DEPENDENCY-BASED UNDERLYING-STRUCTURE TAGGING OF A VERY LARGE CZECH CORPUS

Eva HAJIČOVÁ *

Résumé - Abstract

Le niveau tectogrammatical (sous-jacent à la syntaxe) du scénario à trois niveaux de la Banque d'Arbres de Dépendance de Prague est décrit en détail. Ce niveau d'étiquetage d'un très grand corpus du tchèque s'inscrit dans le cadre théorique de la Description Générative Fonctionnelle, dont les traits distinctifs sont (i) une structure de dépendance syntaxique avec le verbe principal comme racine de l'arbre, et (ii) l'intégration dans les structures tectogrammaticales de la prise en compte systématique de l'articulation topic-focus (structure informationnelle) de la phrase.

The tectogrammatical (underlying syntactic) layer of the three-layer scenario of the Prague Dependency Treebank is described in detail. This layer of tagging of a very large corpus of Czech is based on the theoretical framework of the Functional Generative Description, the distinctive features of which are (i) dependency syntactic structure with the main verb as the root of the tree, and (ii) the integration into the tectogrammatical structures of a systematic account of the topic-focus articulation (information structure) of the sentence.

Mots Clefs - Keywords

Banque d'arbres, dépendance, structure tectogrammaticale, étiquetage

Treebank, dependency, tectogrammatical structure, tagging

* Charles University, Prague. E-mail : hajicova@ufal.mff.cuni.cz. The research reported on in this paper has been carried out within a project supported by the Czech Grant Agency 405-96-K214 and by the Czech Ministry of Education VS 96-151.

INTRODUCTION

Since the very first steps in the formulation of the Praguian project of Czech National Corpus (CNC in the sequel) by a group of Czech linguists from the Charles University in Prague and Masaryk University in Brno, it has been quite clear to all of us that for the outcome of our project to have a broader relevance and a many-sided use we cannot confine ourselves to a mere compilation of a very large corpus of Czech texts. We have been aware that in order to make the corpus really useful for future users, be they linguists or developers of natural language processing systems of any kind, we have to design annotation schemes and develop tools that would allow us to add as much linguistic information as possible. Having the advantage of a long and fruitful tradition of theoretical and computational linguistics and inspired by the research resulting in the Penn Treebank (Marcus M. P. *et al.* 1994), the project group has decided to build the so-called Prague Dependency Treebank (PDT in the sequel).

The following three points are characteristic of PDT:

1. Its theoretical background is a dependency-based syntax (handling the sentence structure as concentrated around the verb and its valency, but containing a further dimension, namely coordination); among the reasons for the choice of a dependency based syntax I would like to stress first of all its relative economy and its perspicuous, immediate correspondence to the empirical data, cf. 2 below and, as for a detailed discussion of the motivation, (Sgall P. *et al.* 1986).
2. The nodes of the dependency tree (more exactly, of a more dimensional network) are labeled by complex symbols (consisting of lexical, morphological and syntactic parts). Thus, the label of every node contains symbols expressing all the information contained in the grammatical position of this word and relevant for a semantic (more exactly, semantico-pragmatic) interpretation. This makes the output representations, or the trees of our treebank, useful not only for practical applications such as parsing, but also for its inclusion into an integrated theoretical description encompassing all layers from the outer (phonetic or graphemic) shape of the sentence to its semantico-pragmatic representation, be it in the form of truth-conditionally based intensional semantics, or in that of a framework paying more attention to the embedding of the sentence in context.
3. The dependency tree is understood as projective (cf. Section 2 below), and its relationships to the morphemic representation of the sentence (a string of symbols) are handled by means of specific rules.

In the present paper, I give first a brief outline of the tagging scenario of the PDT (Sect. 1) and of the tectogrammatical representations (TRs) of the dependency-based Functional Generative Description (FGD), which forms the theoretical background of the tagging scheme on the deep structure level (tectogrammatical tree structures, TGTSs, Sect. 2), focusing then on the differences between TRs and TGTSs (Sect. 3). In the concluding part, some issues will be mentioned that have already emerged as open for further investigation.

1. THE TAGGING SCENARIO OF THE PRAGUE DEPENDENCY TREE-BANK

1.1. The layers of tagging

The annotation scheme of the PDT consists of the following three layers of tagging:

1. The morphemic layer, arrived at by an automatic procedure of POS tagging and by disambiguation of the rich inflectional system of Czech, contains disambiguated values of morphemic categories ((Hajič J. & Hladká B. 1998; Hajič J. & Hladká B. 1997); problems of this step are not discussed in the present paper).
2. Syntactic tags on the analytic layer (analytic tree structures, ATSs), encoding functions of individual word forms (including also e.g. punctuation marks) as they are rendered in the surface shape of the sentence ((Hajič J. ; Hajič J. *et al.* 1998; Hajičová E. *et al.* 1998a; Hajič J. & Hajičová E. 1997); a manual has been prepared for the human annotators in (Bémová A. *et al.* 1997)); at the present time, about 100000 sentences from CNC are tagged on this layer. The layer of analytic syntax does not immediately correspond to a level substantiated by linguistic theory, although it may be viewed as coming close to the level of 'surface syntax' as present in the earlier stages of FGD (see (Sgall P. 1992) as for reasons to abandon this level and thus a multistratal approach). The main difference between 'surface' and the analytical layer is that every function word and punctuation mark gets a node of its own in the syntactic network. We have been led to the inclusion of the analytic layer into the tagging procedure by two reasons: (a) it makes it possible to work with a relatively large set of syntactically tagged sentences without much delay, and (b) it allows for a comparison of the results with the outputs of several tagging and parsing procedures which have been implemented for other languages in different research centres.
3. Syntactic tags on the tectogrammatical layer (TGTSs) capture the deep (underlying, tectogrammatical) structure of the given sentence, i.e. its dependency based syntactic structure proper (see (Hajičová E. 1998b)).

