WDS'05 Proceedings of Contributed Papers, Part I, 72–77, 2005. *ISBN 80-86732-59-2* © *MATFYZPRESS*

Meanings of Morphological Categories on the Tectogrammatical Level

M. Razímová

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. The present paper focuses on representation of morphological meanings on the underlying syntactic level. The concept of semantic counterparts of morphological meanings, the so-called grammatemes, was introduced in Functional Generative Description in the 1960's. We suggest an elaborated system of these grammatemes, which have become a part of the tectogrammatical level of the Prague Dependency Treebank.

Introduction

Functional Generative Description (FGD in the sequel) is a multilevel description of natural language, developed since 60th of the 20th century by Petr Sgall and his collaborators at the Faculty of Mathematics and Physics of the Charles University in Prague.[1] FGD is a stratificational approach; in the current model, it works with the level of phonemics/phonetics, of morphemics and with the underlying syntactic level, the so-called tectogrammatical level. The relation between elements of a higher level and elements of the adjacent lower one is the relation between function and form [see especially *Sgall*, 1967, and *Panevová*, 1980]. In FGD, the sentence is represented as a dependency structure with a verbal core.

The tectogrammatical level, which we would like to concentrate on, is the "most semantic" one. Synonymous sentences should have the same tectogrammatical representation, while there should be a different tectogrammatical representation for each of the meanings of ambiguous sentences. From this point of view, only "semantic" information should be represented on this level, such as the function of the word in the underlying syntactic structure captured by the so-called functor (see below). In FGD, means for capturing the meanings of semantically relevant morphological categories are called grammatemes.[2] The concept of grammatemes was introduced in FGD already at the outset of the formulation of this theory and now it is further elaborated in the context of the Prague Dependency Treebank.

The Prague Dependency Treebank (PDT) has an annotation scenario built on the theoretical base of FGD. There are two versions of PDT: PDT 1.0 contains approx. 100 thousand Czech sentences annotated on the morphological (POS) level and on the level of analytic (surface) syntax between 1996 and 2000 [see *Hajič et al.*, 2001]. The system of grammatemes that we present in this paper is a part of the second version, of PDT 2.0, which is a treebank of about 50 thousand sentences annotated also on the tectogrammatical level between 2000 and 2004.[3]

Classification of nodes of the tectogrammatical tree

Types of tectogrammatical nodes

In PDT 2.0, the sentence is represented by a dependency tree structure built of nodes, representing mainly auto-semantic words and their characteristics (attribute-value pairs), and edges, representing relations between words. We focus here on the attribute called grammateme. Grammatemes are counterparts of those morphological categories which bear tectogrammatically relevant information.

[1] For first systematic formulation of FGD see *Sgall*, 1967, for further developments of this theory e. g. *Panevová*, 1980, *Sgall et al.*, 1986, etc.

[2] E.g. on the tectogrammatical level, the sentences 'He *enjoys* Peter's *newest car*' – 'He *enjoyed* Peter's *new cars*' differ just in three morphological meanings (of tense by the verb *to enjoy*, of degree of comparison by the adjective *new* and of number by the noun *car*).

[3] The PDT 2.0 will be publicly released soon by Linguistic Data Consortium.

From this point of view, grammatemes are not required at all nodes of a tectogrammatical tree since, first, not all words represented by a tectogrammatical node express morphological meanings (e.g. rhematizers or artificial t-lemmas, see below) and, second, not all morphological meanings are semantically relevant.[4]

To differentiate nodes that represent words expressing morphological categories from nodes without these meanings, the classification of tectogrammatical nodes was necessary. This classification is mainly based on information of the following two attributes of the tectogrammatical node:

- attribute t-lemma (for tectogrammatical lemma), which contains the lexical value of the node, mostly represented by a sequence of graphemes; in some cases, the value of this attribute is an artificial lemma (e.g. t-lemma #Gen; at nodes representing a participant of verb, which has no counterpart in the surface sentence structure because of its semantic generality);
- attribute functor, describing the function of the word in the underlying syntactic structure (e.g. functor ACT for the actor of an event, LOC for a free modification with the meaning of location, MOD for modal modification).

