Morphological Meanings in the Prague Dependency Treebank 2.0

Magda Razímová and Zdeněk Žabokrtský *

Institute of Formal and Applied Linguistics, Charles University (MFF),
Malostranské nám. 25, CZ-11800 Prague, Czech Republic
{razimova,zabokrtsky}@ufal.mff.cuni.cz
http://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 is now further elaborated in the context of Prague Dependency Treebank 2.0. We present also a new hierarchical typology of tectogrammatical nodes.

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 levels, 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 level(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). On the tectogrammatical level (TL for short) of Functional Generative Description (FGD, [8], [9]), which we use as the theoretical basis of our work, this means is called grammatemes.[1]

* We would like to thank professor Jarmila Panevová for an extensive linguistic advice. The research reported in this paper has been supported by the projects 1ET101120503, GA-UK 352/2005 and GAČR 201/05/H014.

[1] Just for curiosity: almost the same term 'grammemes' is used for the same notion in the Meaning-Text Theory ([3]), although to a large extent the two approaches were created independently.

The theoretical framework of FGD has been implemented in the Prague Dependency Treebank 2.0 project (PDT, [4]), which aims at complex annotation of large amount of Czech newspaper texts.[2] 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 the results of almost two years of the work on this topic.

2 Tectogrammatical Nodes and Hierarchy of Their Types

2.1 Node Structure

At the TL of PDT, a sentence is represented as a tectogrammatical tree structure, which consists of nodes and edges.[3] Only autosemantic words have 'their own' nodes at the TL, while functional words (such as prepositions, subordinating conjunctions or auxiliary verbs) do not. Tectogrammatical node itself is a complex data structure: each node can be viewed as a set of attribute-value pairs. The attributes capture (besides others)[4] the following information:

- Attribute t-lemma contains the lexical value of the node, represented by a sequence of graphemes, or an 'artificial' t-lemma, containing a special string. The lexical value of the node mostly corresponds to the morphological lemma of the word represented by the node. The artificial t-lemma appears as a t-lemma of a restored node (that has no counterpart in the surface sentence structure, e. g. node with t-lemma #Gen), or it corresponds to a punctuation mark (present in the surface structure; e. g. node with t-lemma #Comma) or to a personal pronoun, no matter whether it is expressed on the surface or not (t-lemma #PersPron). In special cases the t-lemma can be composed of more elements (e.g. the t-lemma of a reflexive verb consists of the verbal infinitive and the reflexive element *se*: c.f. *dohodnout_se* in Fig. 3).
- Attribute functor mostly expresses the dependency relation (deep-syntactic function) between a node and its parent (thus it should be viewed as associated with the edge between the node in question and its parent rather than with the node itself).
- Attribute subfunctor specifies the dependency relation in a more detail.

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

[3] Edges will not be further discussed in this paper, since they represent relations between nodes, whereas grammatemes belong always only to one node. However, suggested classification of nodes has interesting consequences for the classification of edges.

[4] Full documentation of all tectogrammatical attributes will be available in the documentation of PDT 2.0.

- There is a set of coreference attributes, capturing the relation between two nodes which refer to the same entity.
- Attribute tfa serves for the representation of topic-focus articulation of the sentence according to its information structure.
- There is a set of grammateme[5] attributes. Grammatemes are mostly tectogrammatical counterparts of morphological categories (but some of them describe the derivation information).
- Attribute nodetype and sempos specify the type of the node.

The last two attributes serve for node typing, which is necessary if we want to explicitly condition the presence or absence of other attributes (not only grammatemes) in the node in question (for instance, tense should never be present with rhematizer nodes).[6] The proposed hierarchy (sketched in Fig. 1) consists of two levels. The top branching renders fundamental differences in node properties and behavior (Section 2.2), whereas the secondary branching (applicable only on complex nodes, Section 2.3) corresponds to the presence or absence of individual grammatemes (morphological meanings) in the node.

2.2 Division on the First Level – Node Types

Having studied various properties of tectogrammatical nodes, we suggest the following primary classification (in each node, it is captured in attribute nodetype):

- The **root** of the tectogrammatical tree (nodetype=root) is a technical node whose child is the governing node of the sentence structure.
- **Complex nodes** (nodetype=complex) represent autosemantic words on the TL (see Section 2.3 for detailed classification),
- **Atomic nodes** (nodetype=atom) represent words expressing the speaker's position, modal characteristics of the event, rhematizers etc.
- **Roots of coordination and apposition constructions** (nodetype=coap) contain the lemma of a coordinating conjunction or an artificial t-lemma substituting punctuation symbols (e.g. #Comma, #Colon).
- **Dependent nodes of foreign phrases** (nodetype=fphr) bear components of a phrase consisting of foreign words, not determined by Czech grammar; t-lemma of these nodes is identical with the surface (i.e., unlemmatized) form in the surface structure of the sentence.
- **Dependent nodes of phrasemes** (nodetype=dphr) create with their parent node one lexical unit with a meaning that does not follow from the meanings of the dependent node and of its parent.

