anchored by connectives, discourse annotation in PDT 3.0 includes also marking of list structures (as a separate type of discourse structure) and marking of some other text phenomena like article headings, figure, table and chart captions, non-coherent texts like collections of short news etc.

Annotation of rhematizers in role of discourse connectives

Rhematizers (expressions with the tectogrammatical label (functor) RHEM) are in PDT 3.0 considered to be discourse connectives only if they have a connecting function – that means in our approach if they, in a given context, open two positions that are filled by text spans containing at least one verbum finitum each. An example of such context is given in ex. (1) and (2), rhematizers with a connecting function are given in bold.
If a rhematizer only connects noun phrases or both text spans contain verbs with the same or similar meanings (as in ex. (3) and (4)), it is not considered to be a discourse connective, even if it has sentence-initial position as in ex. (5).
Compared to the PDiT 1.0, the PDT 3.0 release newly includes annotation of such rhematizers that together with a conjunctive connective represent a conjunctive relation within a compound sentence. An example of such connective is shown in ex. (6).

Annotation of second relations

Discourse relations with more than one semantic type are now newly annotated with both types – in two separate discourse relations represented by two attributes *discourse*; each of them marks the respective semantic type and connective. (It means that there are two arrows connecting two nodes representing the arguments, each of the arrows marks a different semantic type and connective.)

New attribute *discourse_special*

The newly introduced attribute *discourse_special* captures three special roles of the phrase represented by the node and its subtree; the possible values are: *heading* (article headings; replaces attribute *is_heading* from PDiT), *metatext* (text not belonging to the original newspaper text, produced during the creation of the corpus), and *caption* (for captions of pictures, graphs etc.).

Discourse relations in PDT 3.0 – Distributions				
Semantic type of relation	**Abbreviation**	**Intra-sentential**	**Inter-sentential**	**Total**
concession	***conc***	617	263	880
condition	***cond***	1,350	19	1,369
confrontation	***confr***	345	308	653
conjunction	***conj***	6,109	1,389	7,498
conjunctive alternative	***conjalt***	69	21	90
correction	***corr***	322	123	445
disjunctive alternative	***disjalt***	257	15	272
equivalence	***equiv***	41	64	105
exemplification	***exempl***	28	120	148
explication	***explicat***	100	130	230
pragmatic condition	***f_cond***	15	1	16
pragmatic contrast	***f_opp***	23	27	50
pragmatic reason + result	***f_reason***	12	28	40
generalization	***gener***	9	97	106
gradation	***grad***	241	204	445
opposition	***opp***	1,396	1,800	3,196
other	***other***	1	1	2
precedence + succession	***preced***	591	249	840
purpose	***purp***	413	1	414
reason + result	***reason***	1,601	1,031	2,632
restrictive opposition	***restr***	97	172	269
specification	***spec***	519	111	630
synchrony	***synchr***	174	52	226
Total		**14,330**	**6,226**	**20,556**
			List structures in total	83

Table 4.1: Discourse relations in PDT 3.0

Examples

1. Děti se s některými záležitostmi nechtějí svěřit rodičům, i když žijí v normálně fungující rodině. […] Dnes mají **také** mnozí rodiče méně času na své ratolesti než dřív. (= Children do not want to confide certain matters to their parents, even if they live in a normally functioning family. [...] Today, many parents have **also** less time for their children than before.)
2. Povinností budoucího nájemce tohoto areálu o rozloze 103 tisíc metrů čtverečních bude mj. péče o všechny nemovitosti včetně jejich údržby a oprav. Nájemce bude **také** muset vyřešit podmínky parkování pro návštěvníky tržnice a splnit podmínky Pražského ústavu památkové péče při úpravách objektů vzhledem k tomu, že jde o kulturní památku. (= It will be the duty of the future tenant of this zone with an area of 103,000 meters square, among others, to take care of all properties, including their maintenance and repairs. The tenant will **also** have to solve parking conditions for visitors of the market and to meet the conditions of the Prague National Heritage Institute when rebuilding objects due to the fact that they belong to the cultural heritage.)

3. Podle Mandíkových slov lze komerčně využít zhruba deset hektarů pozemků v železniční stanici Praha- Žižkov. Využít lze **také** prostory stanice Praha-Smíchov. (=According to Mandík, about ten hectares of land in the railway station Prague-Žizkov can be used commercially. **Also** a space of station Praha-Smíchov can be used.)

4. Vyrábějí se zde především tresti do lihovin, limonád, sirupů a pečiva. Firma **také** produkuje cukrářské pasty. (= There is a production of particular essences for spirits, soft drinks, syrups and pastries in this factory. The company **also** produces confectionery pastes.)

5. V okolí Brna a Kyjova se hojně vyskytují muchomůrky zelené. **Také** v Hostivaři a v dalších pražských lesoparcích byl nyní výskyt této houby zaznamenán. (= In the vicinity of Brno and Kyjov, toadstools occur in plenty. **Also** in Hostivař and other forest parks of Prague, the occurrence of these fungi has now been recorded.)

6. Taková odměna může mít skutečně silný motivační účinek pro účastníky **a** může být **také** užitečným přínosem pro firmu, která náklady plně hradí. (= Such a reward may indeed have a strong incentive effect on the participants **and** can **also** be a useful asset for a company that fully pays the costs.)

Annotation procedure

Contrary to the majority of similarly aimed corpus projects (e.g. the above mentioned Penn Discourse Treebank 2.0, [32]), the discourse-related information has been directly on the syntactic trees and technically it is a part of the underlying syntactical – tectogrammatical layer of the PDT. This methodological approach allows us to include discourse-relevant syntactic phenomena annotated earlier (such as e.g. discourse relations expressed by dependent clauses) into the discourse representation, and to take advantage of the syntactic structure itself (resolution of elliptical structures, parentheses, appositions etc.). Also, from the perspective of querying the treebank and visualizing, all the different types of linguistic information ranging from morphology to discourse phenomena are interlinked and available/visible at once.

