Annotation of Morphological Meanings of Verbs Revisited

Jarmila Panevová, Magda Ševčíková

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
{panevova,sevcikova}@ufal.mff.cuni.cz

Abstract

Meanings of morphological categories are an indispensable component of representation of sentence semantics. In the Prague Dependency Treebank 2.0, sentence semantics is represented as a dependency tree consisting of labeled nodes and edges. Meanings of morphological categories are captured as attributes of tree nodes; these attributes are called grammatemes. The present paper focuses on morphological meanings of verbs, i.e. on meanings of the morphological category of tense, mood, aspect etc. After several introductory remarks, seven verbal grammatemes used in the PDT 2.0 annotation scenario are briefly introduced. After that, each of the grammatemes is examined. Three verbal grammatemes of the original set were included in the new set without changes, one of the grammatemes was extended, and three of them were substituted for three new ones. The revised grammateme set is to be included in the forthcoming version of PDT (tentatively called PDT 3.0). Rules for automatic and manual assignment of the revised grammatemes are further discussed in the paper.

1. Introduction

In the present paper, some refinements in annotation of morphological meanings of verbal categories within the Prague Dependency Treebank are suggested. The annotation scenario of the Prague Dependency Treebank version 2.0 (PDT 2.0 in the sequel) was built on the theoretical basis of Functional Generative Description (FGD); see e.g. (Sgall et al., 1986). PDT 2.0 annotation scenario differs from the theoretical approach of FGD in several (though not fundamental) respects, see (Štěpánek, 2006).

In FGD and PDT 2.0, the linguistic meaning of the sentence is represented as a dependency tree structure consisting of nodes and edges with a set of attributes at the so-called tectogrammatical layer. Morphological meanings of verbs (and of other auto-semantic words) are represented as attributes of nodes of the tectogrammatical tree; these attributes are called grammatemes.[1] Grammatemes are mostly counterparts of such morphological categories whose meaning is indispensable for the sentence semantics and which belong to the functional onomatology, see (Mathesius, 1929). In PDT 2.0, fifteen different grammatemes were used; seven of them were designed for rendering morphological meanings of verbs (for more details on PDT 2.0, see Section 2. of the paper).

When designing the annotation scenario of PDT 2.0, we were aware that all requirements of the theoretical background (FGD) cannot be reflected within the representation of sentence. Taking into account these theoretical requirements of FGD as well as new results of linguistic research on the one hand and our experience with the annotation procedure and with the use of PDT 2.0 data within several NLP tasks on the other, we aim now at a revision and refinement of the PDT 2.0 annotation scenario, particularly in the domain of meanings of morphological categories of verbs.

The revised set of verbal grammatemes, which will be incorporated in the new, both updated and extended, version of PDT (tentatively called PDT 3.0), is proposed in Section 3. Three verbal grammatemes of the original set used in PDT 2.0 were included in the new set without changes, one of the grammatemes was extended, and three of them were substituted by three new ones. After an explanation and exemplification of each of the proposed grammatemes and their values, basic annotation rules for manual and/or (semi-)automatic grammateme assignment are discussed in Section 4. Some final remarks are included in Section 5.

2. Current representation of morphological meanings of verbs in PDT 2.0

2.1. Basic characteristics of PDT 2.0

PDT 2.0 is a collection of contemporary Czech newspaper texts from 1990's to which a morphological annotation and annotation at two syntactic layers was assigned, at the analytical layer (layer reflecting the surface shape of a sentence) and at the tectogrammatical layer (layer of the linguistic meaning of the sentence).[2] At the morphological layer, each token (word form or punctuation mark) in each sentence of the source texts is lemmatized and tagged with a positional tag. At the analytical layer, a sentence is represented as a dependency tree with labeled nodes and edges, which correspond to surface-syntactic relations (such as subject, object etc.). One analytical node corresponds to exactly one morphological token.

At the tectogrammatical layer, the meaning of the sentence is represented as a dependency tree structure; tectogrammatical nodes represent auto-semantic words (including pronouns and numerals) whereas functional words, such as

[1] Cf. the term "grammemes" used for the same notion in Meaning–Text Theory (Mel'čuk, 1988).

[2] There is one more, "technical" layer in PDT 2.0 (so-called word layer), at which the source texts are just segmented and labeled with identifiers. This layer is omitted in the present paper.