[5] In this paper we return the term 'grammateme' as used e.g. in [7], thus we use it differently from [2], in which this term covered also subfunctors.

[6] Of course, the idea of formalizing the presence or absence of an attribute in a linguistic data structure by typing the structures is not new – typed feature structures play a central role in unification grammars for a long time. However, no formal typology of tectogrammatical nodes was ever elaborated in PDT (or even in FGD, although its usability was anticipated e.g. in [7]) before the presented work.

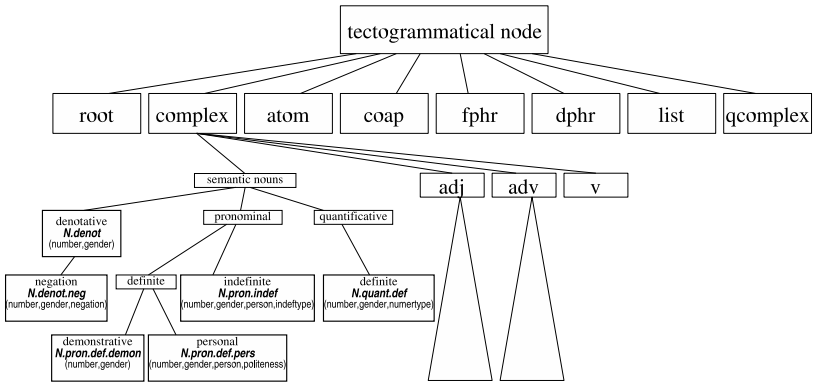

Fig. 1. Type hierarchy of tectogrammatical nodes.

- **Roots of foreign and identification phrases** (nodetype=list) bear one of the artificial t-lemmas #Forn or #Idph (regardless of the functor). The node with t-lemma #Forn is a parent of (above described) dependent nodes of foreign phrases which stand as children nodes of this Forn-node in the order corresponding to the order in the surface structure of the sentence. The node with the t-lemma #Idph plays the role of the governing node of a structure having a function of name (e.g. a title of a book or movie).
- **Quasi-complex nodes** (nodetype=qcomplex) are mostly restored nodes filling empty (but obligatory) valency slots. These nodes receive a substitute t-lemma according to the character of the complementation they stand for, e.g. the quasi-complex node with the substitute t-lemma #Gen plays the role of an inner participant, which was deleted in the surface sentence structure because of its semantic generality.

2.3 Division on the Second Level – Semantic Parts of Speech

Complex nodes (nodetype=complex) are further divided into four basic groups, according to their semantic parts of speech. Semantic parts of speech belong to the TL and correspond to basic onomasiological categories of substance, quality, circumstance and event (see [1]). The semantic parts of speech are semantic nouns (N), semantic adjectives (Adj), semantic adverbs (Adv) and semantic verbs (V). In PDT 2.0, semantic nouns, adjectives and adverbs are further subclassified.[7]

The appurtenance of a tectogrammatical node to the semantic part of speech is stored in the attribute sempos. The value of this attribute delimits the set of

[7] Semantic verbs require a different type of inner classification, which has not been developed yet. This is related to difficult theoretical questions, concerning e.g. the presence or absence of tense in an infinitival verbal expression synonymous with a (tensed) subordinate clause (mentioned also in [3]).

grammatemes that are relevant for the node belonging to the concrete part-of-speech group. The inner structure of semantic nouns is illustrated in the bottom left-hand part of Fig. 1.

The semantic parts of speech are not identical with the 'traditional' parts of speech (i.e. ten parts of speech in the Czech tradition). Traditional nouns, adjectives, adverbs and verbs belong mostly to the corresponding semantic parts of speech (but there are exceptions, mostly due to derivation; see below); traditional pronouns and numerals were distributed to semantic nouns or semantic adjectives according to their function in the tectogrammatical sentence structure, see Fig. 2.[8]

Another reason for differentiating between traditional and semantic parts of speech is that certain derivation relations are distinguished on the TL (in the sense of Kurylowicz's syntactic derivation, see [5]), the occurrence of which results in a change of part of speech. At the TL, the derived word is represented by the t-lemma that it was derived from, and the semantic part of speech corresponds to the t-lemma rather than to the original word. We illustrate this on the example of possessive adjectives and deadjectival adverbs in the following paragraphs.