The annotation procedure consisted of two steps. In the first step, all inter-sentential relations (relations between sentences) and a small part of intra-sentential relations (relations in one sentence) were annotated manually. Intra-sentential relations were only annotated manually in cases when their discourse semantics differed from the tectogrammatical interpretation. In the second step, the remaining intra-sentential relations (the interpretation of which on the tectogrammatical layer was adequate for discourse-level analysis) were automatically extracted and mapped onto the discourse annotation. The automatic part of the annotation was based on extracting relevant information (presence of the relation, scope of the arguments, the connective(s), and a discourse-semantic label) from the deep-syntactic layer of PDT. Both parts of the annotation (the manual and the automatic subparts) underwent consistent checking procedures (see [29]). Table 4.1 shows the distribution of semantic types of discourse relations in PDT 3.0 and the proportion of their intra- and inter-sentential realizations.

In the manual part of annotation, the annotators proceeded from analyzing the raw text (identifying a connective) to marking the discourse relations and all their properties directly on the tectogrammatical trees. A discourse relation between subtrees is marked with a thick orange arrow; the type of the relation is displayed next to the tectogrammatical (deep-syntactic) lemma of the starting node ($reason_n$ in Table 4.1). The connective(s) assigned to the relation shows in green (*Therefore* in Figure 4.1).

For more information on the annotation process see the annotation manual [30].

List of discourse-related annotation attributes in PDT 3.0

Discourse-related annotation is captured mostly in a structured attribute *discourse* at the start node of the relation, additional annotation is captured in attributed *discourse_groups* and *discourse_special*.

- ***discourse/target_node.rf*** – id of the target node, or undefined if there is no target node (e.g. no hypertheme in a list structure)
- ***discourse/type*** – the type of arrow, two possible values: *discourse* (discourse relation), *list* (list entry)
- ***discourse/start_range*** – start range of a discourse arrow; possible values: *n*, where n (non-negative integer) = number of trees to the right of the actual tree belonging to the argument in addition to the node and its subtree (*0* means just the node and its subtree), *group* (an arbitrary set of nodes; see below attributes *discourse/start_group_id* and *discourse_groups*), *forward* (means the node with its subtree plus a non-specified number of the following trees), *backward* (means node with its subtree plus a non-specified number of the preceeding trees)
- ***discourse/target_range*** – target range of a discourse arrow; possible values above
- ***discourse/start_group_id*** – identifier of a group of nodes (positive integer) where the start_range of the arrow is set to "group"; individual nodes belonging to the group keep the group identifier in the attribute *discourse_groups*
- ***discourse/target_group_id*** – identifier of a group of nodes (positive integer) where the target_range of the arrow is set to "group"; individual nodes belonging to the group keep the group identifier in the attribute *discourse_groups*
- ***discourse/discourse_type*** – type of discourse semantic relation, such as *cond* (textual condition); possible values are in the column Abbreviation in tab. 4.1
- ***discourse/t-connectors.rf*** – list of ids of nodes from the tectogrammatical layer that represent a discourse connective
- ***discourse/a-connectors.rf*** – list of ids of nodes from the analytical layer that represent a discourse connective
- ***discourse_groups*** – list of identifiers of groups the given node belongs to
- ***discourse_special*** – three possible values for three special roles of the phrase represented by the node and its subtree: *heading* (replaces attribute *is_heading* from PDiT), *metatext* and *caption*.

Revisions and corrections of the PDiT 1.0 data

Several revisions and corrections have been done compared to the published PDiT 1.0 data:

- corrections of some original arrows in cases where there was a manual (correct) and an automatic (wrong) arrow,
- unnecessary groups removed (error in the removing script),
- correction of values of the attributes *start/target_range* and *start/target_group_id* in cases of removed groups,
- change of direction of automatic intra-sentential arrows derived from the functor CSQ
  some fixes of individual errors.

Visualisation

Figure 4.1 shows the annotation of a discourse relation between the sentences shown above in Example 1. The arrow has assigned a semantic label reason representing the relation of reason and result, with the associated connective *proto* (= *therefore*). Also, the range of the arguments entering the relation is set (*range: 0 -> 0*). In this case, only the two mentioned trees (sentences) enter the relation.

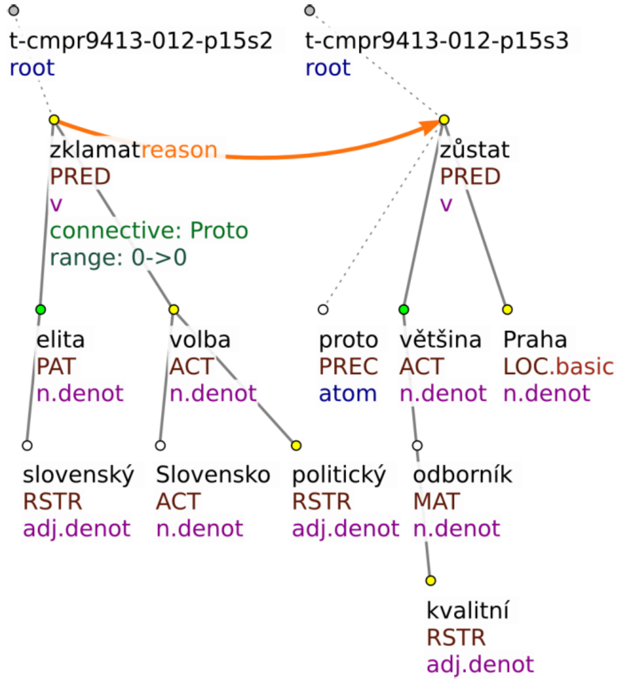

Figure 4.1: *Slovenská elita byla zklamána politickou volbou Slovenska. Proto většina kvalitních odborníků zůstala v Praze*.(= The Slovak elite were disappointed by the political choice of Slovakia. Therefore, most of the quality specialists stayed in Prague.)

