A reusable corpus needs syntactic annotations:

Prague Dependency Treebank

Eva Hajičová and Petr Sgall
Center for Computational Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

e-mail: {hajicova,sgall}@ufal.mff.cuni.cz

Abstract

Prague Dependency Treebank (PDT, i.e. an annotated part of the Czech National Corpus) is conceived as a three-layer system of tags; the individual layers can be characterized as follows: (i) morphemic tagging capturing relatively disambiguated values of morphemic categories based on a full morphemic analysis of Czech; (ii) syntactic tags at the so-called analytical level, capturing the functions of individual word forms; in the analytical tree structures (ATSs), every word token and punctuation mark has a corresponding node and is analyzed as for its POS and morphemic value, as well as for the main syntactic functions ('analytical functors', 'afuns'); among the afuns, Subj, Obj, Adv are not classified in a more subtle way; (iii) syntactic tags at the tectogrammatical level (TGTSs) rendering the underlying (tectogrammatical) structure of the sentence, i.e., its syntactic structure proper (with a detailed classification of underlying syntactic functions).

In the sequel we focus on a brief characterization of the TGTSs and on issues that are specific for the PDT scenario and are crucial, especially from the linguistic point of view. These issues concern (i) the transition from ATSs to TGTSs, (ii) the assignment of the features of the information structure of the sentence (topic-focus articulation), and (iii) a tentative treatment of coreference relations. The TGTSs are based on dependency syntax; the tagging at this level is guided by the following principles: (a) a node of a TGTS represents an autosemantic (lexical) word; the correlates of synsemantic (functional, auxiliary) words are attached to the autosemantic words to which they belong; (b) in the cases of deletion in the surface shape of the sentence, further nodes are supplied into the TGTS to 'recover' a deleted word; (c) no non-projective structures are admitted in the TGTSs (they are supposed to be solved by movement rules between the ATS and the TGTS); (d) not only the direction of the dependence on the governing node (dependence to the left, dependence to the right) is taken into account, but also sister nodes are ordered (from left to right).

1. Introductory remark

Thanks to the pioneering work of a small group of linguists, among whom Geoffrey Leech with his exceptional theoretical involvement and fully competent initiative belongs to the most prominent personalities, linguistic elaboration of large corpora has become the major centre of interest. Its impact for future linguistic studies and applications (in lexicography, stylistics, literary studies and elsewhere) will keep growing, especially with a continuation of the work on tagging the corpora in grammatical and other aspects. A large corpus, if syntactically annotated, can offer a quite new level of investigations, which may use the data gained by semi-automatic tagging procedures and make them more precise by monographic analyses.

The existence of the large Czech National Corpus (initiated by F. Čermák) has allowed for the creation of the Prague Dependency Treebank (PDT), the scheme of the grammatical tagging of which is based on the theoretical linguistic framework of the Functional Generative Description (see Sgall et al. 1986, Hajičová et al. 1998); we believe that a consistent linguistic basis has helped us to develop a complex scenario that covers both the core of language and many of the more or less frequent peripheral phenomena.

2. Morphemic and analytical tagging

The first phases of the tagging procedure (see Hajič 1998) consist of morphemic and "surface" annotations, during which the intermediate 'analytical level' is achieved; the analytical tree structures (ATSs) contain a node for every token

of a word, and even of a punctuation mark, as is often the case in tagging procedures.

Before we come to a characterization of the ATSs, let us devote a few words to the morphemic level, at which each word-form and punctuation mark in the text is assigned the attributes 'word-form', 'lemma' and 'tag'. Tagging is manual with the aid of the full-screen programme sgd working in the environment of Linux (which, however, can be carried on through the mediation of some remote means, e.g. from DOS). Both the entry and the output data for the programme sgd are in the format SGML according to DTD csts. As regards the volume, the aim is to attain, in cooperation with the FI MU Brno, no less than 1 million of annotated word-forms. The programme sgd requires a preliminary morphological treatment of the text, i.e., each word-form from from it is supposed to be accompanied by a list of all possible lemmas and of their (possible) morphological categories. This assignment is done automatically on the basis of an electronic dictionary (at present the vocabulary covers some 98-99% of current newspaper or magazine texts, including names). The remaining word-forms are handled by manual tagging. Typing errors are registered and corrected. In addition to this manual POS tagging, a fully automatic procedure was designed using stochastic modelling, which has been applied to the whole Czech National Corpus; this procedure works with. Due to the rich and complex inflectional morphemics of Czech (with seven morphemic cases and tens of paradigms of declension and conjugation), the number of tags is very high: the procedure works with almost 4000 combinations of morphological values, and its the error rate is about 5%. New procedures are being developed to lower this rate (using combinations of different stochastic and rule-based methods); for more details see Hajič and Hladká (1997).