A significant part of the annotation procedure is carried out automatically, in two steps (see 1.3, paragraphs (a) and (c) below). The annotators involved in the rest of the procedure have a software tool at their disposal that enables them to work with the graphic representation of the trees on the layers 2 and 3, modifying the trees in several respects, esp. in what concerns adding or changing the complex labels of the nodes, or adding and suppressing nodes.

1.2. A characterization of the TGTSs

(a) A node of the TGTS represents an occurrence of an autosemantic (lexical, meaningful) word; the correlates of synsemantic (auxiliary, functional) words are "attached" as indices to the autosemantic words to which they belong (i.e. auxiliary verbs and subordinating conjunctions to the verbs, prepositions to nouns, etc.); coordinating conjunctions remain as nodes of their own (similarly as in the ATSs).

(b) In cases of deletions in the surface shapes of sentences nodes for the deleted autosemantic words are added to the tree structure.

(c) Non-projective structures are not allowed on the tectogrammatical layer of tagging.

(d) Not only the direction of the dependency relation (dependent from the right - dependent from the left), but also the ordering of the sister nodes is specified in the TGTSs.

Each TGTS has the form of a dependency tree with the verb of the main clause as its root (to be more precise, the root of the TGTS is a special node identifying the sentence of which the given structure is the TGTS, and the node of the main verb is the only node incident to this identifier). In case of nominal 'sentences' (i.e. of constructions without a finite verb), three possibilities obtain: (i) the governing verb is added (in case of surface deletions, which is relatively rare), or (ii) a symbol for 'empty verb' ('EV') is added as the governor (e.g. *Od našeho washingtonského zpravodaje* 'From our correspondent from Washington', with the node for 'correspondent' depending on 'EV'), or (iii) the governing nominal node acts as the governor (e.g. with author names).

Each label of a node consists in the following parts:

1. the lexical value proper of the word (represented in a preliminary way just with the usual graphemic form of the word, the 'lemma'),
2. the values of the morphological grammatemes (corresponding primarily to the values of morphological categories such as modality, tense, aspect with verbs, gender and number with nouns, degree of comparison with adjectives),
3. the values of the attribute 'functor', corresponding to (underlying) syntactic functions (Actor, Objective, Means, Locative, etc., see Section 2 below; in our examples, we write the values of functors in upper case letters); as a matter of fact, in case of doubts (since the precise formulation of the criteria can only be achieved later, on the basis of analyses that will have the possibility to use a large tagged corpus as their starting point) the annotators have the possibility to indicate two different values for every functor,
4. the values of the attribute 'syntactic grammateme', corresponding to secondary syntactic functions and combined with some of the functors according to a more subtle (semantic) differentiation of these syntactic relations that is rendered on the surface first of all by prepositions and cases of nouns; this concerns the functors with the meaning of location LOC, DIR-1, DIR-2 and DIR-3 (corresponding to the questions 'where?', 'from where?', 'through which place?' and 'where to?', respectively); thus e.g.

LOC (expressed in Czech by several prepositions combined either with the locative (Loc) or with the instrumental (Instr) case of the noun) is subcategorized into *na*+Loc ('on': *na stole* 'on the table'), *v*+Loc ('in'), *u*+Loc ('by'), *nad*+Instr. ('above'), *pod*+Instr ('under'), *před*+Instr ('in front of'), *za*+Instr ('behind'), *mezi*+Instr ('among'), *mezi*+Instr ('between'), etc. As for functors having a temporal meaning, a similar subcategorization is established with the functor TWHEN (with the grammatemes AFT 'after', BEF 'before', ON 'on Monday', 'next year'). A positive or negative grammateme is attached to ACMP ('with' vs. 'without'), REG ('with regard' vs. 'without regard') and BEN ('for' vs. 'against');

5. the values of a special grammateme capture the basic information about the topic-focus articulation (TFA) of the sentence (see 2 below).

At the present stage, the tentative and preliminary inventory of the tectogrammatical labels for Czech comprises:

(a) 10 attributes for morphological grammatemes, e.g.

number:
- singular
- plural

tense:
- simultaneous
- anterior
- posterior

aspect:
- processual
- complex
- resultative

degrees of comparison:
- positive
- comparative
- superlative

(b) 47 values for the attributes of 'functor' and 'syntactic grammateme', e.g.:

functor Actor, Patient, Addressee, Locative, Means;

syntactic grammateme see point 4 above.

1.3. Illustration

The (preferred) ATS of sentence (1) is given in Fig. 1, its TGTS in Fig. 2 (with many simplifications):

(1)	*Marie*	*nese*	*knihy*	*do*	*knihovny*
	'Mary'	*'is-carrying'*	*'(the)books'*	*'to'*	*'(a)library'*

A simplified ATS of sentence (1), where '*nést*' is the infinitive of '*nese*', AuxP is the syntactic label for a preposition, and the other abbreviations correspond to

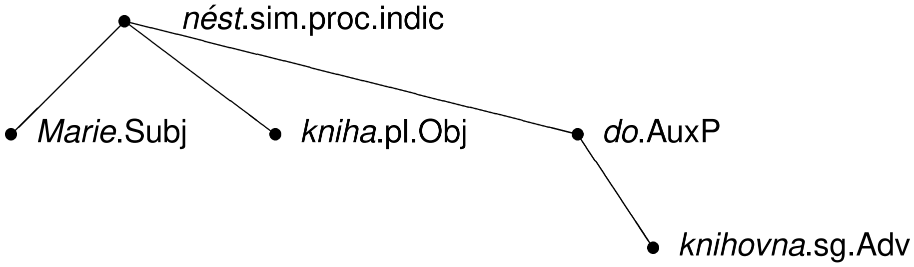

Figure 1. *Simplified ATS of sentence (1)*

morphemic values and to types of dependency on the level of 'analytic' syntax.