5 Genre specification

Description

The PDT data originate from two big Czech daily newspapers (Mladá Fronta, Lidové Noviny), one business weekly (Českomoravský profit) and one scientific journal (Vesmír). As such the corpus can be viewed as journalistic. During various annotation projects, however, we experienced a considerable diversity of the data – the corpus contains in fact texts ranging from TV programs to cultural reviews and also some number of incoherent texts like short news collections.

The manual classification of PDT texts (3,165 documents) according to their genre or text style, newly included in its 3.0 release, should serve the following purposes:

- exclude short and incoherent texts from training sets for modeling of any type of coherence
- more efficient clustering of similar texts types (or ways of text composition) for any NLP experiments, mostly for those working with sentences and larger units (anaphora resolution, text topics and salience, discourse processing, sentiment analysis etc.)
- obtaining gold data for automatic genre/text type clustering

The genre of a document is captured in an attribute *genre* attached to the whole document. Possible values can be found in the column Abbreviation in Table 5.1.

For the related literature, see the publications [30], [31] and [33] in the list of references at the end of this document.

Annotation procedure

Using the previous experience of PDT annotators, we created taxonomy of 20 genre categories in three main classes: monological genres, dialogical genres and other, marginal genres (see Table 5.1). To keep the annotation task as simple as possible, the taxonomy is flat. Also, we only assigned one label to each document, even though the labels combine some formal and content features (e.g. *interview* – deciding is the formal structure and *sports* – deciding is the content. So, for instance, for an interview with an athlete, a label for the prevailing genre is used: if the whole discourse is an interview, it is marked as such; if it is rather a sports report with an embedded short interview with an athlete, it is treated as sports news).

The automatic preannotation of genres used information from the manual annotation of discourse relations, where the annotators had marked corpus documents consisting of a set of short unrelated texts possibly of different genres (these were preannotated as *collections*) and also sentences representing photo, chart or table captions (documents consisting of only one sentence marked previously as captions were preannotated with the genre *caption*). of the remaining documents was performed by eight annotators. 1/5 of the corpus (development test data and evaluation test data) was annotated in parallel by two annotators. Discrepancies were then solved by an arbiter. In case of a substantial disagreement, the problematic genres were checked by the arbiter in all data annotated by the annotators in question.

Genre types in PDT 3.0		
Type	**Abbreviation**	**Description**
Monological genres		
critical review	***review***	of books, films, exhibitions, concerts, theatre etc.
invitation	***invitation***	to concerts, exhibitions, etc.
letters from readers	***letter***	
advice column	***advice***	advice, interpretation of a phenomenon, or instructions (how to report a crime, school of chess, answering letters from readers)
cultural program	***program***	of TV, radio, exhibitions etc.

film/TV program plot description	***plot***	a plot of a film or a TV program
sports news	***sport***	+ sports results
comment	***comment***	Commentary on an actual topic (shorter range), expresses a subjective view
news report	***news***	report on something current, no assessment, includes business results etc.
reflection essay	***essay***	larger report / comment, longer range, more subjective, some current or general topic
overview	***overview***	list of currency rates etc.
description	***description***	of a product, company, services etc.
weather forecast	***weather***	
readers' survey + results	***survey***	survey and its results
Dialogue		
topical interview	***topic_interv***	"actual conversation", i.e. an interview with an expert on a hot topic
interview with a personality	***person_interv***	contains multiple topics, readers are primarily informed about the personality
Other		
collection	***collection***	collection of various texts in one document
caption of photo, table, etc.	***caption***	Descriptions of pictures, graphs, tables etc.
metatext	***metatext***	text resulting from an error in corpus processing
other	***other***	genre is uncertain - especially in isolated sentences

Table 5.1: Genre types in PDT 3.0

Examples

ln95046_021.t., *genre* = *sport*

1. Další Jágrova branka
2. New York -
3. Český hokejista Jaromír Jágr vsítil svůj čtrnáctý gól této sezóny NHL a rozhodl jím o výsledku utkání Pittsburgh - Quebec (5:4).
4. Závěrečná třetina byla nesmírně dramatická, padlo v ní šest branek, přičemž poslední slovo měl právě Jágr, který rozhodl zápas pouhých 22 sekund poté, co Nolan z Quebeku srovnal skóre na 4:4.
5. Po čtyřzápasové pauze zaviněné chřipkou nastoupil i Martin Straka a vstřelil jeden gól.
6. V Miami podlehla Florida týmu New Yorku Rangers 3:5.
7. Za stavu 3:3 v závěrečné třetině prolomil nerozhodný stav Karpovcev, když puk z jeho hokejky skončil po odrazu v soupeřově brance.
8. O konečném vítězství Jezdců 5:3 rozhodl Olczyk.