A certain approach to surface syntax has been specified in the ATSs, i.e. in structural trees the nodes of which are marked with 12 attributes each (see Hajič 1998; Bémová et al. 1997); among them, the attribute 'afun' ('analytical functor') indicates the kind of dependency of the given node on its governing (head) node. For technical reasons, we work with an added root of the tree, on which the main verb of the sentence depends (with 'afun' "pred"), and we use special devices for coordination and apposition constructions, as well as for "distant" dependency (in certain cases in which a deleted head word occurs in the sentence structure), compound (improper) prepositions and conjunctions parenthetic collocations, etc. An ATS contains nodes for all word-forms of the sentence, as well as for all symbols of punctuation.

Among issues that present difficulties for a "surface-syntactic" analysis, there are first of all those concerning the notion of Object. We do not distinguish, in the ATSs, between 'Direct', 'Indirect' and 'Second' Object, and we label as Obj also an infinitive connected with (dependent on) a predicate. Only at the subsequent stage of tagging, in the TGTSs (see Section 3 below) these syntactically different cases are distinguished. Also with adverbials only a sigle 'afun' "Adv" is used in the ATSs, a detailed classification being reserved for the tectogrammatical tagging. Similarly, the analytical representation of numerical expressions (with which a specific function of the Genitive Case problems gets involved in Czech) does not correspond to the actual syntactic patterning (cf. sentences such as *Pět žen tam už sedělo* 'Five women were already sitting there', in which the verb form *sedělo* has Neuter gender, agreeing with the numeral *pět*, rather than with the Feminine noun *žen*, which occurs here in the Genitive case, similarly as in e.g. *vlastnosti žen* 'features of women', where the Genitive clearly functions as an adjunct).

The classification of function words is relatively detailed in the ATSs, comprising, e.g., Pred, Sb, Obj, Adv, Atv (Predicate Complement, e.g. in *Našli ho spícího* 'They found him asleep'), Atr (Adjunct dependent on a noun), Pnom (Predicate Nominal with copula), AuxV auxiliary verb, Coord (Coordinating conjunction), AuxT (Reflexive particle with a 'reflexivum tantum' verb, e.g. *divit se* 'to wonder'), AuxR (Reflexive particle in a 'passive' (General Actor) construction, such as *To se dá dobře pochopit* 'One can easily understand this), AuxP (a primary preposition or a part of a secondary preposition), AuxC (a subordinating conjunction), AuxX (a comma not serving as a coordinating conj.). The annotators have also the option to indicate alternative analyses in certain cases of different possible sentence patterns without a semantic difference, e.g. AtrAtr for an adjunct of any of several preceding nouns, AtrAdv for a structural ambiguity between adverbial and adnominal dependency, or AtrObj for an ambiguity between object and adnominal adjunct without a semantic difference.

3. Dependency as the core of tectogrammatical syntax

3.1. Basic properties of tectogrammatics

A tectogrammatical sentence representation may differ from the corresponding ATS since some nodes (those corresponding to function words and punctuation marks) can be eliminated and some added (representing items deleted in the outer form of the sentence, although present in its underlying structure). Up to now, a sample of about 1000 sentences has been tagged on this level in PDT.

Dependency trees are present both in ATSs and on the level of TRs. However, in the TRs only the nodes corresponding

to lexical (autosemantic) units; function words (or, more exactly, their functions) are represented by indices of the lexical labels, i.e. by syntactic functors and by grammatemes (which mark values of tense, aspect, modalities, number, and of other grammatical categories).

While in ATSs syntactic relations are classified without many subtle differences, such as those between types of objects or of adverbials, the tectogrammatical tree structures (TGTSs) are underlying structures (basically appropriate to serve as input to semantic interpretation, see Sgall et al. 1986; Sgall 1992) and distinguish at least about 40 kinds of syntactic relations (classified in the valency grids included in the lexical entries of the head words as arguments or adjuncts, and obligatory or optional, see Panevová 1974; 1998; a detailed set of instructions for the transition from ATSs to TGTSs can be found in Hajičová et al. 2001). One significant aspect of the TGTSs is their topic-focus articulation with a scale of underlying word order; this aspect is discussed in Section 4 below. Let us just remark here for the sake of illustration that e.g. an adjective prototypically follows its head in a TGTS, even if preceding it on the surface, i.e. in the word order of the morphemic representation (a string without parentheses), cf. *malý* 'small' in (1); see Sgall (1967), Hajičová (1984; 1993).