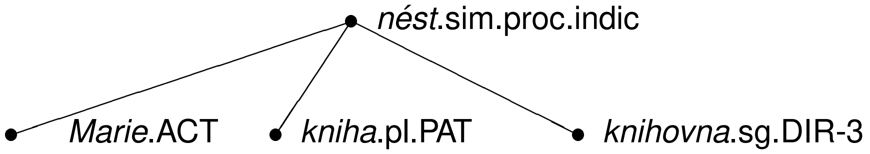

Figure 2. *Simplified TGTS of sentence (1)*

A simplified TGTS of sentence (1), with abbreviated symbols corresponding to the values of grammatemes and functors. The placement of the Patient to the left of its governor corresponds to the reading in which '*knihy*' is contextually bound, see Section 2 below.

1.4. The steps transducing the ATSs to the TGTSs

The procedure changing the ATSs into TGTSs consists of the following three steps:

(a) An automatic procedure has been formulated and implemented that carries out tree-pruning (i.e. that transforms most of the nodes that represent function words and punctuation marks into indices added to the labels of the nodes for autosemantic words), changes some of the morphological symbols (for number, tense, modality, etc.) into grammateme values, and establishes new nodes in some prototypical cases of surface deletions (see below).

(b) A handcrafted annotation procedure is being carried out now the result of which are the basic structural ingredients of TGTSs as specified for the tectogrammatical layer of annotation. According to the current plans, two sets of annotated sentences from the CNC will be achieved within the next two years, namely a smaller "fully tagged" set with complete tectogrammatical labels, and a larger set of "core" TGTSs capturing the functors, the order of nodes (sister nodes with regard to each other and with regard to their governor), and those grammatemes that can be assigned by the automatic procedure (see (a) above and (c) below).

(c) A second phase of the automatic procedure is to operate on the tectogrammatical structures achieved by step (b) and to complement the shape of the TGTSs in several respects, one of which concerns the assignment of specific values of syntactic grammatemes (expressed, in the prototypical case, by prepositions and subordinating conjunctions, see Section 1.2 point 4 above); also the assignment of the value of sentence modality in complex sentences such as (2) or (3) takes place here:

(2)

Marie	*řekla*	*Janě,*	*ať*	*přinese*	*tu*	*knihu*
'Mary'	*'told'*	*'Jane'*	*'that'*	*'she-bring'*	*'that'*	*'book'*

[Mary told Jane to bring that book.]

(3)

Marie	*se*	*ptala*	*Jany,*	*jestli*	*přinesla*	*tu*	*knihu.*
'Mary'	*'refl.'*	*'asked'*	*'Jane'*	*'whether'*	*'she-brought'*	*'that'*	*'book'*

[Mary asked Jane whether she had brought that book.]

It may be of interest to add a remark on our handling of the phenomena of grammatical agreement (concord). Czech has a rich system of verb-subject and adjective-noun agreement, which can be understood to serve as the expression of syntactic relations. While some of the values concerned can be handled as morphemic values not having any direct counterparts in tectogrammatics, others have to be taken into account also in this kind of underlying structure; among the former values there are person and number with verbs, or gender and number with adjectives, including adjectival pronouns and even certain verb forms (i.e. the 'participles' of lexical verbs in passive, conditional and preterite). On the other hand, number and probably also gender of nouns are of direct semantic relevance and have to be present in TRs, as well as numer and gender of those adjectival words that occur without a head noun (or a grammatical antecedent[1]), as illustrated by (4):

(4)

Tomáš.anim.sg	*by.sg*	*z*	*hub.fem.pl,*	*které.fem.pl*
'Tom'	*'would'*	*'from'*	*'mushrooms'*	*'that'*

najde.sg,	*přinesl.anim.sg*	*ty.fem.pl*	*jedlé.fem.pl.*
'he-finds'	*'bring'*	*'those'*	*'edible'*

[From the mushrooms he finds, Tom would bring those that are edible.]

1. By grammatical antecedent I mean that of a relative or reflexive pronoun and the 'controller' (including a noun to which a predicative complement refers in cases such as *Mary* in *Jane found Mary sitting wounded*. - It should also be noted that e.g. pronouns behave as nouns in certain cases or contexts, and as adjectives in others (in Czech e.g. *ten*[*that*] displays different properties in *Ten*.anim.sg *mě*.anim.sg *neviděl*.anim.sg [that-one didn't see me] than in *Ten*.anim.sg *kluk*.anim.sg je *chytrý*.anim.sg [that boy is cute], functioning syntactically as a noun in the former sentence, and as an adjective in the latter.

The symbols for number and gender are given here in capitals in the positions in which they have tectogrammatical counterparts, and they are given in low case letters in those positions where they are relevant just for agreement.

2. TECTOGRAMMATICAL REPRESENTATIONS IN FGD

2.1. General characterization of TRs

The tagging scheme on the deep structure level (TGTSs) is based on the dependency-based theoretical framework of the Functional Generative Description (FGD), namely on its level of the tectogrammatical representations (for motivating discussions and for more details, see e.g. (Sgall P. 1967; Sgall P. 1992; Sgall P. *et al.* 1986) a formalization can be found in (Petkevič V. 1987; Petkevič V. 1995)). It has been shown in which way the class of these representations can be specified by means of a small number of general principles accounting for the core of grammar and by specific rules for peripheral patterns.

The tectogrammatical level can be characterized as the level of linguistic (literal) meaning, i.e. as the structuring of the cognitive content proper to a particular language. On this level, the irregularities of the outer shape of sentences are absent (including synonymy and at least the prototypical cases of ambiguity) and it can thus serve as a useful interface between linguistics in the narrow sense (as the theory of language systems) on one side and such interdisciplinary domains as that of semantic interpretation (logical analysis of language, reference assignment based on inferencing using contextual and other knowledge, further metaphorical and other figurative meanings), that of discourse analysis or text linguistics, and so on, on the other.