Grammateme	Explanation
tense	meanings of the morphological category of tense
iterativeness	whether an event is / is not presented as a repeated action (compatible both with the processual and the complex aspect)
deontmod	modal meanings (necessity, possibility, permission etc.) expressed by modal verbs
aspect	meanings of verbal aspect
resultative	whether an event is / is not presented as a result of the preceding action
dispmod	whether the attitude of the agent to an event is / is not expressed by a special syntactic construction
verbmod	direct counterpart of the morphological category of mood

Table 1: Set of verbal grammatemes implemented in PDT 2.0

prepositions, have no node in the tree.[3] Verb forms consisting of more than one token are represented by a single tectogrammatical node labeled with a lemma corresponding to the infinitive form. Tectogrammatical lemmas of other nodes often correspond to morphological lemmas (e.g. to nominative form for nouns); however, in some cases a basic form from which the word in question was derived is used as the tectogrammatical lemma (for instance, possessive adjectives are represented as their basic nouns; cf. Section 2.2.) or an "artificial" lemma is attached, for instance, in cases of surface deletions, i.e. to nodes which have no counterpart within the surface shape of the sentence (e.g. to a node representing a subject omitted in the sentence).

Besides the tectogrammatical lemma, tectogrammatical nodes are labeled with dependency relations (so-called functors, e.g. actor ACT, addressee ADDR, local specification LOC) and a set of other attributes; some of them (grammatemes and immediately related attributes nodetype and sempos) are described in Section 2.2. Furthermore, valency annotation, annotation of coreference, and annotation of topic-focus articulation are available in tectogrammatical trees as well.

PDT 2.0 data consist of more than 7 thousand manually annotated textual documents, containing altogether more than 115 thousand sentences with nearly 2 million tokens. All these documents were annotated at the morphological layer, 75 % of them also at the analytical layer. Nearly 60 % of the data annotated at the analytical layer (i.e. 45 % of the morphologically annotated data) were annotated also at the tectogrammatical layer, i.e. over 3 thousand documents consisting of more than 49 thousand sentences with more than 830 thousand tokens. The CD-ROM with annotated PDT 2.0 data, a detailed documentation and software tools was publicly released at Linguistic Data Consortium in 2006 (Hajič et al., 2006).

2.2. Verbal grammatemes and their implementation in PDT 2.0

Every grammaticalized morphological category present in Czech language is displayed by two sets of values in PDT 2.0: one set concerns morphological forms and is involved in the morphological tag (e.g. present, preterite and future for the category of tense), the other one represents their meanings (simultaneity, anteriority, posteriority, respectively), and is captured by values of particular morphological grammatemes. As already mentioned, grammatemes are node attributes by means of which such morphological meanings are represented that are indispensable for the meaning of the sentence. Concerning morphological categories of verbs, e.g. tense and aspect are semantically relevant and have thus to be included in the tectogrammatical representation whereas person and number of the verb forms are only imposed by agreement, therefore they have no counterpart at the tectogrammatical layer. Seven grammatemes that were assigned to nodes representing verb forms in PDT 2.0 are displayed in Table 1, a more detailed explanation of the grammatemes is to be found in Section 3.; see also (Mikulová et al., 2006).

Grammatemes were assigned only to nodes that represent words expressing morphological meanings, i.e. nouns, adjectives, verbs, and adverbs as well as pronouns and numerals; grammatemes were not attached to other nodes (for instance, to nodes representing a reconstructed ACT of an infinitive).[4] The fact that a particular node represents a word with morphological meanings is indicated in the node attribute nodetype: with nodes representing nouns, adjectives, verbs, adverbs, pronouns and numerals the value complex was assigned; details on the attribute nodetype (as well as the following sempos) and its values can be found in (Razímová and Žabokrtský, 2006).

After making the distinction between complex nodes (i.e. nodes to which grammatemes are to be assigned) and the other nodes of the tectogrammatical tree, further subclassification of complex nodes was required since not all morphological meanings are relevant for all complex nodes. The groups into which complex nodes are further subdivided are called semantic parts of speech. Four semantic parts of speech were differentiated: semantic nouns, semantic adjectives, semantic verbs and semantic adverbs (according to basic onomasiological categories of substance, quality, event and circumstance, cf. (Dokulil, 1962)).

These groups differ from "traditional" parts of speech especially in the following aspects: firstly, pronouns and numerals were distributed into semantic nouns and adjectives; secondly, adverbs derived from adjectives were

[3] There are several exceptions of technical nature. For instance, coordinating conjunctions, which are used for representation of coordination constructions, are included in the tree.