Translation:

1. Jagr scores again
2. New York –
3. The Czech hockey player Jaromir Jagr scored his fourteenth goal of the NHL season and so decided the result of the match Pittsburgh – Quebec (5:4).
4. The final third was extremely dramatic, six goals were scored, and it was Jagr who had the last word and decided the match just 22 seconds after Nolan from Quebec leveled the score at 4:4.
5. After being absent for four matches because of flu, Martin Straka joined the match and scored a goal.
6. In Miami, Florida was defeated by New York Rangers 3:5.
7. In the state of 3:3 in the final third, Karpovcev broke the tie when the puck from his stick ended up after a bounce in the opponent's net.
8. The final 5:3 victory of the Rangers was decided by Olczyk.

ln95045_056.t, *genre* = *collection*

1. Krátce
2. Návrhy britského premiéra J. Majora a jeho irského partnera J. Burtona na budoucí uspořádání Severního Irska získaly včera podporu britské vlády.
3. Dokument se stane v příštích týdnech předmětem diskusí konstitučních severoirských politických stran.
4. Dvěma hlavními cíli české zahraniční politiky jsou členství v Evropské unii a Severoatlantické alianci, řekl včera český ministr zahraničí Josef Zieleniec ve výboru pro zahraniční věci a zahraniční obchod Poslanecké sněmovny kanadského parlamentu.
5. Dohodu o zastavení palby porušil další ozbrojený konflikt mezi armádou a povstaleckou organizací UNITA, ke kterému došlo u severoangolského města Uige.
6. Irácká vláda nadále v "děsivé" míře a "bez jakýchkoli známek zlepšení" pošlapává lidská práva, konstatuje zvláštní zpravodaj OSN pro Irák Max van der Stoel ve zprávě, která byla včera zveřejněna v ženevském sídle OSN.
7. Zatím nelze říci, kdy bude sestavena nová polská vláda, řekl po setkání představitelů polské vládní koalice, Polské lidové strany a Svazu demokratické levice koaliční kandidát na křeslo premiéra, maršálek Sejmu J. Oleksy.

Translation:

1. Briefly
2. Yesterday, the proposals of the British Prime Minister J. Major and his Irish partner J. Burton on the future organization of Northern Ireland received the support of the British government.
3. The document will be a point of discussions of constitutional Northern Irish political parties.

4. The two main goals of the Czech foreign policy is the membership in the European Union and in NATO, said yesterday the Czech Minister of Foreign Affairs Josef Zieleniec in the Committee of Foreign Affairs and Foreign Trade Chamber of Deputies of the Parliament of Canada.
5. Another armed conflict between the army and the rebel organization Unita, which occurred at north Angola city of Uige, broke the agreement on ceasefire.
6. The Iraqi government keeps in "appalling" extent and "without any signs of improvement" trampling on human rights, says UN special reporter for Iraq, Max van der Stoel in his report, which was published yesterday at the Geneva UN headquarters.
7. So far, it cannot be said when new Polish government would be formed, said the coalition candidate for the seat of Prime Minister Marshal of the Sejm J. Oleksy after a meeting of representatives of Polish government coalition, the Polish People's Party and the Democratic Left Alliance.

ln94211_77.t, *genre = caption*

1. Bývalého generála sovětského strategického letectva nezapře Džochar Dudajev vzorně salutující na slavnostní přehlídce uspořádané při příležitosti třetího výročí vyhlášení nezávislosti Čečenska na Rusku
2. Foto Reuter

Translation:

1. Dzhokhar Dudayev cannot deny being a former general of the Soviet Strategic Air, saluting perfectly at the festive parade organized on the occasion of the third anniversary of the declaration of independence of Chechnya from Russia.
2. Photo Reuter

6 Multiword expressions

Description

All the multiword expressions (MWEs) 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*** - names of books, units of measurement, biological names of plants and animals
- ***address*** - an 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.
For the related literature, see the publications [34], [35] and [36] in the list of references at the end of this document.

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. Table 6.1 below shows how much data was annotated by 1, 2, or 3 annotators in parallel, compared to the size of PDT (t-data).

Anotated data					
Parallel annotation	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%

Table 6.1: Annotation of the the multiword expressions

The data produced by individual annotators is not part of PDT 2.5, but it is freely available at the project web page (http://ufal.mff.cuni.cz/lexemann/mwe/) (http://ufal.mff.cuni.cz/lexemann/mwe/ (http://ufal.mff.cuni.cz/lexemann/mwe/)). 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

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

   * – *date*, *basic_form* "15. července"
   ** – *location*, *basic_form* "Pražský hrad"
   *** – *institution*, *basic_form* "Ústavní soud"

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

   * – *lexeme*, *basic_form*"ústavní soudce"
   ** – *lexeme*, *basic_form* "politická strana"

7 Valency lexicon PDT-Vallex 3.0

Description

Along with the corpus PDT 3.0, there is a new version of valency lexicon PDT-Vallex 3.0. Lexicon PDT-Vallex occurs in parallel with semantic-syntactic annotation of sentences, contains almost exclusively the verbs and their meanings that occurred in the annotated data, i.e. those whose valency frames annotator had to know to be able to correctly annotate the individual obligatory and optional valency modifications in the annotated sentence. The first version of the lexicon PDT-Vallex (version 1.0) was established during the annotation of the corpus PDT 2.0. The lexicon was further extended under other annotation projects.