A tectogrammatical representation (TR) of the sentence basically has the shape of a dependency tree. The edges of the tree denote the dependency relations and the nodes carry complex labels indicating their lexical and morphological values. No nonterminals and no nodes corresponding to function words (auxiliaries, prepositions, conjunctions, articles) are present in the tree. Counterparts of function words (and function morphemes) are parts of the complex symbols of the nodes. Instead of using the notion of phrase, we work with subtrees (i.e. the governor and its dependents, or all its subordinate nodes, where "subordinate" is the transitive closure of "dependent", so that "*b* is subordinated to *c*" means "*b* immediately or through mediation of other nodes depends on *c*"). The left-to-right order of the nodes of the dependency tree is used to represent the topic-focus articulation of the sentence.

TRs meet the strongly restrictive condition of projectivity: a dependency tree is projective if for every three-element set of nodes *a, b, c* present in the tree, it holds that if a depends on *c*, and *b* is placed between *a* and *c* in the left-to-right order, then *b* is subordinated to *c*. The cases of non-projective constructions in the surface structure (which are strongly limited as for their types, though not as for their frequency) can be described by means of movement rules concerning morphemics (see (Sgall P. 1997b; Hajičová E. 1998a) and the writings quoted there).

The orientation of the dependency relation (i.e. the determination of which of the pair of the nodes connected by an edge is the governor and which is the dependent) can be specified on the basis of an operational criterion:[2] the dependent node is that member of the pair that is syntactically omissible, if not in a lexically specified pair of words (as is the case with the endocentric syntagms), then at the level of word classes. Thus e.g. in ((*very*) *slow*) *progress* the syntactic potential of the heads prototypically is identical to that of the whole groups. In *Jim met Sally* nothing can be deleted, but we know from other cases that the verb is never deletable (without a specific context), whereas object can be absent e.g. with *read*, and subject (or, more precisely, its prototypical tectogrammatical counterpart, the Actor or Actor/Bearer) is absent e.g. with *rain* (the E. pronoun *IT* is just a morphemic filler, having no semantic relevance, since no other option is present, and in languages such as Czech or Latin no subject pronoun is present).

As mentioned above, function words do not occupy specific positions in the syntactic structure of the sentence as represented by the TRs. This is substantiated by the fact that articles and prepositions are, as a rule, connected with nouns, auxiliary verbs and conjunctions with verbs, and they cannot be freely modified by other elements of the sentence. Thus it appears not to be appropriate to assign them the same status as to proper (autosemantic) lexical units. Their underlying counterparts thus should be differentiated from those of the autosemantic words and denoted by more economical means than separate nodes. Thus e.g. an embedded clause such as *(We knew) that Jim arrived* is represented by a subtree the head of which is labeled by the lemma of its verb with the functor corresponding to the conjunction *arrive*.ant(erior).indic.PAT

Each of the complex symbols (i.e. the labels of the nodes of the TR) consists of (i) a lexical part, and (ii) a combination of symbols (called grammatemes) for values of grammatical categories such as number, tense, modality, etc., and of those denoting the kinds of syntactic dependency (the valency positions); the latter symbols can equivalently be written as labels of edges (or, in a linear notation, as indices of parentheses). The TR of a sentence thus can be rendered by a string of complex symbols corresponding to lexical occurrences, with every dependent included in a pair of parentheses.

Along with dependency, the TRs include a specification of several further relations. One of these is the topic-focus articulation (TFA), expressed in the surface structure mainly by an interplay of word order and sentence prosody

2. Among the formulations occurring in the discussions that have been going on for decades, this specification has the advantage of being relatively very simple. We are convinced that it can well be compared with specifications of the articulation of a sentence into phrases in constituency- based grammars. Difficulties such as those connected with the position of the infinitive e.g. in Cz. *Slyšel zvonit telefon* [*he heard the phone ringing*, lit. *to ring*] appear to be common to the two approaches.

(esp. the position of intonation centre); in the TRs, TFA is represented by the left-to-right order of the nodes (denoting the so-called communicative dynamism, i.e. the underlying word order) and by the index attached to the verb to denote whether it is contextually bound or non-bound; the nodes to the left of their governor are contextually bound, those to the right are contextually non-bound (for a definition of topic and focus as based on these primitive notions, see (Sgall P. *et al.* 1986)).

Other kinds of syntactic relations are those of coordination (conjunction, disjunction and others) and of apposition. Their interplay with dependency cannot be accounted for with full adequacy by trees if we do not want to neglect the difference between the binary dependency relations and the coordinated (and appositional) constructions, some of which may have an indefinite number of members. However, even a network with a greater number of dimensions, which in this sense can serve as the shape of a TR, can be formally described in the form of its one-to-one linearization (see (Petkevič V. 1995; Sgall P. 1997a)), namely by a string of complex symbols with two kinds of parentheses, one of which denotes dependency (in our notation these are the usual parentheses, a pair of which surrounds every dependent item), the other (brackets in our example) denoting coordination and apposition. The kinds of dependency relations are written as indices of parentheses (attached to the parenthesis that is placed on the side of the head). Also the kinds of coordination are indicated by such indices (on the righthand bracket).

2.2. Valency

The core of syntax in FGD lies in the notion of valency, i.e. of sets of kinds of dependents (see esp. (Panevová J. 1974; Panevová J. 1980; Hajičová E. & Panevová J. 1984)). Within the dependents, arguments or inner participants are differentiated from free modifications (circumstantials, adjuncts) on the basis of the following criteria:

(a) inner participants are bound to certain groups of verbs only;

(b) they occur at most once as dependent on a single verb token.

Five types of inner participants of verbs are distinguished: Actor/Bearer (*Jim runs, sits, sleeps . . . , the brook runs*), Objective (*to build a house; to destroy a house; to see a house; to address someone; to elect the chairman; to choose a spokesman*), Addressee (*to give Marya book*), Effect (*to do sth. as chairman; to elect somebodythe chairman; to choose him as chairman*), Origin (*to make a canoe out of a log*). Valency is not restrictied to verbs; among the inner participants of nouns there is e.g. Material (Partitive, *two baskets of sth.*) and Identity (*the river Danube, the notion of operator*).