For technical reasons, in tagging we use nodes for coordinating conjunctions (as heads of the coordinated items), although this does not exactly correspond to the theoretical specification of the tectogrammatical level (a formal treatment of which, including all combinations of dependency and coordination and based on the detailed specification of the linguistic approach in Sgall et al. 1986, was presented by Petkevič 1995). Therefore we distinguish between tectogrammatical representations proper and Tectogrammatical Tree Structures (TGTSs), see Hajičová (1998); cf. Fig. 1, i.e. a (highly simplified) underlying tree for ex. (1).

(1) Marie a Jan, kteří mají malého syna, žijí v Lomnici.
Mary and John, who have small son, live in Lomnice

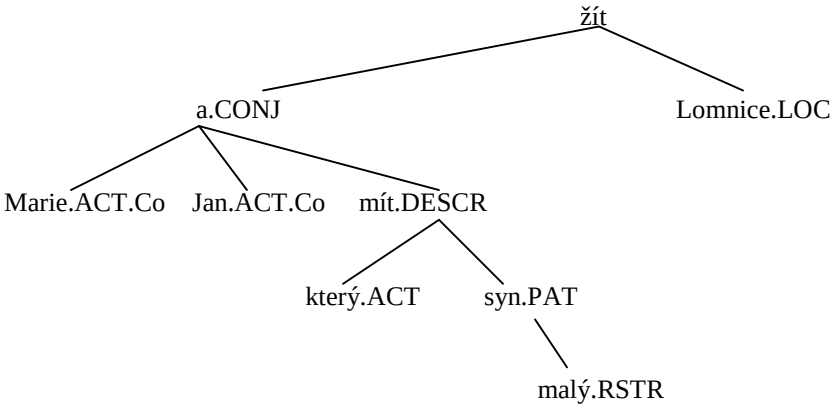

Fig. 1.
A highly simplified TGTS of (1), with functors attached to dependent nodes (Conjunction, Actor/Bearer, Patient, Descriptive adjunct, Co for the Coordinated items).

Every node of a TGTS represents an autosemantic (lexical) word; the correlates of synsemantic (functional, auxiliary) words are attached to the autosemantic words to which they belong either as syntactic functors or as values of 'grammatemes' (i.e. of morphological categories); the latter have been left out in Fig. 1, but in part are mentioned in Section 3.2 below. In the cases of deletion in the surface shape of the sentence, further nodes are supplied into the TGTS to 'recover' a deleted word (e.g. weak pronouns in Subject position - Czech is a pro-drop language - or in coordination constructions such as *červený* (*inkoust*) *a modrý inkoust* 'red (ink) and black ink'.

3.2. Linearized underlying representations

The TGTSs can be unambiguously linearized; e.g. the primary TGTS of (1) can be written as (1'), with each dependent item closed into parentheses; the subscripts (at the parenthesis oriented to the head word) indicate functors:

(1') ((Marie Jan)$_{Conj}$ ($_{Descr}$ (který.Plur)$_{Actor}$ mít ($_{Obj}$ syn.Plur ($_{Restr}$ malý)))) žít ($_{Loc.in}$ Lomnice)

Unmarked grammatemes (Sing, Pres, Declar, etc.) are not written here.

A sentence occurring in PDT can serve as a further example:

(2) Iniciátoři dosud nesehnali potřebných třicet podpisů poslanců.
Initiators hitherto have-not-gathered necessary thirty signatures of-MPs

(2') ((Iniciátor.Plur ($_{Pat}$ on))$_{Act}$ (dosud)$_{Temp.on}$ (Neg)$_{Rhem}$ sehnat.Pret ($_{Pat}$ podpis.Plur ($_{Appurt}$ poslanec.Plur) ($_{Restr}$ třicet) ($_{Desc}$ potřebných))

Note that such a deverbal noun as *iniciátor* has an obligatory Patient. With cases of coreference (anaphora) the data on the antecedent are registered in the label of the coreferential node (see Sect. 3.3 (ii)(c) below).

3.3. The automatic part of the transduction to TGTSs:

A part of the transduction from ATSs to TGTSs can be formulated as general steps, carried out automatically (see Hajičová 1998, Böhmová et al. 1999):

(i) The ATSs constitute the input of an automatic 'pre-processing' module, during which the tree structures are pruned, i.e. the nodes that are marked as auxiliary items in the ATSs get deleted, without losing any important pieces of information these auxiliary items carry. Most of the complex morphemic forms are put together (being placed in the position of the 'highest' of their parts), and the information they convey is added in the form of indices (esp. grammatemes) of the TGTS complex tags. This concerns the values of morphological categories such as tense (Preterite, Future), verbal modality (Conditional), deontic modality (with *musí* 'must', *může* 'can, may' and other modal verbs), diathesis, etc. and aspect, or gender and number with nouns, and degrees of comparison with adjectives and adverbs; they get their values on the basis of their morphemic tags (some asymmetries between forms and their respective functions are solved later, during the manual procedure). The grammateme of sentential modality (with the values ENUNC, INTERR, IMPER, DESID) is specified automatically with all heads of main clauses on the basis of the node standing for the final sentence boundary and of other data (esp. particles) present in the ATS. Also certain syntactic functions are handled by this procedure:

(a) the analytical function Subject with an active verb is converted into the tectogrammatical functor ACT (Actor/Bearer);

(b) the analytical function AuxR, denoting the particle of reflexive passive is converted into a node with the lexical value General and the fuctor ACT.