Extension of valency lexicon

Firstly, the lexicon was extended by the annotation of the Czech part of the Prague Czech-English Dependency Treebank (Hajič et al., 2011; further PCEDT 2.0; the abbreviation of *The Prague Czech-English Dependency Treebank 2.0*, [37]). The corpus PCEDT 2.0 includes articles from the Wall Street Journal (1989) that were translated into Czech for the Czech part of the corpus. There are mostly texts with economic issues. PDT-Vallex was thus widely extended by verbs and meanings of this area (e.g. *nakonfigurovat, podhodnocovat, porcovat medvěda, prát peníze, segmentovat trh, seškrtnout finanční prostředky, srovnat se s* riziky (= configure, underestimate, carve a bear, launder money, market segment, reduce funds, conciliate the risks).
Another great extension of the lexicon was due to the annotation of the Prague Dependency Treebank of Spoken Czech (the PDTSC 2.0, the abbreviation for *The Prague Dependency Treebank of Spoken Czech 2.0*, [38]). The corpus PDTSC 2.0 contains recordings of two types: (i) the Czech part of the corpus that was created as a part of the international project Malach (slightly moderated conversations with the people who survived the holocaust) and (ii) the dialogues that were recorded within the project Companions (the theme of the dialogue is a conversation about personal collection of pictures of one of the participants in the dialogue). The valency lexicon was enriched by the lemmas from the field of general (family) life as *háčkovat, houbařit, koledovat, pošťuchovat se, přebalit dítě, přivdat se, sáňkovat, zavařovat* (= crochet, pick mushrooms, go carol-singing, nudge, change a nappy, marry into, sled, conserve), but also lemmas from authentic testimonies of Holocaust survivors as *proválčit, vybombardovat, odvlíknout, přežít, srocovat se* (= make war, bomb out, abduct, survive, mob).
Under a new annotation of PDT 2.0 data that are now published as a corpus PDT 3.0, there were only little modifications in the valency lexicon. The biggest change was addition of a new frame for verbonominal predicates (*be* + adjective, noun) whose infinitive in the position of actor is controlled by benefactor dependent on the nominal part of the predicate: ACT(.f;aby[.v];že[.v]) PAT(.a1;.a7;.d); e.g.: *Je možné odejít. Je možno odejít.* (= in both cases: It is possible to go away.) *Je pro nás*.BEN *důležité přijít včas.* (= It is important for us.BEN to come on time.) The frame was assigned to 456 verbonominal predicates whose nominal part are lemmas *možný*, *nutný*, *možno*, *nutno* (= possible, necessary). In the next phase of work, the list of adjectives in this function (PAT) will be extended by other types (e.g. *obtížný*, *snadný*, *zajímavý*, *ideální* (= difficult, easy, interesting, ideal) etc.)
Table 7.1 introduces the individual extensions of the valency lexicon expressed in numbers. After the annotation of the corpus PDT 2.0, the valency lexicon contained 5,510 verbal lemmas and 9,191 valency frames. Annotation of other corpora that are more or less comparable with the corpus PDT 2.0 (corpus PCEDT 2.0 contains almost the same number of sentences, but these sentences are longer on average; corpus PDTSC 2.0 then contains a large number of short sentences with many verbs) extended the lexicon always with approximately 1,500 new lemmas and 2,500 new valency frames. (Yet) the latest version of the lexicon contains nearly 8,500 verbal lemmas and 14,500 valency frames.

Number of		**PDT 2.0**	**PCEDT 2.0**	**PDTSC 2.0**
Data	tokens	833,195	1,151,150	742,221
	sentences	49,431	49,208	73,835
	verbal tokens	88,103	118,035	125,271
	assigned lemmas	5,376	4,880	4,628
	assigned frames	7,674	8,285	7,582
Lexicon	lemmas in the lexicon	5,510	7,104	8,459
	frames in the lexicon	9,191	11,933	14,517
		PDT-Vallex 1.0	**PDT-Vallex 2.0**	**PDT-Vallex 3.0**

Table 7.1: Extension of valency lexicon

The annotation of non-standard phenomena in the valency lexicon

Annotation of spoken corpus PDTSC 2.0 required a new modifications in the inscription of valency lemmas. A percent sign (%) was established to indicate different degrees of non-standard phenomena. This sign can be used in the following contexts:

- following a lemma where it indicates non-standard lemma. One sign of % is used for colloquial, expressive or otherwise “strange” lemmas (ex. (1)). Two signs of % are for vulgar lemmas (ex. (2)).
- following the whole frame. Here % denotes non-standard verbal frame, some less usual meaning of the given verbal lemma (ex. (3)). Two signs of % are uses for vulgar verbal meanings (ex. (4)).
- following a sign for the function of a valency member where it indicates non-standard valency member that is usually not used in that meaning and therefore it seems inappropriate (ex. (5)).
- following the form where it signals non-standard formal realization of the given valency member that is usually not used and that would be stylistically inappropriate in a written text (ex.(6)).

Different contexts of using % may be combined within the inscription of a valency lemma. The sign % in both lemma and valency frame captures the cases when one of the valency frames represents a marked meaning for a colloquial form (e.g. *píct (= bake))* of an otherwise ordinary standard verb (*péci* (= bake)), whereas the other valency frames represent unmarked meanings (ex. (7)).

Examples

1. čumět (= stare) **%** ACT(.1) DIR3(*) Čuměla dvě hodiny na obraz. (= She was staring at the picture for two hours.)

2. chlastat (=hit bottle) **%%** ACT(.1) PAT(.4) Začal chlastat alkohol. (=He hit the bottle)

3. bruslit (=be at sea) ACT(.1) PAT(v+6) **%** Bruslil jsem v chemii.(= When it came to chemistry, I was all at sea.)

4. držet (=shut up) ACT(.1) DPHR (hubu) **%%** Drž hubu! (= Shut up!)
5. dobýt (= conquer) ACT (.1) PAT(.4) ?ORIG**%**(od+2) Angličané dobyli Palestinu od Turků.(= The English conquered Palestine from the Turks.)

6. dráždit (= irritate) ACT (.1) ADDR(.4) ?PAT(k+3;na+4**%**) Dráždí mě to na kašel. (= It irritates me cough.)

7. píct (=go out/bake) **%**

   ACT(.1) PAT(s+7) **%** Mohl bych píct s jinou. (= I could go out with another.)
   ACT(.1) PAT(.4) Budeme píct koláče. (= We will bake cakes.)

Related literature – see publications [39], [40] and [41] in the list of references at the end of this document.

B. Modifications and complements on analytical layer

8 Clause segmentation

Description

Analytical trees in PDT 3.0 (originally 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 3.0 was generated by this procedure; the original manually annotated samples are not shipped with PDT 3.0.
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.