There is a rich repertoire of free modifications: mostly for verbs, there are several types of Temporal circumstantials (when, how many times, since when, till when, how long, for how long), Manner, Regard, Extent (*He spent his money to the last penny.*), Norm (*in accordance with*), Criterion (*according to*), Substitution (*instead of*), Accompaniment (*with someone*), Means (Instrument), Difference (*two inches taller*), Benefit (*for someone*), Comparison (*as bright as something; brighter than sth.*), Locative, three types of Directional-1.from where, 2.which way, 3.where to, Condition, Cause, Aim (*in order to, for the sake of*), Concession (*although*), Result (*so that*); dependent mainly on nouns, there are e.g. Appurtenance (*my table, Jim's brother, Mary's car*), Restrictive (*rich man*), Descriptive (*the Swedes, who are a Scandinavian nation*).

A participant or a free modification can be either obligatory or optional with a given head: participants are prototypically obligatory (e.g. Actor and Objective with the verb *meet: Jim met Eve.*), but they can also be optional (e.g. the Addressee with the verb *to read*: *to read a book (to somebody)*). Free modifications are prototypically optional, e.g. *to be sitting* (*somewhere*) (*for a reason*) (*for some time*), but they can also be obligatory (as e.g. Manner with the verb *to behave*: *to behave badly*, or Temporal-how long with the verb *to last*: *to last for a week*, or Direction-to where with the verb *to arrive*: *to arrive at Prague*).

To decide whether a complementation of the verb is obligatory (i.e. present in the underlying structure), although deletable in the surface structure, the so-called 'dialogue test' was formulated by Panevová ((Panevová J. 1974; Panevová J. 1980); see also (Sgall P. *et al.* 1986)). It is based on an assumption that the speaker is obliged to be able to add the information he deleted in his utterance (assuming that it is an information known to the hearer), if he is asked for it. Thus the dialogue in (5) is not coherent, since the speaker A should be able to answer the question posed by the hearer B.

(5) *A: Jerry arrives tomorrow.*
B: Where to?
A: I don't know.

The dialogue test exemplified in (5) indicates that with the verb *arrive* the free modification of Direction-3 ('predictable' for the hearer, i.e. known by the speaker and deleted in the surface shape of the sentence precisely because A believes it to be easily recoverable by B) is obligatory; on the contrary, the dialogue in (6) is coherent, which indicates that with the verb *arrive* the free modification of Cause is not obligatory.

(6) *A: Jerry arrives tomorrow.*
B: Why?
A: I don't know.

The following examples of valency frames illustrate the classification of complementations in FGD (the subscript 'o' stands for 'obligatory'; the symbol of the word class allows to identify a list of free modifications specified for that class by the grammar):

bring	V	Act_o Addr Obj_o Dir-3
change	V	Act_o Obj_o Or Eff_o[3]
give	V	Act_o $Addr_o$ Obj_o
read	V	Act_o Addr Obj
rain	V	
brother	N	$Appurt_o$
glass	N	Material
man	N	
full	A	$Material_o$
green	A	

Along with the information on the valency requirements of each lexical entry, there is also other grammatical information included in the valency frames, such as surface deletability (e.g. Directional-3 with *to arrive* is deletable, Objective with *to meet* is not: *We met there* is a case of reciprocity, rather than of deletion), markers denoting an optional or an obligatory controller (e.g. Actor is an obligatory controller with *to try*, an optional one with *to decide*; Addressee is an optional controller with *to advise*, *to forbid*), and the dependent's ability to occupy certain syntactic positions (e.g. of Subject with Passivization, of a *wh*-element) or to constitute barriers for movement, and subcategorization conditions.

2.3. Topic-focus articulation in TR's

TFA is characterized on the basis of two concepts (discussed in detail in (Sgall P. *et al.* 1986; Hajičová E. 1993; Hajičová E. *et al.* 1998b) and the writings quoted there):

(a) Contextual boundness: Contextually bound (cb) items are grammatical counterparts of expressions carrying so-called given information, and contextually non-bound (nb) items refer in prototypical cases to "new" information; primarily, nb items belong to the focus of the sentence and cb items constitute the topic of the sentence.

(b) Communicative dynamism (CD): In prototypical cases the scale of CD corresponds to the surface word order, but there are secondary cases, e.g. with the most dynamic item (focus proper, which carries the intonation center of the sentence, often a falling stress) occurring elsewhere than at the end of the sentence, or the verb occupying the second position in the uppermost subtree according to language specific rules (which are more or less obligatory in German, optional in Czech).

In the TR's, CD is indicated by the left-to-right order of the nodes of the tree, in which every cb dependent is placed to the left of its head and every nb item is placed to the right of its head. The values f, t, c of the grammateme TFA indicate whether the given item is nb (in the focus), or cb (in the topic), or a contrastive topic, respectively, as indicated in the examples (7) through (9) below; the questions in brackets simulate the context for that reading of the sentence which is considered in these examples.

(7) *Father came home.*
(What about your father?)

(7') *father.t come.t home.f*

(8) *(It is) father (who) came home.*
(Who came home?)

(8') *home.t come.t father.f*

(9) *Mary went home and Fred stayed at school.*
(What about Mary and Fred?)

(9') *Mary.c went.f home.f and Fred.c stayed.f at school.f*

Since up to now the part of CNC that is being tagged contains mainly printed texts, we do not discuss the details of sentence prosody and other specific aspects here. However, it is relevant that at least the position of the intonation center of the sentence is to be identified if the given occurrence of the sentence in a given discourse is to be understood (interpreted) properly.

3. DIFFERENCES BETWEEN TRS AND TGTSS

3.1. Motivation of differences

As has been stated above, TGTSs are based on the theoretical conception of TRs; this does not mean that the resulting tagged structures are 'deteriorated' by being biased to a specific theoretical framework. On the contrary: TGTSs have a theoretically sound and empirically tested basis with very flat structure and other properties favourable for the possibility of comparison with other frameworks; this fact might be taken as an advantage, since it makes it possible to specify the properties of TGTSs in a precise and explicit way. THUS these structures will be a useful source of information also for those who work in other frameworks.