Possessive adjectives as denominative derivates are represented by the t-lemma of their base nouns; sempos of these (traditional) possesive adjectives is 'N' on the TL. E.g. in Fig. 3, the possessive adjective *Mečiarova* (Mečiar`s) is represented by the node with t-lemma *Mečiar* and functor APP (expressing the 'lost' semantic feature of appurtenance).

Deadjectival adverbs are represented by adjectives; their traditional part of speech is 'adverb', while sempos is 'Adj'. E.g. in Fig. 3, *rozumně* (rationally) is represented by the node with t-lemma *rozumný* (rational).

The following types of derivation concern only the traditional pronouns and numerals. A single t-lemma corresponding to the relative pronoun is chosen as the representant of all types of 'indefinite' pronouns (i.e. relative, interrogative, negative etc). E.g. in Fig. 3, the negative pronoun *nic* (nothing) is represented by the t-lemma *co* (something) (which is equal to the relative pronoun), the semantic feature lost from the t-lemma is represented by the value of the grammateme indeftype (in this case value negat).

In a similar way, all types of (definite as well as indefinite) numerals (i.e. basic, ordinal etc.) are represented by the t-lemma corresponding to the basic numeral. The semantic feature of the numeral is marked in the value of the grammateme numertype.

3 Grammatemes and Their Values

Grammatemes belong only to complex nodes. Most grammatemes are tectogrammatical counterparts of morphological categories. Some of them describe deriva-

[8] Naturally, prepositions (which are not represented by a node on the TL) as well as conjunctions, particles and interjections (which belong to other node types than to the complex one) are not grouped into semantic parts of speech.

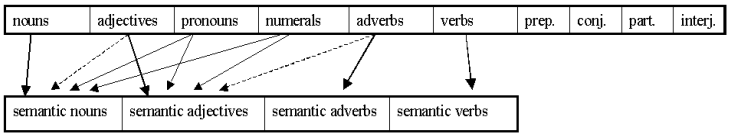

Fig. 2. Relations between traditional and semantic parts of speech. Arrows in bold indicate 'prototypical' relations, dotted arrows represent the classification following the derivation and thin arrows follow the distributing of pronouns and numerals into semantic parts of speech.

tion information. The set of grammatemes which belong to a concrete complex node is delimited by the value of the attribute sempos of this node.

There are 16 grammatemes in the PDT 2.0. We list them in the following paragraphs (the grouping is only tentative).

Grammatemes having their counterpart in a morphological category are the following: (1) number (singular, plural; N);[9] (2) gender (masculine animate, masculine inanimate, feminine, neuter; N); (3) person (1, 2, 3; N); (4) grammateme of degree of comparison degcmp (positive, comparative, superlative, absolute comparative; Adj, Adv); (5) grammateme of verbal modality verbmod (indicative, imperative, conditional; V); (6) aspect (processual, complex; V); (7) tense (simultaneous, anterior, posterior; V).

Grammatemes containing derivation information are the following: (8) numertype (basic, set, kind, ord, frac; N, Adj); (9) indeftype (relat, indef1 to indef6, inter, negat, total1, total2; N, Adj, Adv); (10) negation (neg0, neg1; N, Adj, Adv).

Other grammatemes: (11) grammateme politeness (basic, polite; N); (12) grammateme of deontic modality deontmod (debitive, hortative, volitive, possibilitive, permissive, facultative, declarative; V); (13) grammateme of dispositional modality dispmod (disp0, disp1; V); (14) grammateme resultative (res0, res1; V); (15) grammateme iterativeness (it0, it1; V).

The grammateme of sentence modality (16) sentmod (enunciative, exclamatory, desiderative, imperative, interrogative) differs from the other grammatemes, since its presence is implied by the position of the node in the tree (sentence or direct speech roots and roots of parenthetical constructions) instead of by the value of sempos.

4 Implementation

The procedure for assigning grammatemes (and nodetype and sempos) to nodes of tectogrammatical trees was implemented in ntred[10] environment for accessing the PDT data. Besides almost 2000 lines of Perl code, we created a number of

[9] There is the list of distinguished values in the parenthesis, together with the value of sempos which implies the presence of the given grammateme.