(ii) Another automatic module is being prepared, which will serve after the 'manual' handling of TGTSs (see Section 3.4 below), adding information that can be 'retrieved' automatically in the preliminary version of TGTSs:

(a) the gender and number values are cancelled with word tokens with which they only indicate agreement (adjectives in most positions, certain pronouns, numerals, etc.); thus, an adjective retains its gender value only if this value is not determined by that of a noun (e.g. in *Jen nejlepší budou vybrání* 'Only the best.Plur.Anim will be chosen');

(b) the sentence modality value with 'content' clauses (indirect speech and similar cases) is added into the respective grammateme of the head verbs of these clauses in accordance with the conjunction present, e.g. ENUNC (*že*), IMPER (*ať, nechť, aby*), INTER (*zda* and other interrogative words);

(c) certain additions are carried out which can be specified in this phase of the procedure, e.g.:

(c1) the lemma of the node carrying the functor value ACT is assigned to the grammateme COREF of an occurrence of *se* '-self' that has not yet been treated (i.e. the PAT of an active verb in the prototypical case);

(c2) the remaining nodes without lemmas (in coordinated constructions or in apposition) are assigned the lemmas of their counterparts in the given construction; e.g. in (3)
the node correspnding to the deleted second occurrence of the verb (which has been added "by hand" as governing both *Karel*.ACT and *Milenu*.PAT) gets a lemma identical to that of the lefthand coordinated item;

(3) *Jirka pozval Marii a Karel Milenu*.
Jirka invited Mary and Karel Milena

(c3) the secondary values of syntactic grammatemes (cf. Section 5 below) are added in those cases in which a preposition allows for a reliable choice: ACCOMPANIMENT.WITHOUT (*bez* 'without'), BENEFACTIVE.NEG (*proti* 'against'), DIR3.IN (*do* 'into'), etc.;

(c4) the remaining nodes corresponding to commas, dashes, quotes, etc. are deleted.

In the next stages, the automatic procedure is supposed to be enriched in various respects, to cover at least the most regular phenomena of subdomains such as:

word derivation (up to now only the deverbal adjectives, possessive adjectives and pronouns, and adverbs derived from adjectives are handled on the basis of the lemmas of the source words),

certain elementary ingredients of the build-up of the lexicon, which should contain several kinds of grammatical data especially including the valency frames or grids),

the development of the degrees of activation of the 'stock of shared knowledge' (see Hajičová 1993) as far as derivable from the use of nouns in subsequent utterances in a discourse.

3.4. The intellectual part of underlying tagging

The following operations can only be performed intellectually, before further analysis helps to find reliable criteria to identify specific contexts in which secondary functions occur:

(i) The analytical functions (such as Subject, Object, Adverbial, Attribute), expressed by case endings, subordinating conjunctions and prepositions, are changed into corresponding functors; e.g. Dative with the ATS value 'object' primarily yields ADDRESSEE, with an adverbial it yields BENEFACTIVE, Cz. *aby* 'for' or *na* 'to' yields Objective with ATS objects and AIM or LOC, respectively, with adverbials. The syntactic grammatemes accompanying LOC (corresponding to the primary functions of prepositions such as *v* 'in', *na* 'on', *pod* 'under', *mezi* 'between', and so on) are left for further treatment (the original preposition is retained as the value of a specific attribute in the complex symbol of the noun; cf. Section 3.3 (ii)(c) above as for the subsequent automatic step). The assignment of syntactic grammatemes is limited to those cases in which their values are the prototypical functions of the corresponding morphemic means, such as the prepositions mentioned above; peripheral cases (as well as certain other issues, esp. those mentioned in section (iv) below) are accounted for only in a smaller part of PDT, in the 'model collection'.

(ii) Nodes for the deleted items are 'restored' either as pronouns (including specific symbols for a 'General Participant', for a 'Controllee' and for an 'Empty Verb' (with the non-verbal heads of sentences that are neither Vocatives, nor such pure denominations as nominal headings) or the attribute 'lemma' is left vacant for further treatment (e.g. in coordinations, see Section 3.3 (ii) (c2) above).