However, there are some points in which TGTSs differ from the TRs; the differences have been motivated by an effort, first, to encode complicated relations (other than pure dependency) in a straightforward way (Sect. 3.2, and, second, to preserve also those pieces of information from the surface shape of the sentence that might be of interest for future (mostly linguistic) research (Sect. 3.3).

3.2. Coordination and apposition

To specify TGTSs as two-dimensional trees, coordination and apposition are treated in a way that differs from their treatment in FGD: although coordinating conjuntions belong to function words, they retain their status as nodes (labeled as CONJ, DISJ, etc., with the lexical value of the conjunction) in the

TGTSs; the same holds for the expressions denoting an apposition. In addition, the nodes for the words standing in the coordination (apposition) relation get a special index. Thus, e.g., the bracketted shape of the TGTSs (disregarding other than structural relations) for *staří muži a ženy* ‘old men and women’ is either (CONJ.a (*starý*) (*muž*.CO) (*žena*.CO)) (for the interpretation ‘(*old men*) and (*women*)’) or (CONJ.a ((*starý*) *muž*.CO) ((*starý*.ELID) *žena*.CO)) (for the interpretation ‘(*old men*) and (*old women*)’, with an added node corresponding to the restrictive adjunct ‘*old*’, which has been deleted in the surface shape of the sentence, see below, Sect. 3.3 point 5).

This exception makes it technically possible to work with rooted trees, rather than with networks of more dimensions.

3.3. Further specific points

New attributes for the existing nodes are being established, carrying information that might be interesting for the use of the tagged corpus for further research, though it does not belong to the tectogrammatical level of linguistic description. The following issues belong here:

1. In order to capture collocations as wholes the component parts of a collocation get a positive value of a newly introduced attribute PHRi, where i is the serial number of this collocation in the sentence.
2. Special attributes COREF, CORNUM and CORSNT are introduced (for the time being, only to nouns and pronouns) to capture at least some basic aspects of (esp. textual) coreferential relations.

   The values of these attributes can be characterized as follows:

   COREF: the lemma of the antecedent,

   CORNUM: the serial number of the antecedent if it occurs in the same sentence; else, the value is NA (non-applicable);

   CORSNT: with two values: PREV (if the antecedent occurs in the previous sentences), or else NA.

   In cases of grammatical coreference (such as with the ‘subjects’ of infinitives as complements of the so-called verbs of control), the attribute COREF of the ‘restored’ subject the lemma of the ‘controller’ as its value; the attribute CORNUM then gets the serial number of the controller as its value and CORSNT gets the value NA, cf. e.g. (10):

   (10) *Rodiče* *radili* *Jirkovi* *nechodit* *tam*
   ‘Parents’ *‘adviced’* *‘George’* *‘not-to-go’* *‘there’*

   In the TGTS of (10), there will be a node added as an Actor of the verb *nechodit*, with *Cor* as its lexical value, and with *Jirka* in its COREF, 3 in its CORNUM and NA in its CORSTN.
3. ‘Restored’ nodes standing for elements deleted in the surface structure of the sentence but present in its underlying structure get marked by

one of the following values in the attribute DEL: ELID: the 'restored' element stands alone; e.g. the TGTS (disregarding other than structural relations) for *staří muži a ženy* 'old men and women' (for the interpretation '(old men) and (old women'), with an added node for the deleted occurrence of the restrictive adjunct 'old') should be (CONJ.*a* ((*starý*) *muž*.CO) ((*starý*.ELID) *žena*.CO); ELEX ('expounded' deletion): e.g. the TGTS for *velmi staří muži a ženy* 'very old men and women' (for the interpretation '(very old men) and (very old women)', with an added node for the deleted restrictive adjunct 'old') should be (CONJ.*a* (((*velmi*) *starý*) *muž*.CO) ((*starý*.ELEX) *žena*.CO).

4. Parenthetical items in the sentence without a specific syntactic relation to one of its elements get the functor PAR (e.g. *Jirka myslím*.PAR *přijde pozdě* 'George I-think will come late'). On the other hand, a parenthetical item which exhibits a dependency relation to some element of the sentence obtains a regular functor: e.g. the sentence *Jirka (podle mne) je talentovaný pianista* 'George (according to my view) is a talented pianist' gets the TGTS with *já*.CRIT (e.g. *já* 'I' with the functor of Criterion).
5. A special functor PREC is introduced to denote the syntactic function of those elements of the sentence (with the analytic function of a particle) that as a rule stand at the beginning of the sentence, have a more or less discoursive function of cohesion but do not connect clauses into complex sentences; there belong the particles *tedy*, *tudíž* 'thus', *tj.* 'i.e.', *totiž* 'as a matter of fact', etc. (e.g. *He was ill. Thus he couldn't come there.*)
6. Direct speech is distinguished by an index DSP ('direct speech') attached to the root of the TGTS of the sentence enclosed in quotation marks; if more than a single sentence is in quotation marks, an index DSPP ('part of direct speech') is attached to the root of the TGTS of the first and of the last sentence of such a direct speech.
7. Quoted word(s), if occurring in quotation marks (be they single or double) in the surface shape of the sentence, get the index QUOT, unless they constitute a sentence of its own; e.g. while the noun *pleasure* in *They call it "pleasure"* gets 'QUOT', the verb *tell* in *He told her: "Come back soon"* gets the index DSP indicating direct speech.