[4] In this aspect, assignment of grammatemes differs from that of tectogrammatical lemma and functor, which were attached to each node of the tectogrammatical tree.

treated as semantic adjectives; thirdly, possessive adjectives were classified as semantic nouns; see (Ševčíková and Žabokrtský, 2006). Nevertheless, the group of semantic verbs, with which we are concerned in the present paper, currently corresponds to the traditional word class of verbs. All seven verbal grammatemes were assigned to each node belonging to semantic verbs in PDT 2.0. Number of occurrences of values of all verbal grammatemes in PDT 2.0 are listed in Table 2.[5]

The described principles, on which the annotation of grammatemes at the tectogrammatical layer of PDT 2.0 was based, will be applied when assigning the revised set of verbal grammatemes (introduced in the next section) within PDT 3.0 as well.

Grammateme	Value	# of occurences
tense	ant	31217
	sim	40987
	post	8654
	nil	7166
iterativeness	it0	87919
	it1	105
deontmod	deb	1173
	hrt	3255
	vol	1016
	poss	2777
	perm	92
	fac	95
	decl	79616
aspect	proc	51900
	cpl	35839
	nr	285
resultative	res0	87669
	res1	355
dispmod	disp0	80824
	disp1	9
	nil	7191
verbmod	ind	77145
	cdn	3680
	imp	375
	nil	6824

Table 2: Number of occurrences of values of verbal grammatemes in PDT 2.0 data

[5]Besides "proper" values, which are explained for each grammateme in Section 3., two special values nil and nr were used in PDT 2.0. The value nil (occurring with the grammatemes tense, dispmod, and verbmod in Table 2) was filled in if the verb form represented by the node in question did not express the meaning of the particular grammateme; e.g. in the grammateme tense, the value nil was assigned with nodes representing an infinitive form. The value nr was used if the annotator was not able do choose one of the given grammateme values; concerning the grammateme aspect, the value nr was assigned with verbs which can express both processual and complex events (bi-aspectual verbs) if the annotator could not decide between these two interpretations.

3. Revised set of verbal grammatemes

From the set of verbal grammatemes used in PDT 2.0, the grammatemes tense, iterativeness and deontmod and their values remain untouched (cf. Sections 3.1., 3.2. and 3.3., respectively). A new value was included in the value set of the grammateme aspect (Section 3.4.). The grammatemes dispmod and resultative are canceled and new grammatemes of grammatical diathesis diatgram and of syntactic diathesis diatsynt are included (Section 3.5. and 3.6., respectively). In Section 3.7., the grammateme factmod is described, which partially substitutes the grammateme verbmod used in PDT 2.0.

3.1. The grammateme tense

Three values of the grammateme tense are distinguished: sim for simultaneity, ant for anteriority, and post for posteriority as to their "point of reference" (R). The point of reference R is determined according to the position of the verb (event) in the structure of a complex sentence. The difference between so-called "absolute" tense (the relation of the verb to the point of speech) and "relative" tense (the relation to another event in the complex sentence) is reflected by three recursive rules: for the verb in the main clause, R is always the point of speech; for the verb in a content clause, R is the event of its governing clause; for the verb in an adjunct clause, R is the same as R of its governing clause; for a detailed analysis, see (Panevová et al., 1971). The secondary usage of the tense forms (as, e.g., *praesens historicum* or *preasens pro futuro*) is not covered by the general rules and must be treated individually.

3.2. The grammateme iterativeness

The grammateme iterativeness has two values: it1 for repeated events and it0 for events unmarked for iterativeness. In PDT 2.0, the marked value it1 was assigned only with nodes which represented verbs with special word-forming affixes expressing repetition; cf. the iterative verb form *spává* 'he (usually) sleeps' vs. its noniterative counterpart *spí* 'he sleeps'. In the proposed extended annotation scheme, the iterativeness (compatible with perfect aspect) expressed lexically (for instance, by the adverbs *vždy* 'always', *často* 'often', *pokaždé* 'every time', *denně* 'every day') is presupposed to be treated as well.

3.3. The grammateme deontmod

The values of the grammateme deontmod (for the so-called deontic modality) refer to necessity, possibility, optionality etc. of events. These meanings are expressed prototypically by modal verbs understood as auxiliaries of the autosemantic verbs in FGD as well as in PDT 2.0, and captured by a respective value of the grammateme deontmod. Arguments for such type of representation and delimitation of the deontmod values are given in (Panevová et al., 1971) and (Sgall et al., 1986). Seven values of this grammateme were distinguished in PDT 2.0:

1. value deb for events understood as "necessary" and expressed by the modal verb *muset* 'must / have to'

2. value hrt for "obligatory" events corresponding to the modal verb *mít* 'should / ought to'

3. value vol for "wanted / intended" events expressed by modal verbs *chtít* and *hodlat* 'want'
4. value poss for "possible" events corresponding to modal verbs *moct* and *dát se* 'can'
5. value perm for "permitted" events expressed by the modal verb *smět* 'may'
6. value fac for events understood as "an ability to do something" and corresponding to modal verbs *dovést* and *umět* 'can'
7. value decl for verbs umarked for deontic modality (i.e. auto-semantic verbs not modified by a modal verb)

The values of the grammateme deontmod is to be applied without changes within the revised annotation scenario.