Visualisation

Clause segmentation can be comfortably visualized in TrEd (see Figure 8.1). The new extension for viewing PDT 3.0 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.

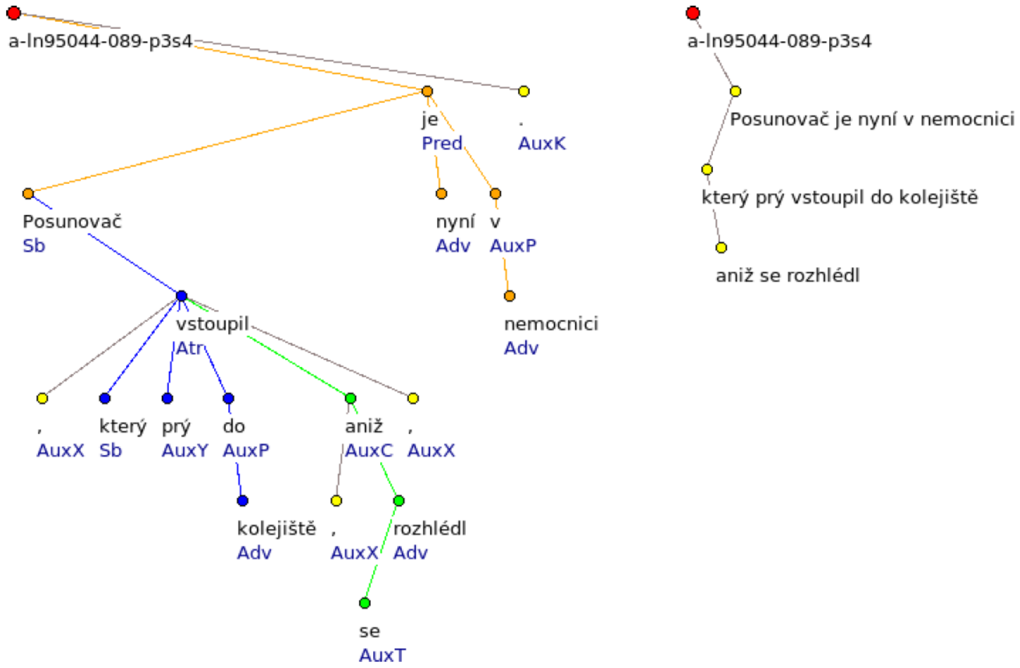

Figure 8.1: Sentence '*Posunovač, který prý vstoupil do kolejiště, aniž se rozhlédl, je nyní v nemocnici.*' represented by two trees: full (and colored) on the left side and with collapsed clauses on the right side

Examples

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

   * – clause boundary, coordinating conjunction joining two clauses

   ** – final punctuation, sentence boundary

2. 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

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

   * – subordinating conjunction

4. 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

We follow the concepts thoroughly formulated in [42] 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 3.0 and all the data is consistently annotated automatically.
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.

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 3.0 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 8.2 for clause counts per sentence and Figure 8.3 for the most common sentence structure *patterns*.