Eight types of tectogrammatical nodes are distinguished:

- root of the tectogrammatical tree is a technical node whose child node is the governing node of the sentence structure;
- complex nodes represent auto-semantic words (see below for further classification);
- atomic nodes represent rhematizers, modal modifications etc.;
- roots of coordination and apposition constructions contain the t-lemma of the coordinating conjunction or a punctuation symbol;
- dependent nodes of foreign phrases bear components of phrases, which do not follow rules of Czech grammar;
- dependent nodes of phrasemes hang on the node representing such phraseme; these nodes constitute a single lexical unit with their parent; the meaning of this unit does not follow from the meanings of its component parts;
- roots of foreign and identification phrases are nodes with special artificial t-lemmas, which play the role of a parent of a foreign phrase (t-lemma #Forn;) or a parent of a phrase having a function of name (t-lemma #Idph;);
- quasi-complex nodes stand mostly for obligatory verbal complementations that are not present in the surface sentence structure; these nodes have artificial t-lemmas.

The appurtenance of the tectogrammatical node to one of the types of nodes is stored in the attribute nodetype. The type hierarchy of tectogrammatical nodes is displayed in Fig. 1.

Semantic parts of speech

Morphological meanings are expressed only by words represented by complex nodes. Further classification of complex nodes was required.

Complex nodes are divided into four subgroups, according to the semantic part of speech. Semantic parts of speech are categories of the tectogrammatical level and correspond to basic onomasiological categories of substance, quality, circumstance and event [see *Dokulil*, 1962]: on the tectogrammatical level, semantic nouns, semantic adjectives, semantic adverbs and semantic verbs are distinguished.

Semantic parts of speech are not identical with “traditional” parts of speech. Relations between traditional and semantic parts of speech are demonstrated in Fig. 2.

The relation between semantic and traditional parts of speech can be illustrated on the example of semantic nouns. The following groups traditionally belonging to different parts of speech belong to the class of semantic nouns: (i) traditional nouns (*sister*), (ii) possessive adjectives (*sister’s* dog), (iii) nominal pronouns (*who* should come?), and (iv) nominal numerals (two *hundred* dogs):

[4] E.g., morphological category of case does not need to have a tectogrammatical counterpart since the information of this category is given by government.

(i) All traditional nouns belong to semantic nouns – we can speak about a "prototypical" relation between traditional and semantic parts of speech.

(ii) On the tectogrammatical level, possessive adjectives are represented by nouns, which they were derived from. This type of syntactic derivation [in sense of *Kuryłowicz*, 1936] is one of derivational types, which are distinguished on the tectogrammatical level. In these cases, the "lost" possessive meaning is stored in the functor APP (appurtenance): e.g. the possessive adjective *sister's* (dog) is represented by the node having t-lemma *sister* and the functor APP and belonging to the class of semantic nouns.

(iii) There is no group such as "semantic pronouns" on the tectogrammatical level. According to their function in the underlying sentence structure, pronouns belong either to semantic nouns or to semantic adjectives. Pronouns that fill the typical role of nouns as agent or patient belong to semantic nouns: e.g. pronoun *which* (as well as *this* and *he*) in the sentence 'This is the program, *which* he uses for animation' is classified as a semantic noun, while *which* in the sentence '*Which* program does he use for animation' is classified as a semantic adjective.

(iv) In the same way as pronouns, numerals are divided into semantic nouns and semantic adjectives. But whereas the same pronoun can belong to both of these semantic parts of speech according to its underlying syntactic function (see *which* in (iii)), each numeral always belongs only to one of them: e.g. numerals as *hundred, thousand* or *million* are always semantic nouns since they never play an adjectival role (and they express the nominal category of number: *million* vs. *millions*); other numerals (e.g. *ten, second*) always belong to semantic adjectives.

