Annotation of Grammatemes in the Prague Dependency Treebank 2.0

Magda Razímová, Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics, Charles University, Prague
Malostranské náměstí 25, Prague 1, 118 00, Czech Republic
{razimova,zabokrtsky}@ufal.mff.cuni.cz

Abstract

In this paper we report our work on the system of grammatemes (mostly semantically-oriented counterparts of morphological categories such as number, degree of comparison, or tense), the concept of which was introduced in Functional Generative Description, and has been recently further elaborated in the layered annotation scenario of the Prague Dependency Treebank 2.0. We present also a hierarchical typology of tectogrammatical nodes, which is used as a formal means for ensuring presence or absence of respective grammatemes.

1. Introduction

Human language, as an extremely complex system, has to be described in a modular way. Many linguistic theories attempt to reach the modularity by decomposing language description into a set of layers, usually linearly ordered along an abstraction axis (from text/sound to semantics/pragmatics). One of the common features of such approaches is that word forms occurring in the original surface expression are substituted (for the sake of higher abstraction) with their lemmas at the higher layer(s). Obviously, the inflectional information contained in the word forms is not present in the lemmas. Some information is 'lost' deliberately and without any harm, since it is only imposed by government (such as case for nouns) or agreement (congruent categories such as person for verbs or gender for adjectives). However, the other part of the inflectional information (such as number for nouns, degree for adjectives or tense for verbs) is semantically indispensable and must be represented by some means, otherwise the sentence representation becomes deficient (naturally, the representations of sentence pairs such as '*Peter met his youngest brother*' and '*Peter meets his young brothers*' must not be identical at any level of abstraction). At the tectogrammatical layer of Functional Generative Description (FGD, (Sgall, 1967), (Sgall et al., 1986)), which we use as the theoretical basis of our work, these means are called grammatemes.[1]

The theoretical framework of FGD has been implemented in the Prague Dependency Treebank 2.0 project (PDT 2.0, (Hajičová et al., 2001)), which aims at a complex annotation of large amount of Czech newspaper texts. Although grammatemes are present in the FGD for decades, in the context of PDT they were paid for a long time a considerably less attention, compared e.g. to valency, topic-focus articulation, or coreference. However, in our opinion grammatemes will play a crucial role in NLP applications of FGD and PDT (e.g., machine translation is impossible without realizing the differences in the above pair of example sentences). That is why we decided to further elaborate the system of grammatemes and to implement it in the PDT 2.0 data. This paper outlines some of the results of more than two years of the work on this topic.

The paper is structured as follows: after introducing the basic properties of the PDT 2.0 with focus on the tectogrammatical layer in Section 2., we will describe the classification of t-layer nodes in Section 3., enumerate and exemplify the individual grammatemes and their values in Section 4. After outlining the basic facts about the (mostly automatic) annotation procedure in Section 5. we will add some final remarks in Section 6.

2. Sentence Representation in the Prague Dependency Treebank 2.0

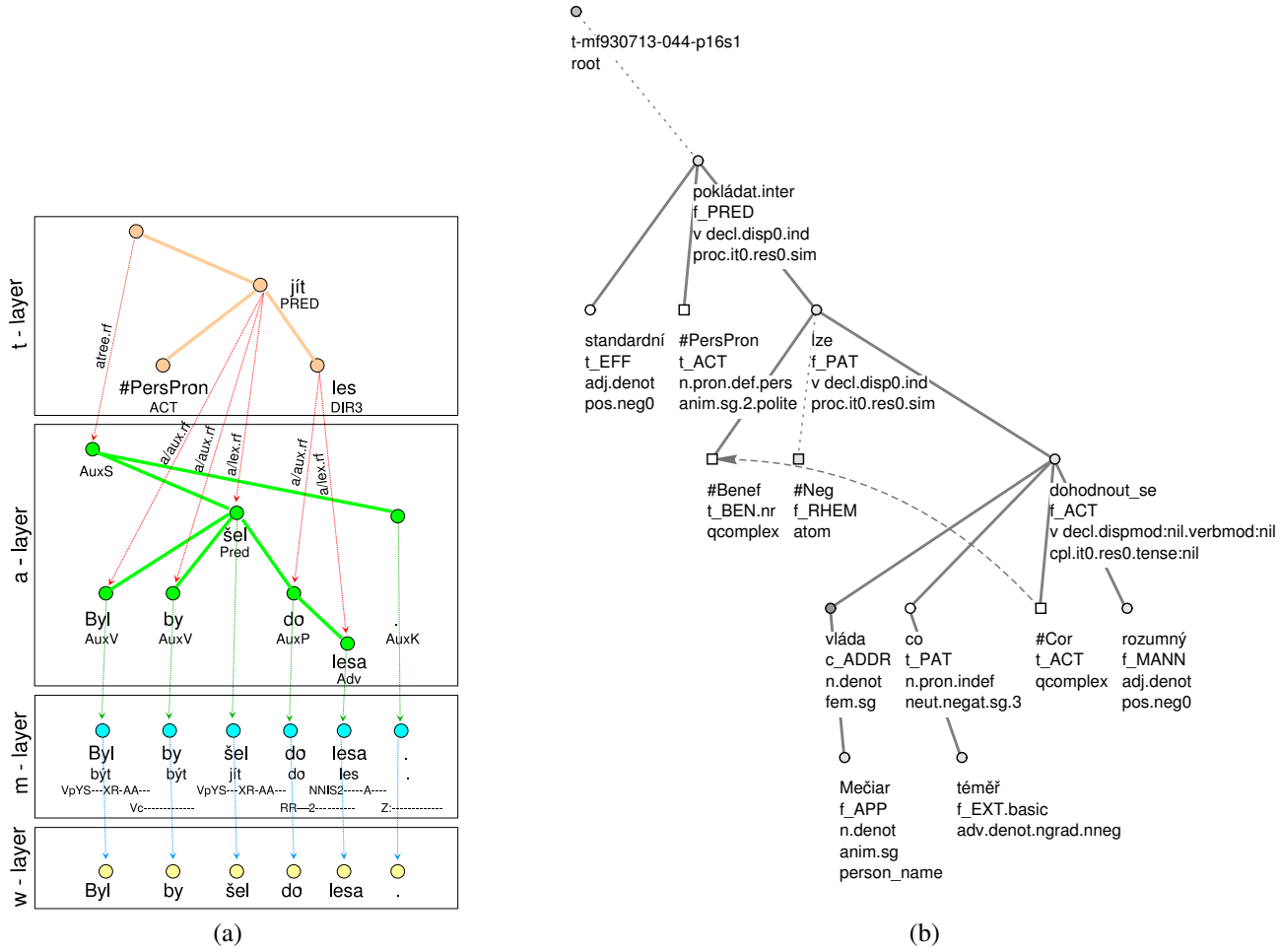
In the Prague Dependency Treebank annotation scenario, three layers of annotation are added to Czech sentences (see Figure 1 (a)):[2]