3.4. The grammateme aspect

The core of the category of aspect is constituted by the opposition of processual events (expressed primarily by the imperfective aspect) and complex events (expressed primarily by the perfective aspect); this opposition is captured by the values proc and cpl, respectively. Since there is no formal counterpart of perfect tenses in Czech language, the verbal aspect as a kind of a morphological category covers partially the lack of formal perfect tenses. Therefore, the value perf (for perfective state) was added into the value set of the grammateme aspect. Besides the forms with the meaning of perfective state, the meaning of the result of the preceding event is expressed also by the forms of resultative diathesis (cf. values res1 and res2 in Section 3.5.) expressing the category of "resultative state".[6] The partial synonymy between these two types of expressions is another reason for enrichment of aspect values with the value perf, though it is not easy to distinguish between these two meanings of the perfective aspectual form; the interpretation depends on context, see the difference between examples (1) and (2). Only (2) may be paraphrased as (3).[7]

(1) ***Roztrhla si*** *šaty o skobu.* [*roztrhnout si*.cpl.act]
lit.: 'She **tore** her dress by the hook.'

(2) ***Roztrhla si*** *šaty, přesto v nich šla do divadla.*
[*roztrhnout si*.perf.act]
lit.: 'She **had torn** her dress but in spite of this she went in it to the theatre.'

(3) ***Měla*** *šaty* ***roztržené****.* [*roztrhnout si*.perf.res2]
lit.: 'She **had** her dress **torn**.'

3.5. The grammateme diatgram

Grammatical diatheses are closely related to the traditional category of verbal voice. The opposition active vs. passive voice constitutes the core of the proposed grammateme diatgram (with corresponding values act and pas, respectively); however, other (secondary) meanings which are productive enough to be considered as grammatical ones are included as values of the grammateme diatgram, namely resultatives (values res1 and res2), deagentization (value deagent), disposition (value disp), and recipient (value recip). The marked grammatical diatheses are characterized by some changes of the verb form (in Czech, it is mostly an analytic form with (semi-)auxiliaries such as *být* 'to be', *mít* 'to have', *dostat* 'to get') and by the shifts of verbal participants into a non-prototypical surface position in the sentence structure: the ACT is mostly shifted from the subject position.

3.5.1. Resultativeness (res1, res2)

Two values for description of resultativeness are introduced: res1, res2. In the constructions with res1 meaning, the ACT position is suppressed;[8] the construction is constituted by the analytical verb form of the auxiliary *být* 'to be' and *-n/-t* participle agreeing with the surface subject. The subject position is filled by a participant different from ACT (ex. (4)) or it is empty if both the ACT and the patient PAT are generalized (ex. (5)) and has a form of neuter sg.
The res2 meaning is a "possessive" variant of res1; it is constituted by the semi-auxiliary verb *mít* 'to have', *-n/-t* participle (which agrees either with the object, or, in objectless sentences, it has unmarked agreement – neuter sg.), the subject position may be interpreted either as an ACT or as an ADDR, see ex. (6) and (7).
The interpretation of the subject position depends on contextual criteria; in many sentences we have to do with ambiguity between these two interpretations. The ontological conditions exclude (or at least make less probable) the interpretation subject=ACT in ex. (8) while the interpretation subject=ACT is obvious in (9). Example (10) illustrates the ambiguity between these two interpretations.

(4) *Oběd* ***je uvařen****.* [*uvařit*.res1]
lit.: 'The lunch **is cooked**.'

(5) ***Je uvařeno****.* [*uvařit*.res1]
lit.: '(It) **is cooked**.'

(6) *Matka* ***měla*** *už oběd* ***uvařen*** *(když přijeli hosté).*
[*uvařit*.res2]
Matka ***měla*** *už oběd* ***uvařen*** *(když se vrátila domů).*
[*uvařit*.res2]
lit.: 'Mother already **had** the lunch **prepared** (when the guests arrived).'
lit.: 'Mother **had** already **had** the lunch **prepared**

[6] The term "resultative state" (výsledný stav) was introduced by (Hausenblas, 1962) as a candidate for a new type of grammatical category of the verb.