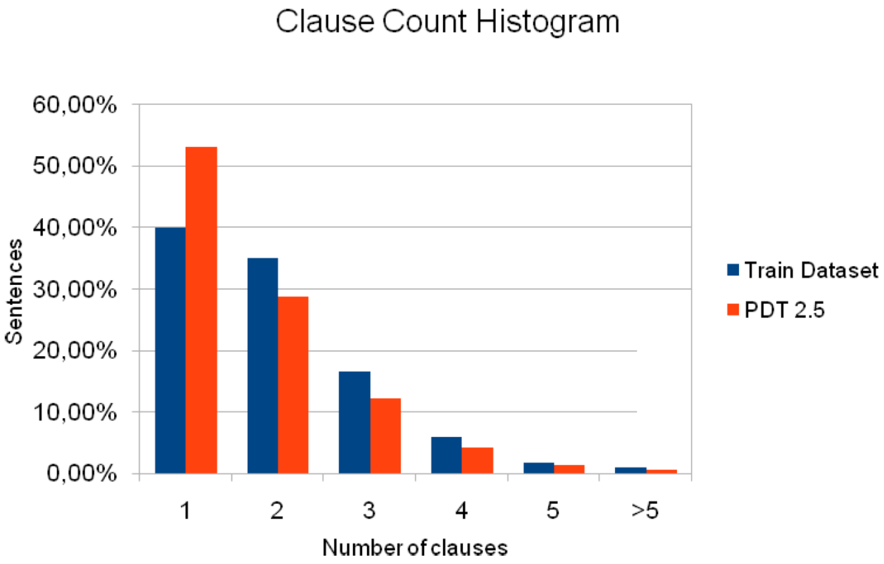

Figure 8.2: Clause Count Histogram

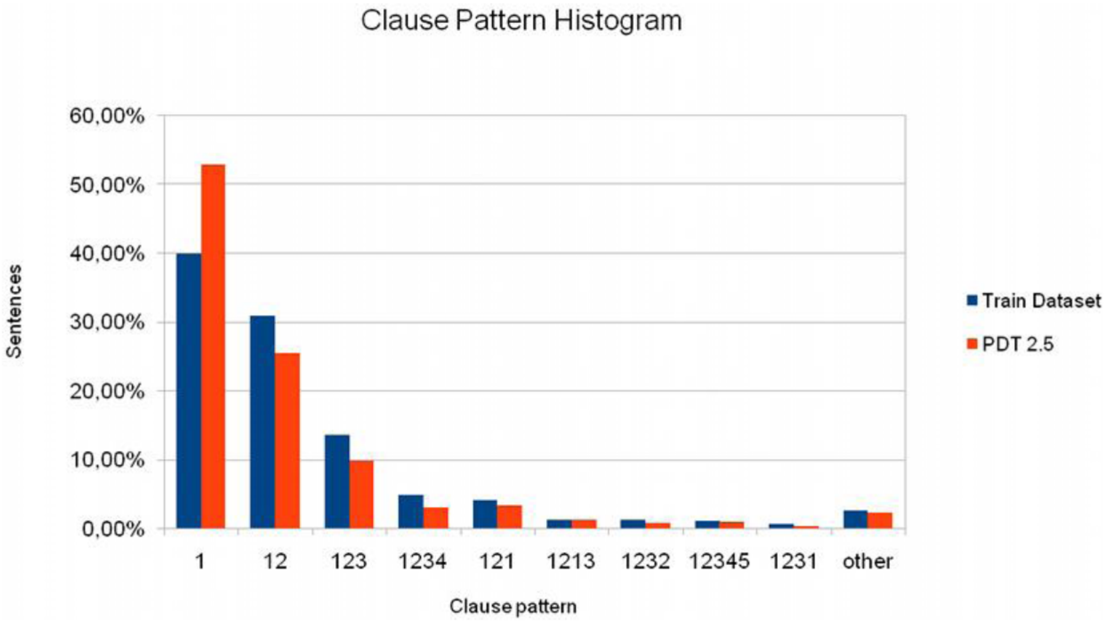

Figure 8.3: For the sake of brevity, clauses are numbered by single digits. For example, the pattern "12" stands for a complex sentence formed by two clauses, the pattern "121" also represents a two-clause sentence but with the second clause embedded, etc.

References

[1] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., Zikánová, Š.: *Prague Dependency Treebank 3.0*. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, 2013. (http://ufal.mff.cuni.cz/pdt3.0/ (http://ufal.mff.cuni.cz/pdt3.0/))

[2] Bejček, E., Panevová, J., Popelka, J., Smejkalová, L., Straňák, P., Ševčíková, M., Štěpánek, J., Toman, J., Žabokrtský, Z., Hajič, J.: *Prague Dependency Treebank 2.5.* Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, 2011. (http://ufal.mff.cuni.cz/pdt2.5/ (http://ufal.mff.cuni.cz/pdt2.5/))

[3] Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., Žabokrtský, Z.: Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, Coling 2012 Organizing Committee, Mumbai, India, pp. 231-246, 2012 (pdf (/pdt3.0/doc/03_COLING_2012.pdf)).

[4] Hajič et al.: *Prague Dependency Treebank 2.0.* Data/software, Linguistic Data Consortium, Philadelphia, PA, USA, 2006. ISBN 1-58563-370-4 (http://www.ldc.upenn.edu (http://www.ldc.upenn.edu))

[5] Mikulová et al.: *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Annotation manual.* Technical report no. TR-2006-30, Univerzita Karlova v Praze, MFF, ÚFAL, Praha, 2005. ISSN 1214-5521 (html (http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/index.html), pdf (/pdt3.0/doc/05_tr30.pdf)).

[6] Poláková, L., Jínová, P., Zikánová, Š., Hajičová, E., Mírovský, J., Nedoluzhko, A., Rysová, M., Pavlíková, V., Zdeňková, J., Pergler, J., Ocelák, R.: *Prague Discourse Treebank 1.0.* Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, 2012. (http://ufal.mff.cuni.cz/pdit/ (http://ufal.mff.cuni.cz/pdit/))

References related to *1.1 Grammateme typgroup*

[7] Panevová, J. – Ševčíková, M.: Delimitation of information between grammatical rules and lexicon. In: *Proceedings of the International Conference on Dependency Linguistics (Depling 2011)*, Universitat Pompeu Fabra, Barcelona, 2011, pp. 173–182. ISBN 978-84-615-1834-0 (pdf (/pdt3.0/doc/07_PUBLISHED.pdf)).

[8] Panevová, J. – Ševčíková, M.: Jak se počítají substantiva v češtině: poznámky ke kategorii čísla. *Slovo a slovesnost*, 72, 2011, pp. 163–176. ISSN 0037-7031 (pdf (/pdt3.0/doc/08_SaS-72-2011-3_Panevova-Sevcikova_163-176.pdf)).

[9] Ševčíková, M. – Panevová, J. – Smejkalová, L.: Specificity of the number of nouns in Czech and its annotation in Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 96, pp. 27–47, 2011. ISSN 0032-6585(pdf (/pdt3.0/doc/09_MAIN.pdf)).

[10] Ševčíková, M. – Panevová, J. – Žabokrtský, Z.: Grammatical number of nouns in Czech: linguistic theory and treebank annotation. In: *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT 2010)*, NEALT Proceedings Series, Vol. 9. Tartu, Estonia, 2010, pp. 211–222. ISSN 1736-8197 (pdf (/pdt3.0/doc/10_tlt9_submission_10.pdf)).

References related to *1.2 Grammateme factmod*

[11] *Mluvnice češtiny II.* Academia, Praha, 1986.

[12] Panevová, J. – Ševčíková, M: Annotation of Morphological Meanings of Verbs Revisited. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta: ELRA, 2010, pp. 1491–1498. ISBN 2-9517408-6-7 (pdf (/pdt3.0/doc/12_395_Paper.pdf)).

[13] Ševčíková, M.: *Funkce kondicionálu z hlediska významové roviny.* UFAL MFF UK, Praha, 2010, 179 pp. ISBN 978-80-904175-2-6.

[14] Ševčíková, M.: The meaning of the conditional mood within the tectogrammatical annotation of Prague Dependency Treebank 2.0. In: *Proceedings of the Slovko 2009 Conference: NLP, Corpus Linguistics, Corpus Based Grammar Research*. Bratislava: Slovenská akadémia vied, pp. 321–330, 2009. ISBN 978-80-7399-875-2 (pdf (/pdt3.0/doc/14_article.pdf)).