(iii) The topic-focus articulation of the sentence is accounted for by means of three values of the corresponding attribute, namely F for 'focus' (more exactly: contextually non-bound), T for non-contrastive (part of) topic (contextually bound) and C for 'contrastive (part of) topic; this part of the annotations is discussed in more detail in section 4 below.

(iv) With possessive adjectives and pronouns dependent on nouns, the number and gender values of their bases are taken as the values of their respective grammatemes:

jeho 'his' gets the values SING, ANIMATE (or INANIMATE or NEUTER, according to the context, i.e. to the gender of the antecedent,
její 'her' gets SING, FEMININE,
jejich gets PLUR and the appropriate gender,
můj 'my' gets SING and either ANIM or FEM,
matčin 'mother's' gets SING, FEM, and so on.

The annotators use a specially designed 'user-friendly' software that enables them to work directly with diagrammatic shapes of trees.

4. Topic-Focus articulation

Topic-focus articulation (TFA) is treated after the structure of TGTSs has been built; however, when making the decisions about the values of the attribute of TFA, in certain cases the surface shape of the sentence and its ATS (which is hidden on the screen, but always accessible for the annotators) has to be taken into account.

The principles of the description of TFA in PDT, which is seen as based on the 'aboutness' relation (Focus being asserted to hold about Topic in a positive declarative sentence), are as follows (cf. Sgall et al. 1986, Hajičová et al. 1998):

(a) the label of every node of a TGTS has an index concerning the contextual boundness of the given word token: F (contextually non-bound, primarily in Focus), T (contextually bound, primarily in Topic), or C (contrastive (part of) Topic); in the underlying word order, every node assigned F follows its head node and every node assigned T or C precedes its head (exceptions are listed in point (i)(e) below);
(b) not only the just mentioned orientation of the dependence on the governing node (dependence to the left with T or C, dependence to the right with F) is taken into account, but also sister nodes are ordered (from left to right);
(c) no non-projective structures are admitted in the TGTSs (they are supposed to be solved by movement rules between the ATS and the TGTS); as for the tagging of non-projective ATSs, see section (ii) below. The automatic preprocessing procedure, preceding the manual tagging, assigns F to every node; this value is changed manually into T or C if necessary, according to the instructions below. If the verb has a complex form in the ATS, then in the TGTS its node is typically placed in the position that in the ATS is occupied by the lexical, rather than by the auxiliary, verb (i.e. the position of the infinitive or the participle is decisive).

Further general rules for the transduction to TGTSs:

(i) (a) every node that depends on the verb from the left in the ATS is assigned T, as a rule; the value F remains only where it can be clearly recognized that "new information" or a new relation is involved; sentential stress would be placed there in the spoken form of the sentence (falling pitch, intonation centre), e.g. *TÁTA přišel* 'FATHER came'; in the written Czech texts (especially in the intellectualized, technical ones, although not e.g. in written representations of speech), the rightmost position gets F quite regularly;

the value C is asigned to the contrastive part of the topic (which expresses an element taken from a set of alternatives; often this concerns sentences in which the verb is preceded by a part different from its Actor), e.g.: *Jedině s úspěšnými vzory*.C *se můžeme poměřovat*. Lit. 'Only with successful models.C (we) can compare ourselves'; *Jirkovi*.C *to Martin nedal*. Lit. 'To George it Martin did-not-give'; *Janu*.C *Marie neviděla*. Lit. 'Jane(Acc.) Mary(Nom.) did-not-see';

(b) in the prototypical case, a node that depends on the verb from the right and occupies the rightmost position in an ATS, gets the value F;

(c) any verb and what is between it and the rightmost position (see (b) above) gets, in principle, F (this also concerns semantically "poor" verbs, such as *být* 'be', *mít* 'have', *činit* 'do' etc.); if the node represents a unit repeated (not necessarily verbatim) from the preceding text (within the sentence as well as from previous context), it obtains T;

(d) the nodes that are more deeply dependent (such as an attribute, etc.) are, as a rule, assigned F (for exceptions, see (c) above, on repeated units);

(e) as a rule it holds, in the TGTSs, that every node having T or C depends on the left from its head node, and every node having F depends on the right from its head, with the following exceptions:
a focus sensitive particle with F precedes its head node if the latter has F (see (iii)(k) below);
a 'proxy focus' has T, but is placed to the right of its head in the TGTS, see (ii)(b) below; another case in which an item with T follows its head is mentioned in (iii)(g);
the node of a coordinating conjunction is not assigned any value in its attribute TFA, i.e. it is assigned NIL by the automatic preprocessing; the preprocessing inserts F as the TFA values of all the other nodes, which then are changed to T or C, in accordance with the instructions specified here.