[7] In our examples, the sentence is written in italics. The tectogrammatical lemma by which the verb form/s in bold is/are represented in the tectogrammatical tree is given in square brackets after the example sentence. The tectogrammatical lemma is followed by the value of the particular grammateme which is appropriate with regard to the context of the given sentence. In examples (1) to (3), both the values of the grammateme aspect and diatgram are specified since these two grammatemes are closely related; diatgram is explained in Section 3.5. The source of the example is cited in round brackets ("CNC" for Czech National Corpus; `http://ucnk.ff.cuni.cz`); if no source is cited, it is an example created by the authors of the paper.

[8] By the presence of the ACT in similar structures as *Oběd byl uvařen prvotřídním kuchařem* 'The lunch was prepared by a first-class cook' the fact that the sentence expresses a passive voice is signalized.

aspect / diatgram	proc	cpl	perf
act	*Bratr píše dopis.* lit: 'Brother writes / is writing a letter.'	*Bratr napsal dopis.* lit: 'Brother wrote a letter.'	*Bratr napsal dopis.* lit: 'Brother has written a letter.'
pas	*Dopis byl psán Napoleonem.* lit: 'The letter was (being) written by Napoleon.'	*Dopis byl napsán Napoleonem u Borodina.* lit: 'The letter was written by Napoleon near Borodino.'	*Dopis byl napsán, odešli ho.* lit: 'The letter has been written, send it away.'
res1	–	–	*Oběd je uvařen. / Dopis je napsán.* lit: 'The lunch is cooked. / The letter is written.'
res2	–	–	*Matka měla oběd uvařen.* lit.: 'Mother had the lunch prepared. / Mother had had the lunch prepared.'
deagent	*Dopisy se dnes píší na počítači.* lit: 'Today, the letters are being written on computers.'	*Citace se napíšou kurzivou.* lit: 'Quotations will be written in italics.'	*Bábovka se snědla celá.* lit: 'The cake has been eaten whole.'
disp	*Eseje se (mu) píšou snadno.* lit: 'Essays are easy (for him) to write.'	*Esej se (mu) napíše snadno.* lit: 'An essay will be easy (for him) to write.'	–
recip	*Bratr dostává (od otce) vynadáno.* 'lit: Brother gets a scolding (from his father).'	*Bratr dostal (od otce) vynadáno.* lit: 'Brother got a scolding (from his father).'	*Bratr dostal (od otce) vynadáno.* lit: 'Brother has got a scolding (from his father).'

Table 3: Verb forms corresponding to particular combinations of the values of the grammatemes diatgram (listed in the 1st column) and the values of aspect (in the 1st line). If the combination od the given values is not realized in Czech, the symbol "–" is used.

(when she arrived home).'

(7) *Matka už **má uvařeno**.* [*uvařit*.res2]
lit.: 'Mother **has** already **cooked**. / Somebody **has** already **cooked** for mother.'

(8) *Pacient **měl zasaženy** vnitřní orgány.*
[*zasáhnout*.res2] (CNC)
lit: 'The patient **had** his inner organs **afflicted**.'

(9) *O mnoho víc **neměl nalétáno** ani čtyřiadvacetiletý pilot.* [*nalétat*.res2] (CNC)
lit.: 'The twenty four years old pilot **had not** yet **flown** much more.'

(10) *. . . ženu kriminalisté nalezli v jejím bytě, **měla** kolem krku **omotáno** vodítko na psa.*
[*omotat*.res2] (CNC)
lit.: '. . . criminalists have found the lady in her flat, she **had** a dog-lead **wrapped** around her throat.'

3.5.2. Deagentization (deagent)

The reflexive form of the verb with the suppressing of the ACT/subject position is used for this type of diathesis. The non-lexically specified human ACT which is typical for such type of action is generalized and cannot be expressed on the surface,[9] see ex. (11).

(11) *Tyto potraviny a bavlna **se** v České republice **nepěstují**, a tak jejich dovoz naše zemědělce neohrozí.* [*pěstovat*.deagent] (PDT 2.0)
lit.: 'This food and cotton **are not grown** in Czech Republic, so that import of them will not endanger our farmers.'

3.5.3. Disposition constructions (disp)

The reflexive form of a verb with a shift of participant (the ACT is not in the subject position) accompanied by an evaluative adverb such as *dobře* 'well', *snadno* 'easily', *pomalu* 'slowly' is called here a disposition construction.[10] The ACT is not excluded; however, if present, it is expressed by a dative form. This position characterizes the ACT as positively or negatively disposed to this action, see ex. (12).