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

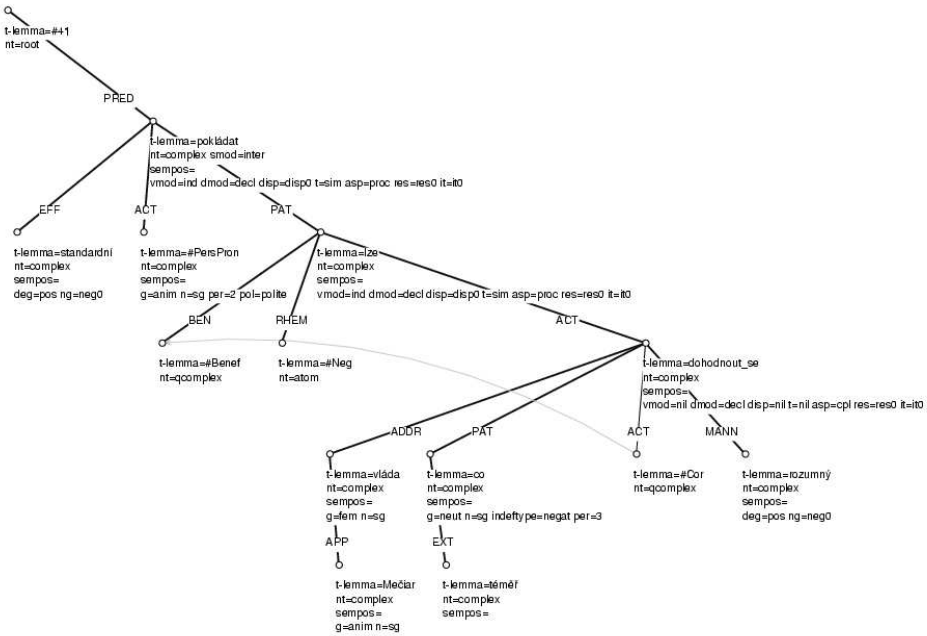

Fig. 3. Simplified tectogrammatical representation (only t-lemma, functor, nodetype, sempos, and grammatemes are depicted) of the sentence: "*Pokládáte za standardní, když se s Mečiarovou vládou nelze téměř na ničem rozumně dohodnout?* " (Do you find it standard if almost nothing can be agreed on with Mečiar`s government?).

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 on the two 'lower' levels of the PDT (morphological and analytical), most of the annotation could have been done automatically with a highly satisfactory precision. We needed only around 5 man-months of human annotation for solving very specific issues.

For the lack of space, a detailed description of the whole procedure could not be included into this paper. Just to demonstrate that grammatemes are not just dummy copies of what was already present in the morphological tag of the node, we give two examples. (1) Deleted pronouns in subject positions (which must be restored at the TL) 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). (2) 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 on all tectogrammatical data of the PDT: 3,168 newspaper texts containing 49,442 sentences with 833,357 tokens (word forms and punctuation marks). All these data, enriched with node classification and grammateme annotation, will be included in PDT 2.0 distribution.

5 Conclusions

We believe that two important goals have been achieved in the present prospect: (1) We suggested a formal classification of tectogrammatical nodes and described its the consequences on the system of grammatemes, and thus the tectogrammatical tree structures become formalizable e.g. by typed feature structures. (2) We implemented an automatic and highly-complex procedure for capturing the node classification, the system of grammatemes and derivations, and verified it on a 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.

In the near future, we plan to separate the grammatemes, which bear the derivational information ('derivemes', 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 nouns etc.). The set of derivemes will be extended in this way. The next issue is the problem of subclassification of semantic verbs.

References

1. Dokulil, M.: Tvoření slov v češtině I. Praha, Academia (1962)
2. Hajičová, E., Panevová, J., Sgall, P. Manuál pro tektogramatické značkování. Technical Report ÚFAL-TR-7 (1999)
3. Kahane, S.: The Meaning-Text Theory. In: Dependency and Valency. An International Handbook of Contemporary Research (2003)
4. Hajičová E. et al: The Current Status of the Prague Dependency Treebank. Proceeings of the 4th Internation Conference Text, Speech and Dialogue, LNAI2166, Springer (2001)
5. Kurylowicz, J.: Dérivation lexicale et dérivation syntaxique. Bulletin de la Société de liguistique de Paris, 37, (1936)
6. Panevová J.: Formy a funkce ve stavbě české věty. Praha, Academia (1980)
7. Petkevič, V.: Underlying Structure of Sentence Based on Dependency: Formal description of sentence in the Functional Generative Description of Sentence, FF UK, Prague (1995)
8. Sgall, P.: Generativní popis jazyka a česká deklinace. Praha, Academia (1967)
9. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Praha, Academia (1986)