4. SOME ISSUES FOR FURTHER INVESTIGATION

A complete tectogrammatical tagging of a large corpus is, of course, a very demanding task and it is no wonder then that the specifications we are now working with (and which are briefly summarized above) cannot cover all subtle oppositions that should be distinguished in the representation of the meaning of the sentence. However, we have found our task very stimulating and leading to new insights concerning issues some of which either (i) are technically complex and could not yet been entirely integrated into our apparatus (as e.g. a fully automatic handling of PP attachment and other cases of morpho-syntactic ambiguity), or (ii) have not yet been analyzed in any existing grammar or monograph of Czech. Let us briefly mention here some of the open questions that are still waiting for a monographic inquiry; some of them concern the theoretical framework, some are connected only with the decisions concerning tagging and its ambiguities:

(i) Issues concerning types of valency slots:

1. Some of the functors seem to cover more than one type of syntactic relations and thus might be more subtly differentiated; this concerns e.g. the Locative (cf. the difference between *zranil se v lese* 'he injured himself in the forest' and *zranil se na ruce* 'he injured himself on his hand', the latter being in some sense closer to an inner participant (argument). However, the question remains of how to classify the Locatives *v kuchyni* 'in the kitchen', *jednání uvnitř koalice* 'discussions within the coalition' and the modification of Means (cf. the difference between *psát rukou* 'to write with hand', *na stroji* 'on the typewriter', *tužkou* 'with a pencil' and *pohnout rukou* 'to move one's hand').
2. A similar question concerns the relation between the functor Dir-3 and the so-called Intent (e.g. *šla nakoupit* 'she went shopping'). A further question of this kind is that of the difference between e.g. *pojmenovat nějak* 'to give some name' and *pojmenovat po kom* 'to name (something, somebody) after somebody'; the latter example certainly is not just an instance of the functor Manner.
3. In the valency frames of nouns, we work with the modifications of Restrictive (adjunct) and Identity. It is then an open question how to distinguish among such examples as *pan* N. 'Mister N.', *poslanec* N. 'the deputy N.', *termín sloveso* 'the term verb'. The following criterion might be applied: with two adjacent (congruent, or non-declined) nouns the second noun functions as an Identity modifier if (a) it is non-declined (e.g. *parníkem* [Instr.] *Hradčany* [Nom.] 'with the steamer Hradčany') or (b) it can be put (without a change of meaning) into a genitive case (e.g. *pojem subjekt/u* 'the notion (of) subject-Nom./Gen.'). In all other cases the first noun would then be classified as a Restrictive adjunct. However, even with this rule some intermediate cases remain: e.g. in the combinations of first name - family name, the family name may also be in genitive (esp. if the first name has a shape of a nickname: *Jan Novák*, but *Honzík Novák*/Nom.Sg. or *Nováků*/Gen.Pl.).
4. In the domain of TFA it is necessary to make more precise the notion of contrastive topic. It is also necessary to make sure whether for a given language a distinction between contrastive and non-contrastive (parts of) focus is grammatically determined, and to pay much more attention to the study of the systemic ordering of kinds of dependents ('canonical order'). The boundary line between the syntactic function of focus sensitive particles (rhematisers, focalizers) and those of other subclasses of Attitude

adverbials has to be systematically studied; up to now we distinguish between RHEM (rhematizer), ETHD (ethical dative), INTF (intensifier), ATT (attitudinal adjunct). The primary (prototypical) and secondary (marked) positions of overt focalizers (for a most recent treatment, see (Hajičová E. *et al.* 1998b)) should be taken into account. It also should be considered whether some of these functors should not be reclassified as syntactic grammatemes belonging to a single functor (in accordance with an older proposal by P. Sgall).

(ii) Coordination constructions:

It should be further investigated under which conditions a coordination construction is to be understood as a coordination of sentences or of their parts (up to now we handle such examples as *Sedlák a Bureš objevili virus L.* 'S. and B. discovered the virus L' as a narrow coordination, although such a sentence does not exclude that each of the persons discovered the virus separately).

(iii) Deletions:

Up to now, we do not restore the governing verb of a whole sentence the deletion of which is registered in the analytic trees by means of an extra node labeled as ExD. In those cases where perhaps also some of the dependents of the restored node should be restored (cf. the symbol EXPN above), we are not yet capable to specify under which conditions this restoration should take place (to avoid repeating what does not belong to the deleted position).

(iv) Issues of the lexicon and word formation:

We are aware that lexical semantics is a domain to be investigated, but up to now we only can point out some of its open problems:

1. We have not yet started to analyze questions of the composition of lexical meaning (degrees of hyponyms, etc.) and of its parts or features.
2. In the subdomain of word formation, up to now we have only worked with some of the most productive affixes and their roles (negation, some postverbal nouns and adjectives, postadjectival adverbs, possessive adjectives and pronouns).

The boundary lines of some of these groups (and of many other) have not been drawn with full adequacy. For instance, the intransitive verbs derived by the 'reflexive' particle *se*, such as *šířit se* 'expand' are treated as specific lexical units; their relationship to the base forms is only to be found in the analytic trees.

(v) Coreference in discourse:

It has been mentioned above that only the elementary cases of textual anaphora are recorded. It will be necessary to look for a more complete application of the considerations of our previous research, cf. (Hajičová E. 1993), i.e. to work more systematically with the degrees of salience of the items contained in the stock of information shared by the speaker and (according to the speaker's assumptions) the hearer(s).

(vi) Graphic symbols:

Our treatment of dashes, quotes and quoted words, direct speech, and so on, as well as of the difference between a full stop and a semicolon as marking sentence boundaries of different strengths, or of the boudaries between paragraphs, is only preliminary. It has to be studied to which degree and in which ways the corresponding graphic symbols contribute to the underlying structure of sentences (which also characterizes some aspects of the sentence as occupying certain positions in the disocurse pattern).

CONCLUSION

If compared to the prevailing present-day trends in parsing and tagging, the present scenario of PDT is promising in going deeper in the sense of including much of semantically relevant phenomena. The aim of tagging under this approach is not only to check the grammatical structure of sentences (and their well-formedness), making choice of the reading to be preferred, but also to provide an adequate input for the semantico-pragmatic interpretation of sentences and of their specification in what concerns their embedding in context.