- morphological layer (m-layer), on which each token is lemmatized and POS-tagged,
- analytical layer (a-layer), on which a sentence is represented as a rooted ordered tree with labeled nodes and edges, corresponding to the surface-syntactic relations; one a-layer node corresponds to exactly one m-layer token,
- tectogrammatical layer (t-layer), which will be briefly described later in this section.

The full version of the PDT 2.0 data consists of 7,129 manually annotated textual documents, containing altogether 116,065 sentences with 1,960,657 tokens (word forms and punctuation marks). All these documents are annotated at the m-layer. 75 % of the m-layer data are annotated at the a-layer (5,338 documents, 87,980 sentences, 1,504,847 tokens). 59 % of the a-layer data are annotated also at the t-layer (i.e. 44 % of the m-layer data; 3,168 documents,

[1] Just for curiosity: almost the same term 'grammemes' is used for the same notion in the Meaning-Text Theory (Mel'čuk, 1988), although to a large extent the two approaches were created independently.

[2] Technically, there is also one more layer below these three layers which is called w-layer (word layer); on this layer the original raw-text is only segmented into documents, paragraphs and tokens and all these units are enriched with identifiers.



(a) (b)

Figure 1: (a) PDT 2.0 annotation layers (and the layer interlinking) illustrated (in a simplified fashion) on the sentence *Byl by šel do lesa.* ([He] would have gone into forest.), (b) tectogrammatical representation of the sentence: *Pokládáte za standardní, když se s Mečiarovou vládou nelze téměř na ničem rozumně dohodnout?* (Do you find it standard if almost nothing can be reasonably agreed on with Mečiar's government?)

49,442 sentences, 833,357 tokens).[3] The annotation at the t-layer started in 2000 and was divided into four areas:

a. building the dependency tree structure of the sentence including labeling of dependency relations and valency annotation,

b. topic / focus annotation,

c. annotation of coreference (i.e. relations between nodes referring to the same entity),

d. annotation of grammatemes and related attributes, the description of which is the main objective of this paper.

After the annotation of data had finished in 2004, an extensive cross-layer checking took over a year. The CD-ROM including the final annotation of PDT 2.0-data, a detailed documentation as well as software tools is to be publicly released by Linguistic Data Consortium in 2006.[4]

[3]The previous version of the treebank, PDT 1.0, was smaller and contained only m-layer and a-layer annotation (Hajič et al., 2001).

[4]See `http://ufal.mff.cuni.cz/pdt2.0/`

At the t-layer, the sentence is represented as a dependency tree structure built of nodes and edges (see Figure 1 (b)). Tectogrammatical nodes (t-nodes) represent auto-semantic words (including pronouns and numerals) while functional words such as prepositions have no node in the tree (with some exception of technical nature: e.g. coordinating conjunctions used for representation of coordination constructions are present in the tree structure). Each t-node is a complex data structure – it can be viewed as a set of attribute-value pairs, or even as a typed feature structure as used in unification grammars such as HPSG (Pollard and Sag, 1994).