The specific inner structure of the class of semantic nouns is indicated by the appurtenance of the given noun to a specific subgroup: there is a subgroup of semantic nouns with denotative meaning (for

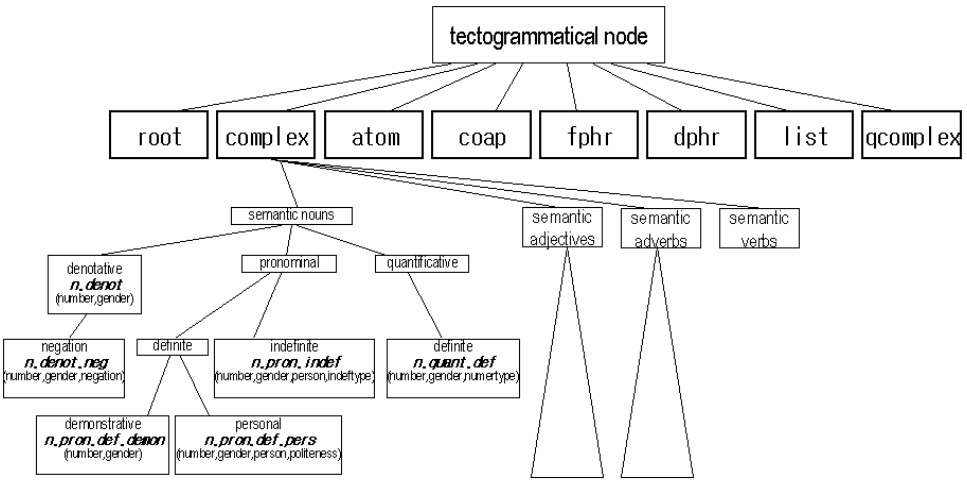

Fig.1. System of types of tectogrammatical nodes: there are values of attribute nodetype on the second level. Only complex nodes are further subdivided into four semantic parts of speech (semantic nouns, semantic adjectives, semantic adverbs and semantic verbs). The three first semantic parts of speech are further subclassified. For more details of the inner structure of semantic nouns, see Fig. 3.

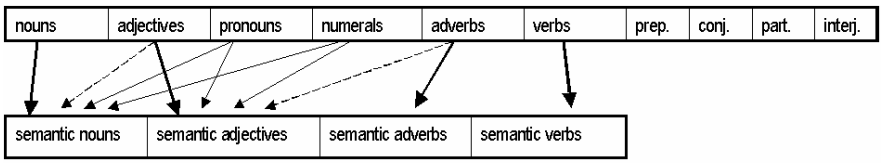

Fig. 2. Relations of "traditional" parts of speech to their semantic counterparts: arrows in bold denote a "prototypical" relation, thin arrows indicate the distribution of pronouns and numerals into semantic parts of speech and dotted arrows stand for the classification according to derivational relations.

traditional nouns and possessive adjectives), a subgroup of semantic nouns of pronominal character (mostly for nominal pronouns) and a subgroup of semantic nouns with the meaning of quantification (mainly for nominal numerals). These subgroups can be further divided.

The appurtenance of a tectogrammatical node to a concrete subgroup of the semantic part of speech is stored in the attribute sempos. The inner structure of semantic nouns with all the subgroups and corresponding values of sempos is displayed in Fig. 3.

The inner structure of the classes of semantic adjectives and semantic adverbs follows the same principles as the inner structure of semantic nouns. De-adjectival adverbs belong to semantic adjectives: they are represented by the t-lemma of their basic adjective and the lost semantic feature is expressed by the functor (e.g. the adverb in the sentence 'He runs *quickly*' is represented by the node having t-lemma *quick* and the functor MANN (for "manner") and belonging to semantic adjectives).[5] However, semantic verbs require another kind of subdivision, which has not been developed yet.