(12) *Krásně **se** nám **bruslilo**.* [*bruslit*.disp] (PDT 2.0)
lit.: 'It **was pleasent** for us **to skate**.'

[9] We prefer the term "general" for the subject/ACT usually called "arbitrary" in generative grammar because the subject in these contexts is typical rather than arbitrary.

[10] See also (Dokulil, 1941), sometimes this construction is understood as mediopassive.

3.5.4. Recipient (passive) constructions (recip)

In these constructions, the prominent syntactic position (of subject) is filled by a participant other than an ACT; from the point of view of semantics it is a recipient, expressing usually an ADDR (in dative with three-argument verbs), sometimes a PAT (in dative with two-argument verbs); the (semi-)auxiliary verb *dostat* 'to get' (and marginally *mít* 'to have' as well) forms the analytical form with *-n/-t* participle and with agreement with the surface subject; see ex. (13). Though this construction is productive enough, it has some limitations; semantic groups of these verbs compatible with this value are given in (Daneš, 1985).

(13) *Je to asi taková práce, jako kdybyste **dostal napsán** konečný součet dlouhé řady čísel.*
[*napsat*.recip] (CNC)
lit.: 'It is a similar effort as if you **got** a finite count of a long string of numbers **written**.'

3.5.5. Some correspondences between grammateme values and Czech forms

Though due to their systemic character the resultative and recipient diatheses are considered as grammaticalized categories, they are not regularly derived from any arbitrary verb. The features *+res*, *+recip* must be included in the lexical information about the verb in the lexicon.
The restrictions on the formation of passive, deagentive and disposition constructions are of grammatical nature and they are described elsewhere.
Possible combinations of relevant values of grammatemmes and their exemplification are given in Tables 3 and 4.

3.6. The grammateme diatsynt

The reciprocal constructions of the type (14) and (16) are annotated as a syntactic diathesis of the hypothetical structures (15) and (17), respectively. In comparison with the hypothetical basic structure the number of valency participants in the reciprocal diathesis is reduced, one participant as a part of reciprocal action is shifted into another position (see (14)) or is covered by the plural noun (see (16)) in this position. Theoretical arguments as well as technical details of this description are given in (Panevová, 2007) and (Panevová and Mikulová, 2007).

(14) *Jan a Marie **se objali**.*
lit.: 'John and Mary embraced each other.'

(15) *Jan objal Marii a Marie objala Jana.*
lit.: 'John embraced Mary and Mary embraced John.'

(16) *Obě strany **se** vzájemně **obviňovaly** z používání černé magie.* (CNC)
lit.: 'Both sides blamed each other of using the black magic.'

(17) *Jedna strana obviňuje druhou stranu z používání černé magie* and vice versa
lit.: 'One side blames the other side of using the black magic' and the other way round

3.7. The grammateme factmod

The grammateme factmod (for factual modality; see bellow) partially substitutes the grammateme verbmod used in PDT 2.0. The grammateme verbmod was included in the annotation scenario of PDT 2.0 as a tentative solution of the domain of verbal modality, which requires further investigation. Three values which were defined for the grammateme verbmod (i.e. ind, cdn and imp) directly corresponded to morphological moods and did not reflect the meaning of the morphological category of mood.
After a detailed linguistic analysis (Ševčíková, 2009), a substantial difference between the functions of the indicative and conditional on the one hand and the imperative on the other turned out. The indicative and conditional express modality which affects the meaning of the verb concerned (we speak about factual modality) whereas the imperative is a marker of illocutionary force (in Czech as well as in many other languages; see (Bybee, 1985)), which concerns the sentence as a whole.[11] Thus, only the indicative and conditional have to be captured by a verbal grammateme while the imperative is to be included in an attribute of illocutionary modality (the analysis of which goes beyond the scope of this paper).
The indicative renders unconditioned (real, asserted) events while by the conditional conditioned (unreal, hypothetical) events are expressed. Beside this semantic opposition, which is subsumed under the term of factual modality, the grammateme factmod captures also the difference between two types of conditioned events, between the potential ones (expressed prototypically by the so-called present conditional) and the irreal ones (expressed by the so-called past conditional unambiguously, but often by the present conditional, which leads to ambiguity). Three values of the grammateme factmod are therefore proposed: potential for potential events, irreal for irreal events and asserted for asserted, unconditioned events; see example sentences (18) with the indicative verb form, (19) with present conditional, and (20) with past conditional, respectively.