For the purpose of our contribution, the most important attributes are the attribute t-lemma (tectogrammatical lemma), attribute functor, grammatemes and the classifying attributes nodetype and sempos. The annotation of attributes t-lemma and functor belongs to the area marked above as (a); these attributes will be introduced in the next paragraphs. Grammatemes and the attributes nodetype and sempos – all of them coming under the area (d) – will be characterized from the standpoint of annotation in Section 3. (The annotation of attributes belonging to the areas

(b) and (c) goes beyond the scope of this paper.)
The attribute t-lemma contains the lexical value of the t-node, or an 'artificial' lemma. The lexical value of the t-node is mostly a sequence of graphemes corresponding to the 'normalized' form of the represented word (i.e. infinitive for verbs or nominative form for nouns). In some cases, the t-lemma corresponds to the basic word from which the represented word was derived, e.g. in Figure 1 (b), the possessive adjective *Mečiarova* (*Mečiar's*) is represented by the t-lemma *Mečiar*, or the adverb *rozumně* (*reasonably*) is represented by the adjectival t-lemma *rozumný* (*reasonable*). The artificial t-lemma appears at t-nodes that have no counterpart in the surface sentence structure (e.g. the t-lemma #Gen at a verbal complementation not occurring in the surface structure because of its semantic generality), or it corresponds to personal pronouns, no matter whether expressed on the surface or not (e.g. the t-lemma #PersPron at the t-node in Figure 1 (b)). The dependency relation between the t-node in question and its parent t-node is stored in the attribute functor, e.g. functor EFF at the t-node with t-lemma *standardní* (*standard*), which plays the role of an effect of the predicate in the sentence displayed in Figure 1 (b).

3. Two-level Typing of Tectogrammatical Nodes

While the attributes t-lemma and functor are attached to each t-node of the tectogrammatical tree, grammatemes are relevant only for some of them. The reason for this difference consists in the fact that only some words represented by t-nodes bear morphological meanings.

3.1. Types of Tectogrammatical Nodes

To differentiate t-nodes that bear morphological meanings from those without such meanings, a classification of t-nodes was necessary. Based on the information captured by the above mentioned attributes t-lemma and functor, eight types of t-nodes were distinguished. The appurtenance of the t-node to one of the types is stored in the attribute nodetype.[5]

- **Complex nodes** (nodetype='complex') as the most important node type should be named in the first place: since they represent nouns, adjectives, verbs, adverbs and also pronouns and numerals (i.e. words expressing morphological meanings), they are the only ones with which grammatemes are to be assigned.

The other seven types of t-nodes and the corresponding values of the attribute nodetype are as follows:

- **The root of the tectogrammatical tree** (nodetype='root') is a technical t-node the child t-node of which is the governing t-node of the sentence structure.
- **Atomic nodes** (nodetype='atom') are t-nodes with functors RHEM, MOD etc. – they represent rhematizers, modal modifications etc.
- **Roots of coordination and apposition constructions** (nodetype='coap') contain the t-lemma of the coordinating conjunction or an artificial t-lemma of a punctuation symbol (e.g. #Comma).
- **Parts of foreign phrases** (nodetype='fphr') are components of phrases that do not follow rules of Czech grammar (labeled by a special functor FPHR in the tree).
- **Dependent parts of phrasemes** (nodetype='dphr') represent words that constitute a single lexical unit with their parent t-node (labeled by a special functor DPHR in the tree); the meaning of this unit does not follow from the meanings of its component parts.
- **Roots of foreign and identification phrases** (nodetype='list') are nodes with special artificial t-lemmas (#Forn and #Idph), which play the role of a parent of a foreign phrase (i.e. of nodes with nodetype='fphr' – see above) or the role of a parent of a phrase having a function of a proper name.
- So called **quasi-complex nodes** (nodetype= 'qcomplex') stand mostly for obligatory verbal complementations that are not present in the surface sentence structure (i.e. they have the same functors as complex nodes but, unlike them, quasi-complex t-nodes have artificial t-lemmas, e.g. #Gen).

3.2. Semantic Parts of Speech

Not all morphological meanings (chosen as tectogrammatically pertinent) are relevant for all complex t-nodes (cf., for example, the category of tense at nouns or the degree of comparison at verbs). As we did not want to introduce any 'negative' value to identify the non-presence of the given morphological meaning at a t-node (i.e., if all grammatemes would be annotated at each complex t-node, the negative value would be filled in at the irrelevant ones), the attribute sempos for sorting the t-nodes according to morphological meanings they bear had to be introduced into the attribute system.
The groups into which the complex t-nodes were further divided are called semantic parts of speech. According to basic onomasiological categories of substance, quality, event and circumstance (Dokulil, 1962), four semantic parts of speech were distinguished: semantic nouns, semantic adjectives, semantic verbs and semantic adverbs. These groups are not identical with the 'traditional' parts of speech: while ten traditional parts of speech are discerned in Czech and the appurtenance of the word to one of them is captured by a morphological tag (i.e. by an attribute of m-layer in the PDT 2.0), the 'only' four semantic parts of speech are categories of the t-layer and are captured by the attribute sempos (values n, adj, v and adv). The relations between semantic and traditional parts of speech are demonstrated in Figure 2. We would like to illustrate them on the example of semantic adjectives in more detail.
The following groups traditionally belonging to different parts of speech count among the semantic adjectives: (i) traditional adjectives, (ii) deadjectival adverbs, (iii) adjectival pronouns, and (iv) adjectival numerals.