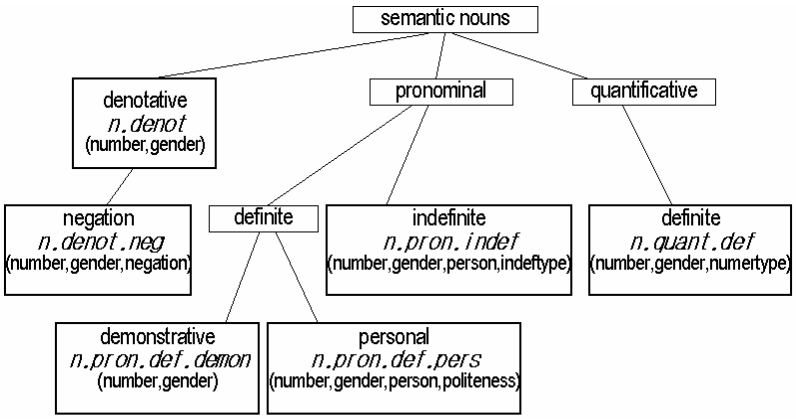

Fig. 3. Inner structure of the class of semantic nouns according to semantic features of words belonging to this semantic part of speech: the way of subdivision is demonstrated by a tree-like structure. In the leafs of the structure, there are the "final" subgroups of semantic nouns: there are the values of attribute sempos on the second line and the grammatemes belonging to the subgroup on the third line in the box.

Grammatemes and their values

There are 16 grammatemes on the tectogrammatical level. Most of them are counterparts of morphological categories that bear information relevant to the tectogrammatical level. Some of the grammatemes indicate the derivational information. The set of grammatemes that are required at the single node is captured by the value of the attribute sempos. There are grammatemes appearing at each node belonging to the given semantic part of speech (regardless to the subgroup; e.g. the grammateme number with semantic nouns), while others belong only to special subgroups (e.g. the grammateme indeftype only at the subgroup of indefinite pronominal or quantificative semantic nouns, adjectives or adverbs), see Fig. 3. Instead of listing all grammatemes and their values, we introduce three grammatemes in a more detail: grammateme number, grammateme indeftype and grammateme tense.[6]

As mentioned above, grammateme number is assigned to each node belonging to the class of semantic nouns. The values of this grammateme, sg (for singular) and pl (for plural), prototypically correspond to the morphological category of number (e.g. *dog*.sg, *dogs*.pl). However, there are cases of asymmetry between the value of grammateme and the morphological value – the grammateme value is

[5] With other types of derivation analyzed on the tectogrammatical level the lost semantic feature is stored as a value of the grammateme. This type of "derivational" grammatemes will be introduced in section "Grammatemes and their values".

[6] Detailed description of the system of grammatemes will be available in the manual of the PDT 2.0.

chosen according to the "semantic" number; this is the case e.g. with pluralia tantum: the morphological number of Czech noun *kalhoty* (trousers) is always "plural", but there is a distinctions between singular and plural on the tectogrammatical level: tectogrammatical singular in 'jedny *kalhoty*' (one pair of *trousers*) against tectogrammatical plural in 'dvoje *kalhoty*' (two pairs of *trousers*). Polite usage of the 2nd person pronouns *vy* (tectogrammatical singular in '*vy* jste přišel', '*you* came' said politely to a single person, against a "true" plural '*vy* jste přijeli', '*you* came') displays another type of asymmetry between morphological and tectogrammatical category of number.

Some grammatemes do not have their counterparts in the morphological category; they describe derivational information if the inverse operation to derivation is used on the tectogrammatical level. This is the case of indefinite pronouns (as well as pronominal numerals and adverbs), which are represented by the t-lemma corresponding to the relative pronoun (numeral or adverb, respectively). The semantic feature that would be lost by such a representation is captured by the grammateme indeftype; e.g. the negative pronoun *nikdo* (*nobody*) is represented by the node with t-lemma *kdo* (*somebody*), the "negative" semantic feature is represented by the value negat in the grammateme indeftype. All values of the grammateme indeftype and forms, which can be represented only by four t-lemmas, are displayed in Fig. 4.