[15] Ševčíková, M.: Kondicionál přítomný jako součást explicitních performativních formulí. *Korpus – gramatika – axiologie*, Vol. 1, No. 1, pp. 41–62, 2010. ISSN 1804-137X (pdf (/pdt3.0/doc/15_Sevcikova_KGA.pdf)).

References related to *1.3 Grammateme diatgram*

[16] Panevová, J.: O rezultativnosti (zejména) v češtině. In: *Gramatika i leksika u slovenskim jezicina*. Novi Sad, Beograd: Matica Srbska, Institut za srpski jezik. pp. 165 – 176, 2011.

[17] Panevová, J. – Ševčíková, M.: Delimitation of Information between Grammatical Rules and Lexicon. In: *Linguistic Aspects of Dependency* (Wanner, L., Gerdes, K, eds.). John Benjamins Publ. House, Amsterdam/the Netherland, pp.1- 20, 2013.

[18] Panevová, J. – Ševčíková, M.: The Role of Grammatical Constraints in Lexical Komponent in Functional Generative Description. In: *Proceedings of the 6th International Conference on Meaning-Text Theory*. Praha, pp. 134-143, 2013 (pdf (/pdt3.0/doc/18_Panevova-Sevc_MTT13.pdf)).

References related to *1.4 The sentmod attribute*

[19] Ševčíková, M. – Mírovský, J.: Sentence Modality Assignment in the Prague Dependency Treebank. In: *Proceedings of the 15th International Conference Text, Speech and Dialogue (TSD 2012)*. Springer, Berlin, pp. 56–63, 2012. ISBN 978-3-642-32789-6, ISSN 0302-9743 (pdf (/pdt3.0/doc/19_tsd391a.pdf)).

References related to *2 Modification of the annotation of lemma #Benef*

[20] Panevová, J.: „Být posel dobrých zpráv je mi příjemné“ (Několik poznámek k infinitivním konstrukcím). In: *Karlík a továrna na lingvistiku. Prof. Petru Karlíkovi k 60. Narozeninám*. (eds. A. Bičan, J. Klaška, P. Macurová, J. Zmrzlíková). Host/Masarykova univerzita, Brno, s. 345 – 354, 2010.

[21] Panevová, J.: On Syntax and Semantics of Czech Infinitival Constructions: A Case Study. In: *Slovo i jazyk. Sbornik statej k vosmidesjatiletiju akademika Ju. D. Apresjana.* Jazyki slavjanskich kul'tur, Moskva, pp. 541 – 551, 2011.

[22] Panevová, J.: Infinitiv ve funkci atributu. In: *Kapitoly z české gramatiky* (ed. F. Štícha), Academia., Praha, s. 945 – 960, 2011.

[23] Panevová, J. a kol.: *Mluvnice spisovné češtiny 2. Syntax na základě anotovaného korpusu (kap. 5).* Karolinum, Praha (in press).

References related to *3 Coreference and bridging relations*

[24] Nedoluzhko, A.: *Rozšířená textová koreference a asociační anafora. Koncepce anotace českých dat v Pražském závislostním korpusu*. UFAL MFF UK, ISBN 978-80-904571-2-6, Praha, 2011.

[25] Nedoluzhko, A., Mírovský, J.: *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank. Annotation manual.* Technical report No. 44, UFAL MFF UK, Prague, 2011 (pdf (/pdt3.0/doc/25_tr44.pdf)).

[26] Nedoluzhko, A., Mírovský, J.; Novák, M.: A Coreferentially annotated Corpus and Anaphora Resolution for Czech. In: *Computational Linguistics and Intellectual Technologies*. ABBYY, Moscow, Russia, pp. 467-475, 2013. ISBN 978-1-937284-58-9 (pdf (/pdt3.0/doc/26_Dialog2013.pdf)).

[27] Nedoluzhko, A.: Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In: *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*. Omnipress, Inc, Sofia, Bulgaria, pp. 103-111, 2013. ISBN 978-1-937284-58-9 (pdf (/pdt3.0/doc/27_W13-2313.pdf)).

[28] Nedoluzhko, A., Mírovský, J.: How Dependency Trees and Tectogrammatics Help Annotating Coreference and Bridging Relations in Prague Dependency Treebank. In: *Proceedings of the Second International Conference on Dependency Linguistics, Depling 2013*. Matfyzpress, Charles University in Prague, Prague, pp. 244-251, 2013. ISBN 978-80-7378-240-5 (pdf (/pdt3.0/doc/28_W13-3727.pdf)).

References related to *4 Discourse relations* and *5 Genre specification*

[29] Jínová, P., Mírovský, J., Poláková, L.: Analyzing the Most Common Errors in the Discourse Annotation of the Prague Dependency Treebank. In: *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, Edicoes Colibri, Lisboa, Portugal, pp. 127-132, 2012 (callto:127-132,%202012). ISBN 978-989-689-274-6 (callto:978-989-689-274-6) (pdf (/pdt3.0/doc/29_Proceedings_121-126.pdf)).

[30] Poláková, L., Jínová, P., Zikánová, Š., Bedřichová, Z., Mírovský, J., Rysová, M., Zdeňková, J., Pavlíková, V., Hajičová, E.: *Manual for Annotation of Discourse Relations in Prague Dependency Treebank.* Technical report no. 2012/47, UFAL MFF UK, Praha, 2012 (pdf (/pdt3.0/doc/30_tr47.pdf)).

[31] Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., Hajičová, E.: Introducing the Prague Discourse Treebank 1.0. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Asian Federation of Natural Language Processing, pp. 91-99, 2013. ISBN 978-4-9907348-0-0 (callto:978-4-9907348-0-0)(pdf (/pdt3.0/doc/31_IJCNLP011.pdf)).