[5] Some of the nodetype values are present in Figure 1 (b). If none of the nodetype values is indicated with the t-node, the nodetype is 'complex'.

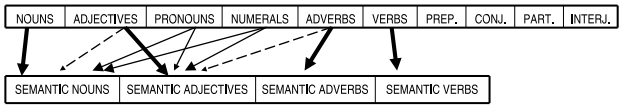


Figure 2: Relations of traditional parts of speech to their semantic counterparts. Arrows in bold denote a prototypical relation, thin arrows indicate the distribution of pronouns and numerals into semantic parts of speech and dotted arrows stand for the classification according to derivational relations.

(i) Traditional adjectives, e.g. *standardní* (*standard*) in Figure 1 (b), are mostly regarded as semantic adjectives (with the already mentioned exception of possessive adjectives converted to nouns).
(ii) At the t-layer, deadjectival adverbs, e.g. *rozumně* (*reasonably*) in Figure 1 (b), are represented by the t-lemma of the corresponding adjective, here by the t-lemma *rozumný* (*reasonable*). In this way, a derivational relation is followed: the word is represented by its basic word. Other types of derivational relations analyzed in PDT 2.0 will be introduced in the next sections.
(iii) and (iv) Since there are no groups such as 'semantic pronouns' or 'semantic numerals' at the t-layer, these words were distributed into semantic nouns and adjectives according to their function they fill in the sentence. While pronouns and numerals filling typical positions of nouns (such as agent or patient) belong to semantic nouns, pronouns and numerals playing an adjectival role are classified as semantic adjectives. For examples of nominal usage of the pronoun *který* (*which*) and of the numeral *sto* (*hundred*) see sentences (1), and (2) respectively:

(1) *Kurz, který*.n *jsem si vybral, je špatný.*
The course that I have chosen is bad.

(2) *Už vedl sto*.n *kurzů.*
He has already taught one hundred courses.

For examples of adjectival usage of the pronoun *který* (*which*) and of the numeral *tři* (*three*) see sentences (3), and (4) respectively:

(3) *Který*.adj *kurz si mám vybrat?*
Which course should I choose?

(4) *Vyučuje tři*.adj *kurzy.*
He teaches three courses.

The subgroups of semantic adjectives presented above are viewed as constituting the inner structure of this class. Also the classes of semantic nouns and semantic adverbs were sub-classified in a similar way. (Semantic verbs cannot be subdivided by the same principles as the other semantic parts of speech.)[6] The appurtenance of a t-node to a concrete subgroup of semantic parts of speech is captured as a detailed value of the attribute sempos (e.g. adj.denot or adj.quant.def in Figure 3).

[6]The sub-classification of semantic verbs is one of our future aims; properties of verbal systems in other languages (as studied e.g. in (Bybee, 1985)) will be considered.

The t-node hierarchy including the detailed subclassification of semantic adjectives is displayed in Figure 3.

4. Grammatemes and Their Values

There are 15 grammatemes at the t-layer of PDT 2.0. Grammatemes number, gender, person and politeness were assigned to t-nodes belonging to the subclasses of semantic nouns. The grammatemes degcmp, negation, numertype and indeftype were annotated with semantic nouns as well as with semantic adjectives, the latter two of them also with semantic adverbs. The other seven grammatemes belong to semantic verbs: tense, aspect, verbmod, deontmod, dispmod, resultative, and iterativeness.
All the grammatemes will be explained and exemplified in the following subsections one by one. A separate subsection is devoted to a more detailed discussion about pronominal words.

4.1. Number

The grammateme **number** is the tectogrammatical counterpart of the morphological category of number – the grammateme values, sg (for singular) and pl (for plural), mostly correspond to the values of this morphological category, e.g. the noun *vláda*.sg (*government*) in Figure 1 (b) is in singular while *vlády*.pl (*governments*) would be plural. However, as the grammateme captures the 'semantic' number, its value differs from that of the morphological category in some cases: e.g. while the morphological number of pluralia tantum is always 'plural' (e.g. the Czech word *dveře*, *door*), the tectogrammatical singular in a sentence like (5) is discerned from the tectogrammatical plural in the sentence (6) – at these nouns, the decision by an annotator was necessary; if such a decision were not possible on the basis of context (e.g. in the sentence (7)), a special value nr ('not recognized') was assigned.

(5) *Neotevírej tyto dveře*.sg
Do not open this door.

(6) *Šel dlouhou chodbou*
He walked through a long corridor
a minul několikery dveře.pl
and passed several doors.

(7) *Otevřel dveře*.nr
He opened the door/doors.

4.2. Gender

In PDT 2.0, values of the grammateme **gender** correspond to the morphological gender: anim (for masculine animate), inan (for masculine inanimate), fem (for feminine), and neut (for neuter).