(18) *Rekonstrukce bytu **stojí** milion.* [*stát*.asserted]
lit.: 'Reconstruction of the flat **costs** a million.'

(19) *Rekonstrukce bytu **by stála** milion.*
[*stát*.potential]
lit.: 'Reconstruction of the flat **would cost** a million.

(20) *Rekonstrukce bytu **by byla stála** milion.*
[*stát*.irreal]
lit.: 'Reconstruction of the flat **would have cost** a million.'

[11] In a compound sentence, the involved clauses can express different illocutionary forces, e.g. *Zavři dveře a já otevřu okno* 'Close the door and I open the window'. However, this issue goes beyond the scope of the present paper.

aspect / tense	proc	cpl	perf
sim	*vaří* lit.: 'she is cooking / she cooks'	–	*má uvařeno / je uvařeno* lit.: 'she has (the meal) cooked / (it) is cooked'
anter	*vařila* lit.: 'she was cooking / she cooked'	*uvařila* lit.: 'she cooked'	*měla uvařeno / je uvařeno* *uvařila* lit.: 'she had (the meal) cooked / (it) was cooked' lit.: 'she had cooked (the meal)'
poster	*bude vařit* lit.: 'she will be cooking'	*uvaří* lit.: 'she will cook'	*bude mít uvařeno / bude uvařeno* *uvaří* lit.: 'she will have (the meal) cooked / (it) will be cooked' lit.: 'she will cook (the meal)'

Table 4: Verb forms corresponding to particular combinations of the values of the grammatemes tense (in 1st column) and the values of aspect (in the 1st line). If the combination od the given values is not realized in Czech, the symbol "–" is used.

4. Assignment of the revised set of verbal grammatemes

The revised set of verbal grammatemes is to be assigned to nodes of tectogrammatical trees according to the values of the attributes nodetype and sempos as described in Section 2.2.; i.e. each of the grammatemes will be assigned to each node belonging to semantic verbs. For those verbal grammatemes which were taken over from the PDT 2.0 annotation scenario without changes (grammatemes tense, iterativeness, and deontmod), the assignment procedures used during annotation of PDT 2.0 will be applied also within the annotation of PDT 3.0. Concerning the grammateme aspect, the existing semi-automatic assignment procedure has to be reconsidered with regard to the fact that a new value perf was added to the value set of this grammateme. Rules for annotation of the newly proposed grammatemes diatgram, diatsynt, and factmod are to be specified.

- Values of the grammatemes tense and iterativeness will be assigned automatically. Rules for assigning values of the grammateme tense are based on information involved in the morphological tag. Complex sentences with an embedded clause dependent on a content clause (determined as special types of object and subject clauses) will be checked manually.
  Concerning the grammateme iterativeness, the value it1 will be assigned with nodes representing verbs with particular word-formation affixes. For iterativeness expressed by perfective forms, a new part of algorithm should be designed. With remaining nodes, the value it0 is to be filled in.

- Values of the grammateme deontmod can be assigned automatically using existing rules based on correspondences between modal verbs and the values of this grammateme. The modal verbs are, of course, used for the other domains of the modality, namely for the epistemic one; however, because of the intersection of grammatical and lexical means for an expression of epistemic modality, we have not yet included these issues in the new scenario. Therefore, all occurrences of modal verbs will still be assigned as if they express deontic modality.

- The core values of the grammateme aspect (proc and cpl) will be assigned automatically using lists of Czech verbs expressing processual and complex events, respectively. Since there are bi-aspectual verbs in Czech, manual annotation will be necessary to make a decision between the given values (if such a decision is not possible, the value nr is filled in). After that, tectogrammatical nodes with the value cpl will be checked manually whether they express a perfective event; if so, the value will be changed to perf.

- The values of the new grammatemes diatgram will be assigned according to the following rules:

  - The value res1 is to be assigned with nodes corresponding to the combination of the verb *být* in the form of 3rd person sg. and an *-n/-t* participle (see ex. (5)). The difference between the value res1 as in ex. (4) and the value pas must be treated manually.
  - Deagentive constructions as well as dispositional constructions (values deagent and disp, respectively) are marked syntactically by the presence of a node corresponding to a generalized ACT in the former case, by the coocurrence of optional ACT in dative and obligatory evaluative adverb (enumerated in a special list) in the latter case.
  - Cooccurrence of forms of the verb *dostat* 'to get' and the *-n/-t* participle is a prerequisite for assigning the value recip. Manual checking by an annotator is needed.