T-lemma:	*kdo*	*co*	*který*	*jaký*
Values of grammateme indeftype:				
relat	*kdo*	*co*	*který, jenž*	*jaký*
indef1	*někdo*	*něco*	*některý*	*nějaký*
indef2	*kdosi, kdos*	*cosi, cos*	*kterýsi*	*jakýsi*
indef3	*kdokoli(v)*	*cokoli(v)…*	*kterýkoli(v)*	*jakýkoli(v)*
indef4	*ledakdo, leckdo…*	*ledaco, lecco…*	*leckterý, ledakterý*	*lecjaký, ledajaký*
indef5	*kdekdo*	*kdeco*	*kdekterý*	*kdejaký*
indef6	*málokdo, kdovíkdo…*	*máloco…*	*málokterý…*	*všelijaký…*
inter	*kdo, kdopak…*	*co, copak…*	*který, kterýpak*	*jaký, jakýpak*
negat	*nikdo*	*nic*	*žádný*	*nijaký*
total1	*všechen*	*všechen, všechno, vše*	–	–
total2	–	–	*každý*	–

Fig.4. The grammateme indeftype 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 nad 4th column) by only four t-lemmas on the tectogrammatical level: e.g. the node with the t-lemma co and the value indef4 of grammateme indeftype represents the word ledaco (whatever), the same node with value negat is the tectogramamtical counterpart of the word nic (nothing).

The grammateme tense is used here to demonstrate the special character of verbal gramamtemes. It may obtain one of the following three values: sim (simultaneous with the moment of speech/with other event), ant (anterior to the moment of speech/to other event), post (posterior to the moment of speech/to other event). Since the class of semantic verbs has not been subclassified yet, this grammateme occurs with all verbal nodes in the PDT 2.0. However, the values are relevant only for some of them: for nodes representing a non-imperative verbal form or a transgressive. E.g. grammateme tense with the node with t-lemma *spát* (*to sleep*) representing the analytic future form *bude spát* (*he will sleep*)[7] gets the value post, while the value sim is assigned to the node representing the transgressive in 'Hlasitě *plačíc,* odcházela…' (Lit.: Loudly *crying* she-was-leaving…). For imperative verbal forms special value nil was introduced.

[7] On the tectogrammatical level, the verbal form consisting of more word forms is represented by a single node, whose t-lemma is identical with the infinitive form of the auto-semantic verb.

Conclusion

In our contribution, we have introduced the system of grammatemes as a representation of morphological meanings in the underlying syntactic level of the large richly annotated Prague Dependency Treebank. The complex nodes express morphological meanings that are relevant for the tectogrammatical representation. Nodes of this type were further divided according to the onomasiological categories of substance, quality, circumstance and event. Whereas at the present stage of research semantic nouns, adjectives and adverbs are subclassified, inner subclassification of semantic verbs is still to be carried out in the nearest future.

Another area for future work is to separate the grammatemes that bear the derivation information (see the grammateme indeftype) from the grammatemes having their counterpart in a morphological category. The long-term aim is to describe further types of derivation: we should concentrate on productive types of derivation (diminutive building, building of feminine nouns etc.). The set of grammatemes bearing derivational information (which could be called "derivemes") will be extended in this way.

Acknowledgments. I would like to thank to my supervisor Professor Jarmila Panevová for an extensive linguistic advice. Special thanks are also due to Zdeněk Žabokrtský, co-author of the grammateme system. The work presented in this contribution was supported by the projects MŠMT ČR No. LC 536, GAČR 201/05/H014 and GA-UK 352/2005.

References

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

Hajič, J., E. Hajičová, P. Pajas, J. Panevová, P. Sgall and B. Vidová-Hladká, Prague Dependency Treebank 1.0 (Final Production Label), CDROM CAT: LDC2001T10, ISBN 1-58563-212-0, 2001.

Kuryłowicz, J., Dérivation lexicale et dérivation syntaxique, *Bulletin de la Société de liguistique de Paris*, 1936, 37, pp. 79–92.

Panevová, J., Formy a funkce ve stavbě české věty, Prague, Academia, 1980.

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

Sgall, P., E. Hajičová, and J. Panevová, The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Dordrecht, Reidel – Prague, Academia, 1986.