4.3. Person and Politeness

The grammatemes **person** and **politeness** have been assigned to one subclass of semantic nouns that contains personal pronouns. These words are represented by the artificial t-lemma #PersPron at the t-layer (e.g. in the Figure 1 (b), where the t-node with the t-lemma #PersPron represents the actor that is not present in the surface sentence structure). The values of the former grammateme (1, 2, 3) distinguish among the 1st, 2nd and 3rd person pronouns;

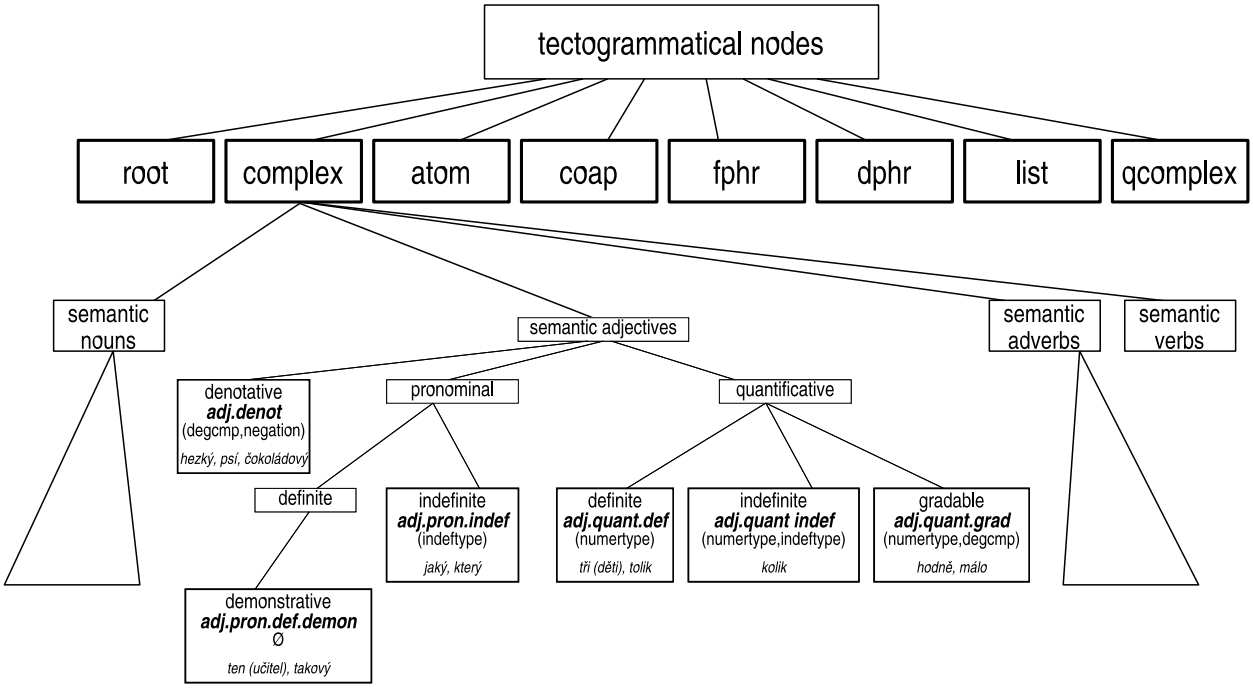


Figure 3: Hierarchy of t-nodes. The first branching renders the nodetype distinctions. Then, only complex t-nodes are further subdivided into four semantic parts of speech. Semantic nouns, semantic adjectives and semantic adverbs are further subclassified. Due to space limitations, only the subclassification of semantic adjectives is displayed in detail. In the leaf t-nodes of this subclassification, the values of attribute sempos is given on the second line and the list of grammatemes associated with the given class follows on the third line in the boxes.

the values of the latter one (basic, polite) discern the common from the polite usage of 2nd person pronouns. The surface pronoun is derived from the combination of t-lemma and values of grammatemes number, gender, person and politeness. E.g., the pronoun *vy* (*you*) in the sentence (8) is derived from the tectogrammatical representation #PersPron+**pl**+anim+2+**basic** in contrast to the same pronoun in the sentence (9) that is derived from the representation #PersPron+**sg**+anim+2+**polite**.

(8) *Vy jste vybrali dobrý kurz.*
'You have chosen a good course'
(- said to a group of persons)

(9) *Vy jste vybral dobrý kurz.*
'You have chosen a good course'
(- said politely to a single person)

4.4. Degree of Comparison

The grammateme **degcmp** corresponds to the morphological category of degree of comparison. Besides the values pos (for positive), comp (comparative) and sup (superlative), a special value acomp for comparative forms of adjectives/adverbs without a comparative meaning (so called 'absolute comparative', also 'elative') was established. The common usage of comparative forms such as *Jan je starší*.comp *než ona* (*Jan is elder than her*) was distinguished from the absolute usage e.g. in *starší*.acomp *muž* (*an elder man*) by the manual annotation.