It will of course take some time before the part of PDT equipped with tectogrammatical tags is large enough to be of actual relevance either for practical applications or for further studies. The expected application of statistically based methods should lead to a more general and efficient shape of the procedure, but even then the tagging will contain many errors of most different kinds. However, it will relatively soon be useful for authors of future monographic inquiries into Czech grammatical and textual phenomena and their relationships to those of other languages. We asume that these authors will find the sources of their analyses in this form to make it possible to achieve more systematic insight into the studied issues, to remove the individual errors and to amend the procedures.

We are aware that many questions remain open, see Section 4 above. It may be of interest to finishing by one very specific issue of this kind: if the (head verb of) direct speech is understood as the object of the verb in the introducing sentence e.g. in *He said: "I am tired."*, then how to handle a direct speech consisting of more than one sentence (e.g. *"I am tired. I cannot continue."*)? This and many other puzzles connected with the freedom of language (the speakers being free to decide for any deviation of the norm they only can think of) make it necessary to look for descriptive methods adequate to account not only for the norm, but also for most different deviations (cf. (Sgall P. in press)).

REFERENCES

BÉMOVÁ, Alla ; BURÁŇOVÁ, Eva ; HAJIČ, Jan ; KÁRNíK, Jiří ; PAJAS, Petr ; PANEVOVÁ, Jarmila ; ŠTĚPÁNEK, Jan ; UREŠOVÁ, Zdena (1997) : *Anotace na analytické rovině: návod pro anotátory [Annotations on the analytic level: instructions for the annotators.]*, Rapport technique n ° 3, ÚFAL, Charles University, Prague.

HAJIČ, Jan (1998) : "Building a syntactically annotated corpus: The prague dependency treebank", *in Issues of Valency and Meaning,* E. Hajičová (ed.), Prague, Karolinum, pp. 106-132.

HAJIČ, Jan ; HAJIČOVÁ, Eva (1997) : "Syntactic tagging in the prague tree bank", *in Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe",* R. Marcinkeviciene & N. Volz (eds.), pp. 55-68, Kaunas.

HAJIČ, Jan ; HLADKÁ, Barbora (1997) : "Probabilistic and rule-based tagger of an inflective language - a comparison", *in Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 111-118, Washington, D.C.

HAJIČ, Jan ; HLADKÁ, Barbora (1998) : "Czech language processing — pos tagging", *in Proceedings of the First International Conference on Language Resources & Evaluation*, Granada, Spain.

HAJIČ, Jan ; HAJIČOVÁ, Eva ; PANEVOVÁ, Jarmila ; SGALL, Petr (1998) : "Syntax v českém národním korpusu", *Slovo a slovesnost*, n ° 59, pp. 168-177.

HAJIČOVÁ, Eva ; PANEVOVÁ, Jarmila (1984) : "Valency (case) frames of verbs", Amsterdam: Benjamins, Prague: Academia, pp. 147-188.

HAJIČOVÁ, Eva ; PANEVOVÁ, Jarmila ; SGALL, Petr (1998a) : "Language resources need annotations to make them really reusable: The prague dependency treebank", *in Proceedings of the First International Conference on Language Resources & Evaluation*, pp. 713-718, Granada, Spain.

HAJIČOVÁ, Eva ; PARTEE, B. ; SGALL, Petr (1998b) : *Topic-focus articulation, tripartite structures, and semantic content*, Amsterdam, Kluwer.

HAJIČOVÁ, Eva (1993) : *Issues of Sentence Structure and Discourse Patterns*, Prague, Charles University.

HAJIČOVÁ, Eva (1998a) : "Movement rules revisited", *in Processing of Dependency-Based Grammars, Proceedings from the Workshop, COLING/ACL,* S. Kahane & A. Polguère (eds.), pp. 49-57, Montreal.

HAJIČOVÁ, Eva (1998b) : "Prague dependency treebank: From analytic to tectogrammatical annotations", *in Proceedings of the Conference TSD 98*, Brno.

MARCUS, M. P. ; KIM, G. ; MARCINKIEWICZ M. A. (1994) : "The penn treebank: Annotating predicate argument structure", *in Proceedings of the ARPA Human Language Technology Workshop,* M. Kaufmann (ed.), San Francisco.

PANEVOVÁ, Jarmila (1974) : "On verbal frames in functional generative description", *Prague Bulletin of Mathematical Linguistics*, n˚22, pp. 3-40; 23(1975):17-52.

PANEVOVÁ, Jarmila (1980) : *Formy a funkce ve stavbě české věty. [Forms and Functions in the Structure of the Czech Sentence]*, Prague, Academia.

PETKEVIČ, Vladimír (1987) : "A new dependency based specification of underlying representations of sentences", *Theoretical Linguistics*, n˚14, pp. 143-172.

PETKEVIČ, Vladimír (1995) : "A new formal specification of underlying representations", *Theoretical Linguistics*, n˚21, pp. 7-61.

SGALL, Petr ; HAJIČOVÁ, Eva ; PANEVOVÁ, Jarmila (1986) : *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Dordrecht:Reidel; Prague: Academia.

SGALL, Petr (1967) : *Generativní popis jazyka a česká deklinace [Generative Description of Czech and Czech Declension]*, Prague, Academia.

SGALL, Petr (1992) : "Underlying structure of sentences and its relations to semantics", *in Wiener Slawistischer Almanach,* T. Reuther (ed.), Wien, Gesellschaft zur Förderung slawistischer Studien, pp. 273-282.

SGALL, Petr (1997a) : "Valency and underlying structure. an alternative view on dependency", *in Recent Trends in Meaning-Text Theory,* L. Wanner (ed.), Amsterdam, Benjamins, pp. 149-166.

SGALL, Petr (1997b) : "On the usefulness of movement rules", *in Actes du 16e Congrès International des Linguistes (Paris 20-25 juillet 1997),* B. Caron (ed.), Elsevier Sciences, Oxford.

SGALL, Petr (in press) : "The freedom of language", to appear in *Prague Linguistic Circle Papers 4.*