- The values of the proposed grammateme diatsynt will be inferred from the tree structure; for details, see (Mikulová et al., 2006).

- Assignment of the proposed grammateme factmod substantially differs from the assignment of the previous grammateme verbmod. As a direct counterpart of the morphological category of mood, the grammateme verbmod was assigned automatically using information involved in the morphological tag of the particular verb form (or, if a complex verb form occured, a combination of features from morphological tags of each of the involved tokens was considered).
  Since the grammateme factmod was proposed for those cases in which indicative and conditional verb forms express factual modality, the meanings of factual modality should be distinguished from other meanings of these moods.[12] However, with regard to the lack of formal features on the basis of which the meanings of factual modality can be distinguished from the other ones, all occurrences of the indicative and conditional will be assigned as if they express factual modality.
  To assign the values of the grammateme factmod described above, both automatic and manual annotation will be used. The decision of a human annotator will be needed especially to resolve the ambiguity of the present conditional which can express both potential and irreal events.

5. Final remarks

In the present paper, a revised set of verbal grammatemes was introduced which is intended to be used in the annotation scenario of PDT 3.0. The grammatemes tense, iterativeness, and deontmod can be assigned automatically whereas the other grammatemes require a detailed manual checking of automatically assigned values (concerning at least some of the values of these grammatemes).
Linguistic data based on the scenario revised in the domain of verbal morphological categories will serve as a solid base for practical testing of new theoretical proposals.
However, even the new annotation scenario cannot cover all issues connected with morphological meanings of verbal categories. Some of these issues were mentioned in the paper: for instance, epistemic modality or the correspondence between imperative forms and illocutionary acts of order, command etc. These topics remain open for further elaboration.

Acknowledgements
The research reported on in this paper has been supported by the project GA P406/2010/0875.

6. References

J. L. Bybee. 1985. *Morphology: A Study of the Relation between Meaning and Form*. Benjamins, Philadelphia.

F. Daneš. 1985. *Věta a text*. Academia, Prague.

M. Dokulil. 1941. Morfologické kategorie pasiva ve spisovných jazycích severských ve srovnání se spisovnou češtinou. In L. Zatočil, editor, *Hrst studií a vzpomínek*, pages 77–99, Brno. Odbočka Jednoty českých filologů.

M. Dokulil. 1962. *Tvoření slov v češtině*. Academia, Prague.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, and M. Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. Linguistic Data Consortium, CAT LDC2006T01, ISBN 1-58563-370-4.

K. Hausenblas. 1962. Slovesná kategorie výsledného stavu v dnešní češtině. *Naše řeč*, 48:13–28.

V. Mathesius. 1929. Funkční lingvistika. In *Sborník přednášek pronesených na Prvém sjezdu československých profesorů filosofie, filologie a historie v Praze 3.–7. dubna 1929*, pages 118–130, Prague.

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

M. Mikulová, A. Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá, and Z. Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 2006/30, Institute of Formal and Applied Linguistics, Prague.

J. Panevová and M. Mikulová. 2007. On Reciprocity. *The Prague Bulletin of Mathematical Linguistics*, 87:27–40.

J. Panevová, E. Benešová, and P. Sgall. 1971. *Čas a modalita v češtině*. Univerzita Karlova, Prague.

J. Panevová. 2007. Znovu o reciprocitě. *Slovo a slovesnost*, 68:91–100.

M. Razímová and Z. Žabokrtský. 2006. Annotation of Grammatemes in the Prague Dependency Treebank 2.0. In E. Atwell and N. Ide, editors, *Proceedings of the LREC 2006 Workshop on Annotation Science*, pages 12–19, Genova. ELRA.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

M. Ševčíková and Z. Žabokrtský. 2006. Systematic Parameterized Description of Pro-forms in the Prague Dependency Treebank 2.0. In J. Hajič and J. Nivre, editors, *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 175–186, Prague. Institute of Formal and Applied Linguistics.

M. Ševčíková. 2009. *Funkce kondicionálu z hlediska významové roviny*. Institute of Formal and Applied Linguistics, Prague.

J. Štěpánek. 2006. *Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat)*. Ph.D. thesis, Charles University in Prague.

[12] For instance, besides the factual modality the conditional can express deontic and epistemic modality as well. Cf. in the sentence *Kouřil bych* 'I would smoke', which actually means "I want to smoke", the conditional expresses speaker's will to do something (and belongs thus to deontic modality); an example of the epistemic use of the conditional is the sentence *Šel bych* 'I would go', which can be used instead of "I will probably go" in an appropriate context.