4.5. Types of Numeral and Pronominal Expressions

Neither the grammateme **numertype** nor **indeftype** have a counterpart in the traditional set of morphological categories. They capture information on derivational relations among numerals, and pronominal words respectively, analyzed at the t-layer: derived words are represented by the t-lemma of its basic word and the feature that would be lost by such a representation is captured by values of these grammatemes. As all types of numerals are seen as derivations from the corresponding basic numeral and thus represented by its t-lemma, the grammateme numertype captures the type of the numeral in question. The surface numeral is then derived from the t-lemma and the value of this grammateme, e.g. the ordinal numeral *třetí* (*the third*) is derived form the following tectogrammatical representation: t-lemma *tři* (*three*) + numertype='ord' (for ordinal). Besides the value ord, the value set of this grammateme involves four other values: basic for basic numerals (*tři kurzy–three courses*), frac for fractional numerals (*třetina kurzu–the third of the course*), kind for numerals concerning the number of kinds/sorts (*trojí víno–three sorts of wine*), and set for numerals with meaning of the number of sets (*troje klíče–three sets of keys*).

In a similar vein, indefinite, negative, interrogative, and relative pronouns are represented by the t-lemma corresponding to the relative pronoun – the specific semantic feature is stored in the grammateme indeftype. Surface pronouns are derived from the lemma and the value of this grammateme: e.g. the indefinite pronoun *někdo* (*somebody*) and the negative pronoun *nikdo* (*nobody*) are derived from the

following tectogrammatical representations: t-lemma *kdo* + indeftype='indef', and t-lemma *kdo* + indeftype='negat' respectively.[7] Such representation of derivational relations makes it possible to represent all these words by a very small set of t-lemmas. The question of applying similar principles to pronominal words in other languages will be mentioned in Subsection 4.11.

4.6. Negation

Also the grammateme **negation** captures a lexical information needed for derivation of surface forms: it enables to represent both, the positive and the negative forms of adjectives, adverbs and (temporarily, only a group of) nouns by a single t-node with the same t-lemma – e.g. the adjective *standardní* (*standard*) in Figure 1 (b) as well as its negative form *nestandardní* (*non-standard*) are represented by the t-node with t-lemma *standardní* and the absence/presence of negation is captured by the value of the grammateme: the value neg0 was assigned to the t-node representing the positive form, the value neg1 to the t-node corresponding to the negative form.[8]

4.7. Tense

The grammateme **tense** corresponds to the morphological category of tense. The values sim (simultaneous with the moment of speech/with other event), ant (anterior to the moment of speech/to other event), and post (posterior to the moment of speech/to other event)[9] have been assigned automatically.

4.8. Aspect

The grammateme **aspect** is the tectogrammatical counterpart of the category of aspect. As there are verbs in Czech that can express both, imperfective and perfective aspects by the same forms (so called bi-aspectual verbs), manual annotation was necessary to make a decision with these verbs.

4.9. Verbal Modalities

There are three grammatemes concerning modality. The grammateme **verbmod** captures if the represented verbal form expresses the indicative (value ind), the imperative (imp), or the conditional mood (cdn). Since modal verbs do not have a t-node of their own at the t-layer (for explanation see (Panevová et al., 1971)), the deontic modality expressed by these verbs is stored in the grammateme **deontmod**, e.g. the predicate of the sentence *Už může odejít* (*He can already leave*) is represented by a t-node with t-lemma *odejít* (*to leave*) and the modality is stored as the value poss (for possibilitive) in the grammateme deontmod. The last of the modality grammatemes, the grammateme **dispmod**, concerns the so-called dispositional modality. This type of modality is represented by a special syntactic construction involving a 'reflexive-passive' verb construction, a dative form of a noun/personal pronoun playing the role of agent, and a modal adverb, e.g. the sentence (10):

(10) *Studentům se ta kniha čte dobře.*
Lit. *To students the book reads well.*
It is easy for the students to read the book.

4.10. Resultative and Iterativeness

While the grammateme **resultative** (values res1, res0) reflects the fact whether the event is/is not presented as a resultant state, the last verbal grammateme **iterativeness** indicates whether the event is/is not viewed as a repeated (multiplied) action (values it1, it0).

4.11. Pronominal Words at the T-layer

In this chapter, we would like to provide a deeper view into the principles of representation of pronominal words at the t-layer of PDT 2.0, and then to outline how this representation can be applied to such words in English or German.
As already mentioned above, pronouns are represented by a minimal set of t-lemmas at the t-layer. Personal pronouns by a single (artificial) t-lemma #PersPron; grammatemes assigned to the t-nodes of personal pronouns were presented in the previous chapter. Indefinite, negative, in-

T-lemma:	*kdo*	*co*	*který*	*jaký*
indefype:				
relat	kdo	co	který, jenž	jaký
indef1	někdo	něco	některý	nějaký
indef2	kdosi kdos	cosi cos	kterýsi	jakýsi
indef3	kdokoli kdokoliv	cokoli cokoliv	kterýkoli kterýkoliv	jakýkoli jakýkoliv
indef4	ledakdo leckdo	ledaco lecco	leckterý ledakterý	lecjaký ledajaký
indef5	kdekdo	kdeco	kdekterý	kdejaký
indef6	kdovíkdo málokdo	kdovíco máloco	kdovíkterý málokterý	kdovíjaký všelijaký
inter	kdo kdopak	co copak	který kterýpak	jaký jakýpak
negat	nikdo	nic	žádný	nijaký
total1	všechen	všechno vše	-	-
total2	-	-	každý	-

Table 1: The indeftype grammateme has actually eleven values (1st column in the table). It makes it possible to represent all semantic variants of pronouns *kdo* (*somebody*), *co* (*something*), *který* (*that*) and *jaký* (*what*) (in the 2nd, 3rd, 4th and 5th column) by only four t-lemmas at the t-layer.