(ii) The treatment of structures that are non-projective at the analytical level (i.e. of structures containing discontinuous subtrees):

(a) the node occurring at the leftmost position of a non-projective construction in an ATS will be assigned the symbol

C(ontrast) instead of T (for exceptions, see point (c) below), that is to say, its contrastive use is assumed; this node is placed in the projectively leftmost position (all nodes that depend on it follow and obtain T or F according to the rules concerning the remaining nodes); clitics will also be placed in a projective way, yet they obtain index T (rather than C); typical examples are (4) and (5):

(4) K jásotu.C není nejmenší důvod.
For jubilation (there) is-not the-slightest reason'

(5) Jirka ti ho dnes nezačne číst.
George you.Dat it.Accus today will-not-start to-read
'George will not start to read it to you today.'

A contrastive node depending to the left gets C, even if it is not in a non-projective position, e.g.: *Jeho*.C *jsem neznal; ji*.C *jsem poznal hned.* 'Him I didn't know; as for her, I recognized her at once'.

(b) If the first F node is deeply embedded (i.e. it does not depend directly on the verb) and follows the verb, its governing node (having the index T), a 'proxy focus' in the sense of Hajičová et al. (1998), is placed (as an exception) to the right of its own governing node; an example is (6), which may serve as an answer to *Kterého učitele potkal Pavel?* 'Which teacher did Paul meet?':

(6) Pavel.T potkal.T učitele.T angličtiny.F.
Paul met the-teacher of-English

(c) the node n placed in a non-projective position to the left of its head node is usually assigned C, but it may get T if it is not placed at the beginning of a clause (or if other factors show that it is not the topic proper, i.e. that it would be difficult to understand the clause as 'speaking about n') and at the same time, the governing word is 'quasimodal', be it a simple word (*plánovat* 'to plan', *rozhodnout_se* 'to decide to'), or a phraseme, with a quasimodal meaning (*mít čest* 'to have the honour to', *pokládat si za čest* 'to consider it a honour to', *projevit zájem o* 'to express interest in'); a list of such phrasemes and quasimodal verbs is being compiled, step by step; the words dependent on a phraseme are placed as dependent on its head, cf. (7):

(7) *Včera*.T *o Marii*.T *projevil zájem Rudolf*.F
Yesterday about Mary expressed interest Rudolf

(d) a predicate complement at the beginning of a clause gets C, as e.g. in (8):

(8) *Jako správné*.C *prověřil*.F *čas*.F *rozhodnutí*.F *Martino*.F.
As right verified time decision.Acc Martha's
'The time showed that it was Martha's decision that was right.'

(iii) The assignment of the values T and F is further guided by the following points:

(a) The value T is assigned to the nodes restored in a TGTS as having been deleted (elided) in the outer form of the sentence; this concerns above all cases of coreference (even with verbs); exceptions occur with coordination, see point (g) below.
(b) Indexical expressions - *já, my, teď, můj, náš, tady, letos, zítra* 'I, we, now, my, our, here, this year, tomorrow,' etc., obtain, as a rule, the value T; it is only with contrast, as bearers of intonation centre (sentential stress, mostly at the end of the sentence) that they are assigned F.

(c) General pronominal words such as *někdo, něco, jednou, nějaký* 'someone, something, once, some' get the value F (with the same exception as with the verb, see (i)(c) above);

(d) An adjunct expressed by an adjective, noun or pronoun (except for indexical expressions) and dependent on a noun is assigned F in the prototypical case (though it may stand to the left of its head noun in the surface word order); the value T is assigned only if this attribute is repeated or obvious from the preceding context:
starý dům 'old house' --> *dům starý.F*, but:
dům starý.F a starý.T park 'old house and old park'.

If there are more than one adjuncts to the left of their governing noun, they are 'moved' from the left to the right in a mirror-like way, cf. the following examples, with which the prime indicates the TFA values and the underlying word

order:

(9) Koupila hezký malý obrázek.
she-bought nice small picture
(9') Koupila obrázek.F malý.F hezký.F

(10) Prozradíte další elitní jména.
You-will-disclose further distinguished names
(10') Prozradíte.F elitní.T jména.T další.F

Note that in (10') *jména*.T stands to the right of the verb, as a proxy focus, see point (ii)(b) above.

(e) The value T is assigned to the weak forms of pronouns (*tě, ti, ho, mu, mi,* mostly also *mě* 'of you, to you, him, to him, to me')*;* the same holds for *mně, ji, jí, jim, jej, nás,* etc. 'to me, her, to her, to me, him, us' unless sentential stress is placed on them, and for other clitics.

(f) The strong forms of pronouns (*tebe* 'you.Accus', *jemu*.Dat 'him', etc.) obtain the value F if they (or their prepositions, as the case may be) carry sentential stress when the sentence is pronounced (falling, i.e., the intonation centre, not only an optional rising, contrastive, stress - this would be assigned C); pronouns following a preposition have always strong forms, so that the form is not decisive in these cases for the assignment of the TFA value and they can get the value T or C, as e.g. in (11), with capitals denoting the intonation centre:

(11)(a) Pro něj.C to přinesu ZÍTRA.
for him it I´ll-bring tomorrow.
'For him I'll bring it TOMORROW.'

(b) Zítra to přinesu PRO něj.F.
tomorrow it I´ll-bring for him
'Tomorrow I´ll bring it for HIM.'

(c) Zítra to pro něj.T PŘINESU.
tomorrow it for him I´ll-bring
'Tomorrow I´ll-BRING it for him'.

(d) ZÍTRA to pro něj.T přinesu.
tomorrow it for him I´ll-bring
'TOMORROW I´ll-bring it for him.'

(g) As mentioned above, the symbols treated as heads of coordination structures (i.e. the conjunctions) have neither T nor F; typically the coordinated nodes with T precede the conjunction and those with F follow it; only in a case of restored deletion it can happen that the left part of the pair has F while the right one has T, e.g. *červené*-F [*víno*-F] a *víno*-T *bílé*-F 'red [wine] and white wine', where the first occurrence of the noun is elided; this, however, has F, while the second, undeleted, occurrence has T; in such a case, both the nodes (with F and T) are placed to the right of the conjunction, F preceding T.

(h) The inserted node is always (with the exception just stated in (g)) placed to the left of its governing node and it has index T.

(i) As regards complex sentences, the following rules hold:
(i1) in a coordinated (compound) sentence, each coordinated clause has a TFA of its own, cf. (12):

(12)(a) Tom.T přinesl.F knihy.F a pak.T [on.T] odnesl.F noviny.F
Tom brought books and then [he] took-away newspapers

(b) Knihy.C [on.T] odnesl.F a noviny.C [on.T] přinesl.F.
books [he] took-away and newspapers [he] brought

(i2) as a rule, the head verb in a direct speech clause gets F, and the verb in the introductory clause gets either F or T, in accordance with the context; this holds true also in cases in which the introductory clause follows the direct speech:

(13)(a) Jirka.T řekl.F (or T): "Je.F dobře.F"
Jirka said it's fine
(b) "Je.F dobře.F," řekl.F/T Jirka.F/T
it´s fine said Jirka

(i3) In a complex sentence (with subordination), the dependent clause usually preserves its position to the left (with T) or to the right (with F) according to the surface ordering unless its intonation or stress are marked. If the verb of an adverbial clause is placed before the governing verb of the main clause and is in a contrastive position (i.e. it has a rising intonation contour, at least as an optional feature), the dependent verb gets C and remains to the left, as in e.g. in (14):

(14) Protože se jim kniha líbila.C, všichni ji dočetli.F
because Refl them book pleased all it read-through
'Since they enjoyed the book, all of them read it to the end.'

(i4) If some node in a relative clause obtains C, then this node should be placed to the left of the relative word introducing the relative clause.

(j) Among adverbial adjuncts ('free modifications'), MANN and MOD (Manner and Mode) have as a rule F (even if they are placed to the left of their heads on the surface, as the so-called adverbial attribute: *rychle řekl* lit. 'quickly he-said'*; snad řekl* 'perhaps he-said'*; dobře řekl...* lit. 'well he-said'); ATT (Attitude) has usually T (*Naštěstí přišli* 'Fortunately they arrived').

(k) The position of a focus sensitive particle, including Neg, and the value of its TFA attribute are decided as follows:

(k1) if its governing node has the value F, then the particle is assigned F and depends on its head from the left, cf. *Dnes Jirka nepřišel* 'Today Jirka has not turned up' (with the verb included in the 'focus' of the particle);

(k2) if its head has T, the particle can depend from the left and then it has T (*Dnes Jirka nepřišel proto, že je nemocen* 'Today Jirka has not turned up because he is ill'), or from the right (with F), when the head is outside the focus of the particle (*...nepřišel s omluvou, ale s vysvětlením* '...he has not turned up with an excuse, but with an explanation'), or when, as the case may be, there is nothing else than the rhematizer in F.

(l) If in a verb complex an auxiliary verb (including the modal verbs proper) belongs evidently to the focus, the lexical verb is placed in the focus part and is assigned F, even if its lexical value is contextually bound; thus, (15) gets the representation sketched in (15'):

(15) Uděláno to už MÁ.
Done it already he-has
'He HAS already done it'

(15') (on.t) (to.T) (už.T) (udělat.Perf.F)

5. Conclusions and prospects

Almost 100 000 sentences from the Czech National Corpus have obtained their 'analytical' annotations, and we expect to get about 5000 sentences annotated by their TRs before the end of the year 2001.

Neither the automatic nor the manual part of the tagging can achieve a complete shape of tectogrammatical representations. A number of questions concerning different types of grammatical information requires further examination. In some cases, this concerns issues needing more refined classification and, therefore, further empirical research; thus, e.g., the disambiguation of the functions of prepositions and conjunctions can only be completed after lists of nouns and verbs with specific syntactic properties are established. However, the annotated corpus will offer a suitable starting point for monographic analysis of the problem concerned. In other cases technical adjustments are needed for problems a more complete solution of which is known, but has not been applied at this stage owing to their exacting nature. An overview of both kinds of these open questions is being prepared for publication.

We do hope that even though achieved in this preliminary way, the syntactic (as well as the morphemic) annotations of

parts of the Czech National Corpus will be useful in that the present form of the Prague Dependency Treebank makes it possible for interested researchers to collect large sets of data relevant for their research in a much more easy and rapid way than was possible using manual excerption. Any monographic research oriented at an issue from grammar, style or present-day development of Czech can then bring proposals how to amend the classification of the given set of data and how to bring the annotation procedure to a higher level, perhaps also to raise the degree of its automation. Whenever possible, also statistical methods will be used; specific combined procedures are being tested, based on statistical and structural approaches.

A theoretically substantiated labelling of the underlying representations can be gained in this way, distinguishing between different kinds of objects and adverbials, between meanings of function morphemes, topic and focus, and so on. The result will be much more complex than that of a parser or tagger of the usual kinds: not only the grammatical well-formedness and some kind of surface structure will be checked, but disambiguated representations of sentences will be achieved, which would constitute an appropriate input for a procedure of semantic(-pragmatic) interpretation. Although these representations will be underspecified in the points in which the sentence structure is not fully specific (i.e. in cases of indistinctness, with "systematic ambiguity", scopes of quantifiers, and so on), they may constitute a large set of sentence analyses that would document the degree of adequateness of the underlying linguistic framework, in our case, the Functional Generative Description, or of its future modifications.

References

Bémová A et al 1997 *Anotace na analytické rovině: Návod pro anotátory* [Annotations on the analytical level: Instructions for the annotators.] Technical report UFAL TR-1997-03. Prague, Charles University. English translation: Technical report UFAL/CKL TR-2001-09.

Hajič J 1998 Building a syntactically annotated corpus: The Prague Dependency Treebank. In Hajičová E. (ed), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*. Prague Karolinum, pp 106-132.

Hajič J and B Hladká 1997 Probabilistic and rule-based tagger of an inflective language – a comparison. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C. pp 111-118.

Hajičová E 1984 Presupposition and allegation revisited. *Journal of Pragmatics* 8:155-167; amplified in: Sgall (1984), 99-122.

Hajičová E 1993 *Issues of sentence structure and discourse patterns.* Prague, Charles University.

Hajičová E 1998 Prague Dependency Treebank: From analytic to tectogrammatical annotations. In P. Sojka et al. (eds.) *Text, speech, dialogue. Proceedings of the Conference TSD* 98. Brno, Masaryk University, pp 45-50.

Hajičová E, Panevová J and P Sgall 2001 *A manual for tectogrammatical tagging of the Prague Dependency Treebank.* Technical report UFAL/CKL TR-2001-10.

Hajičová E, Partee B H and P Sgall 1998 *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Dordrecht, Kluwer.

Panevová J 1974 On verbal frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics* 22:3-40, 23(1975):17-52; a revised version in *Prague Studies in Mathematical Linguistics* 6, 1978, pp 227-254.

Panevová J 1998 Ještě k teorii valence [Valency theory revisited]. *Slovo a Slovesnost* 59:1-14.

Petkevič V 1995 A new formal specification of underlying structures. *Theoretical Linguistics* 21:7-61.

Sgall P 1967 Functional sentence perspective in a generative description. *Prague Studies in Mathematical Linguistics*, 2, pp. 203-225.

Sgall P (ed.) 1984 *Contributions to functional syntax, semantics and language comprehension.* Amsterdam: Benjamins - Prague: Academia.

Sgall P 1992 Underlying structure of sentences and its relations to semantics. In T. Reuther (ed.) *Wiener Slawistischer Almanach.* Sonderband 33, pp 273-282.

Sgall P, Hajičová E and J Panevová 1986 *The meaning of the sentence in its semantic and pragmatic aspects*, ed. by J L Mey. Dordrecht:Reidel - Prague:Academia.