[7] A similar treatment of indefinite and negative pronouns as of two subtypes of the same entity can be found in (Helbig, 2001).

[8] Unlike this representation, negative verbal forms (verbal negation is expressed also by the prefix *ne-* in Czech) are represented by a sub-tree consisting of a t-node with a verbal t-lemma the child of which is a t-node with the artificial t-lemma #Neg; cf. the representation of the negated verb *nelze* ((it) *can not be*) by two t-nodes, with the t-lemmas *lze* ((it) *can be*) and #Neg, in Figure 1 (b). The explanation can be found in (Hajičová, 1975).

[9] As the class of semantic verbs has not been sub-classified yet and all verbal grammatemes were annotated with each verbal t-node, a special value nil was inserted into the value system for cases when the represented word does not express a feature captured by the grammateme (cf. the value of grammateme tense at a t-node representing an infinitive form).

terrogative and relative pronouns are all represented by a t-lemma corresponding to the relative pronoun. In this way, only four lemmas – i.e. *kdo* (*somebody*), *co* (*something*), *který* (*which*) and *jaký* (*what*) – are sufficient to represent all Czech pronouns of named types at the t-layer. The pronouns with corresponding values of the grammateme indeftype are displayed in Table 1.

Since the semantic features stored in the grammateme indeftype are expressed also by other words of pronominal character in Czech, e.g. by pronominal adverbs *nikde* (*nowhere*) or *nějak* (*somehow*), or by an indefinite numeral *několik* (*a few*), we can use this grammateme also for the tectogrammatical representation of these words.[10]

As the groups of pronominal words are unproductive classes with (at least to a certain extent) transparent derivational relations not only in Czech, but also in other languages, we believe that similar regularities to those captured in Czech by the indeftype grammateme can be found also elsewhere. However, as it is obvious from the preliminary sketch of several English and German pronouns classified in Table 2,[11] the application of our scheme to other languages will not be straightforward and various subtle differences have to be taken into account. For instance, there is only one negative form *nikdo* corresponding to the t-lemma *kdo* in Czech, therefore the present system provides no means for distinguishing German negative pronouns *niemand* and *niergendjemand*. A new question arises also in the case of English *anybody* when used in negative clauses, which has no counterpart in Czech or German.

	English	English	German	German
T-lemma	*who*	*what*	*wer*	*was*
indefype:				
relat	who	what	wer	was
indef1	somebody	something	jemand	etwas
indef2	-	-	irgendjemand	irgendetwas
indef3	whoever	whatever	-	-
inter	who	what	wer	was
negat	nobody	nothing	niemand	nichts
total1	all	everything	alle	alles
total2	each	each	jeder	jedes

Table 2: Selected English and German pronouns preliminarily classified according to the indeftype grammateme.

5. Implementation

The procedure for assigning grammatemes (and nodetype and sempos) to nodes of tectogrammatical trees was implemented in ntred[12] environment for processing the PDT data. Besides almost 2000 lines of Perl code, we formulated a number of rules for grammateme assignment written in a text file using a special economic notation (roughly 2000 lines again), and numerous lexical resources (e.g. special-purpose list of verbs or adverbs). As we intensively used all information available also at the two 'lower' levels of the PDT (morphological and analytical), most of the annotation could have been done automatically with a highly satisfactory precision.

It should be emphasized that the inter-layer links played a key role in the procedure. As it is clear from Figure 1 (a), it would not be possible to set e.g. the value of the number grammateme of the (already lemmatized) t-node *les* (forest) without having the access to the morphological tag of the corresponding m-layer unit in the given sentence, or to find out that the verb *jít* (to go) is in conditional mood (verbmod=cdn) without knowing that the corresponding a-layer complex verb form subgraph contains the node *by*.

Due to the fact that a lot of effort had been spent on checking and correcting of the inter-layer pointers in PDT 2.0, finally we needed only around 5 man-months of human annotation for solving just the very specific issues (as mentioned at single grammatemes in the previous section).

Now we would like to show a fragment of the above mentioned rules. For a given t-node: if the lemma of the corresponding m-node is *který* (which), the t-node itself is not in the attributive syntactic position and participates in grammatical coreference (i.e., it forms a relative construction), then sempos=n.pron.indef, indeftype=relat, and the values of the grammatemes gender and number are inherited from the coreference antecedent. This rule would be applied on the sentence (1).

To further demonstrate that grammatemes are not just dummy copies of what was already present in the morphological tag of the node, we give two examples:

- Deleted pronouns in subject positions (which must be restored at the t-layer) might inherit their gender and/or number from the agreement with the governing verb (possibly complex verbal form), or from an adjective (if the governor was copula), or from its antecedent (in the sense of textual coreference).

- Future verbal tense in Czech can be realized using simple inflection (perfectives), or auxiliary verb (imperfectives), or prefixing (lexically limited).

The procedure was repeatedly tested on the PDT data, which was extremely important for debugging and further improvements of the procedure. Final version of the procedure was applied to all the available tectogrammatical data (as for its size, recall the second paragraph in Section 2.). This data, enriched with node classification and grammateme annotation, will be included in PDT 2.0 distribution.

Due to the highly structured nature of the task, it is difficult to present the results of the annotation procedure from the quantitative viewpoint. However, at least the distribution of the values of nodetype and sempos are shown in Tables 3 and 4.

[10] The indeftype grammateme is applied to indefinite numerals together with the above-mentioned grammateme numertype – thus only a single t-lemma *kolik* (*how many*) represent words of different nature: e.g. *několik´at´y* (*not the first*), *kolikr´at* (*how many times*) etc.

[11] We chose English and German, because, first, the two languages are the most familiar to the present authors, and second, certain experiments concerning their t-layer have already been performed, see e.g. (Cinková, 2004) or (Kučerová and Žabokrtský, 2002).

[12] `http://ufal.mff.cuni.cz/~pajas`

complex	550947
root	49442
qcomplex	46015
coap	35747
atom	34035
fphr	4549
list	2512
dphr	1282

Table 3: Values of nodetype sorted according to the number of occurences in the PDT 2.0 t-layer data.

n.denot	236926
adj.denot	100877
v	88037
n.pron.def.pers	32903
adj.quant.def	19441
n.denot.neg	18831
n.pron.indef	11343
adv.denot.ngrad.nneg	8947
n.quant.def	7994
adj.pron.def.demon	5746
n.pron.def.demon	4759
adj.pron.indef	3383
adv.pron.indef	3107
adv.pron.def	2928
adj.quant.grad	1865
adv.denot.grad.neg	1315
adv.denot.grad.nneg	1139
adv.denot.ngrad.neg	751
adj.quant.indef	655

Table 4: Detailed values of sempos sorted according to the number of occurences in the PDT 2.0 t-layer data.

6. Conclusion

We believe that two important novel goals have been achieved in the present enterprise:

- We proposed a formal classification of tectogrammatical nodes and described its consequences on the system of grammatemes, and thus the tectogrammatical tree structures become formalizable e.g. by typed feature structures.
- We implemented an automatic and highly-complex procedure for capturing the node classification, the system of grammatemes and derivations, and verified it on large-scale data, namely on the whole tectogrammatical data of PDT 2.0. Thus the results of our work will be soon publicly available.

In the paper we do not compare our achievements with related work, since we are simply not aware of a comparably structured annotation on comparably large data in any other publicly available treebank. For instance, to our knowledge no other treebank attemps at reducing the (semantically redundant) morphological attributes imposed only by agreement, or at specifying verbal tense for a complex verb form as for a whole, or at representing a noun (or a personal pronoun) and the corresponding possessive adjective (or possessive pronoun, respectively) in a unified fashion. However, from the theoretical viewpoint the presented model bears some resemblances with the system of grammemes in the deep-syntactic level of the already mentioned Meaning-Text Theory (Mel'čuk, 1988).

In the near future, we plan to separate the grammatemes that bear the derivational information (such as numertype) from the grammatemes having their direct counterpart in traditional morphological categories. The long-term aim is to describe further types of derivation: we should concentrate on productive types of derivation (diminutive formation, formation of feminine counterparts of agentive nouns etc.). The set of 'derivational' grammatemes will be extended in this way. The next issue is the problem of subclassification of semantic verbs. The challenging topic is also the study of grammatemes in other languages.

Acknowledgements

The research reported in this paper was supported by the projects 1ET101120503, GA-UK 352/2005 and GD201/05/H014. We would also like to thank professors Jarmila Panevová and Eva Hajičová for numerous comments on the draft of the paper.

7. References

Joan L. Bybee. 1985. *Morphology: A study of the relation between meaning and form*. Benjamins, Philadelphia.

Silvie Cinková. 2004. Manuál pro tektogramatickou anotaci angličtiny. Technical report, ÚFAL/CKL MFF UK.

Miloš Dokulil. 1962. *Tvoření slov v češtině I*. Academia, Prague.

Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová Hladká. 2001. Prague Dependency Treebank 1.0.

Eva Hajičová, Jan Hajič, Barbora Vidová-Hladká, Martin Holub, Petr Pajas, Veronika Kolářová-Řezníčková, and Petr Sgall. 2001. The Current Status of the Prague Dependency Treebank. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pages 11–20, Berlin, Heidelberg, New York. Springer-Verlag.

Eva Hajičová. 1975. *Negace a presupozice ve významové stavbě věty*. Academia, Prague.

Hermann Helbig. 2001. *Die semantische Struktur natürlicher Sprache*. Springer-Verlag, Berlin, Heidelberg, New York.

Ivona Kučerová and Zdeněk Žabokrtský. 2002. Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees. *Prague Bulletin of Mathematical Linguistics*, (78):77–94.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

Jarmila Panevová, Eva Benešová, and Petr Sgall. 1971. *Čas a modalita v češtině*. Univerzita Karlova, Prague.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague.