Prague Dependency Treebank: From analytic to tectogrammatical annotations

Eva Hajičová
Charles University, Prague

Abstract

The Prague Dependency Treebank is conceived of as an annotated corpus of written Czech, comprising three layers of annotations. In the present paper, we focus on a more detailed description of the structure and contents of the tectogrammatical syntactic trees (underlying sentence representations) and a specification of the transition from the analytic syntactic tree to the tectogrammatical one.

1. An Overview of the Prague Dependency Treebank

The project called The Prague Dependency Tree Bank (PDT) was inspired by the activities resulting in the Penn Treebank: the aim is to achieve a complex annotation of (a part of) the Czech National Corpus (CNC), the creation of which is under progress at the Department of Czech National Corpus at the Faculty of Philosophy, Charles University (the corpus currently comprises about 70 million tokens or word forms) so that the corpus (or at least a representative part of it) would contain not only data on part-of-speech appurtenance of the individual lexical occurrences, but also information on grammatical (syntactic) values.

The annotation scheme of PTB works with three layers of tagging:

(1) Morphological (POS) tagging (with more than 3000 labels) including disambiguated lemmas (described in detail in Hajič & Hladká, 1997).

(2) Analytic syntactic tagging (see Hajič, 1998) the result of which are dependency trees with nodes labelled (in addition to

the POS tags and lemmas, see (1) above) by the word forms together with tags representing the syntactic relations; the analytic layer is conceived of as a step towards (underlying, tectogrammatical) syntactic representations and in the current phase of the project, about 30000 Czech sentences taken from the corpus have been annotated manually on this layer.

(3) Tectogrammatical tagging, resulting in dependency trees the nodes of which are labelled by the autosemantic lexical items of the sentence with tags representing the syntactico-semantic (tectogrammatical, TR's, in the sense of the Functional Generative Description, see e.g. Sgall et al., 1986) relations such as Actor/Bearer, Patient, Addressee, Effect, Origin, and circumstantial modifications of different kinds.

2. The Tectogrammatical Layer of Annotations

The form of the tectogrammatical syntactic trees (TSTs) is basically conceived of in accordance with the theoretical assumptions of FGD. The TSTs have the shape of a dependency tree with the verb as the root of the tree and its daughter nodes representing nodes depending on the governor (on each layer of the tree). The two dimensions of the tree represent the syntactic structure of the sentence (the vertical dimension) and the topic-focus articulation of the sentence, based on the underlying word order (the horizontal dimension).

In comparison to the analytic trees, the TSTs are guided by the following principles:

(a) a single node of a TST may be a representation of more than one word; only autosemantic words have a node of their own, while the correlates of functional words (auxiliaries, prepositions etc.) are attached as indices to the autosemantic words to which they "belong" (auxiliaries and conjunctions to lexical verbs, prepositions to nouns, etc.);

(b) nodes are added in case of clearly specified deletions on the surface level;

(c) non-projectivity (i.e. crossing of edges with each other or with perpendiculars incident to nodes) is not allowed;

(d) analytic functions are substituted by tectogrammatical functions (such as Actor/Bearer, Patient, Addressee, Origin, Effect, different kinds of Circumstantials);

(e) at least some basic features of the information structure of the sentences (Topic-Focus Articulation) are added.

The tentative and preliminary inventory of the tectogrammatical labels (based on the detailed studies of Sgall, 1967; Sgall et al., 1986; Petkevič, 1995) comprises 10 grammatemes (i.e. attributes taking as their value the values of morphological categories such as Tense, Modality, Aspect etc.) and 47 functors (i.e. syntactic relations).

3. The transition from the analytic to the tectogrammatical layer of tagging

To illustrate the two syntactic layers of tagging briefly characterized above, we present in Figs. 1 and 2 the analytic syntactic tree and the tectogrammatical syntactic tree of sentence (1) as taken from our corpus (sentence No. 22); for the sake of transparency, we omit the POS tags in both trees and add only the grammatemes of modalities and the functors in the TST.

(1) Všechny domy musí být stavěny s kryty , ale dosud
All houses must be built with shelters, but hitherto

pouze proti konvenčnímu útoku .
only against conventional attack.

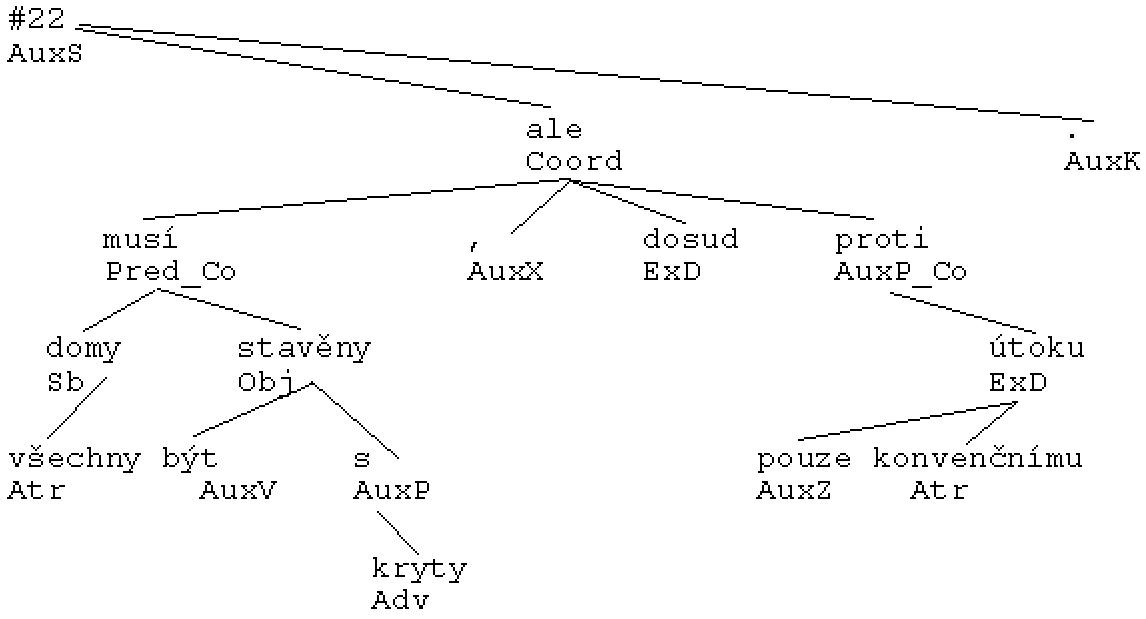

Fig.1 The analytic syntactic tree of the sentence (1)

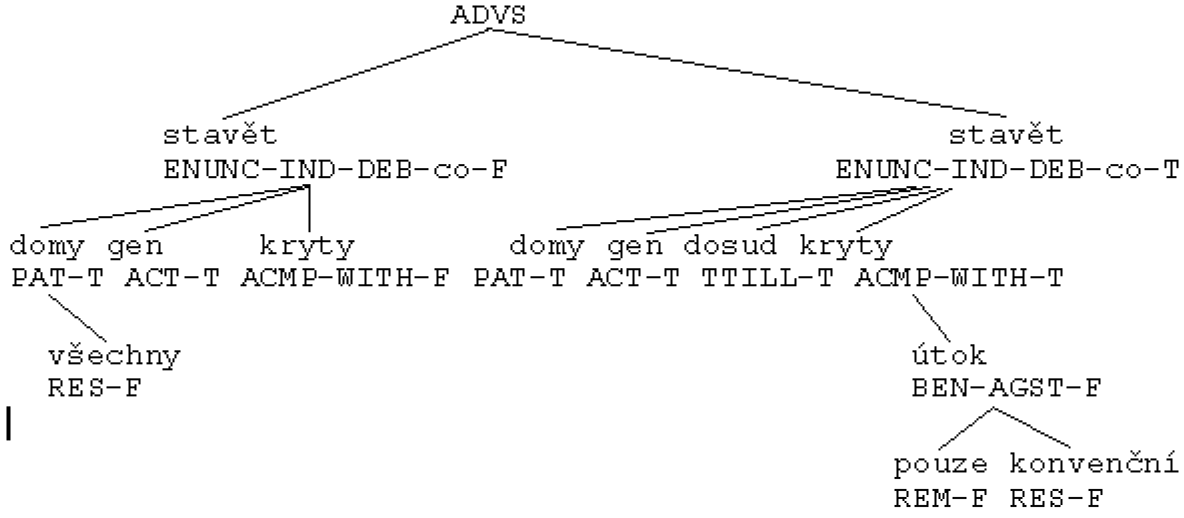

Fig.2 The tectogrammatical syntactic tree of the sentence (1)

The procedure that allows for a transduction of the ASTs to the TSTs is conceived of in two phases:

(A) An automatic tree pruning with the following steps:

(1) deletion of the AuxS and the auxiliary nodes for punctuation marks having only the function of delimiting the boundaries of clauses or of sentence parts (see the nodes AuxS for the root of the tree and AuxX for comma in Fig. 1);

(2) transformation of some of the 'auxiliary' nodes of ASTs into values of some parts of the complex labels in TSTs; this concerns the nodes having as their "lexical" values in the AST a punctuation mark (at the end of the sentence), a preposition, a hypotactic conjunction and an auxiliary AuxV. Thus, in our example in Fig.1, the AuxK is transformed into the value ENUNC of the attribute Sentmod attached to the main verb (the root of the TST); the AuxP's (*proti* 'against' and *s* 'with') are attached to the nouns they belong to (i.e. which 'depend' on them in the AST); the AuxV *být* 'be' is attached to the verb *stavěny* 'built'.

(B) A handcrafted procedure (again, with the help of specifically designed software tools) some of the steps of which can be tentatively characterized as follows:

(1) merging the periphrastic verb forms (placing the result in the position of their head nodes) and adding (to these head nodes) the values of grammatemes that capture the respective morphological meanings: thus the Pred_Co *musí* 'must' and the Obj *být stavěny* 'be built' are merged into a single node label stavět-ENUNC-IND-DEB;

(2) determination of the values of functors
and syntactic grammatemes:

(a) transformation of the prepositions (and hypotactic conjunctions) adjoined to the nouns (and verbs) they belong to during the automatic tree pruning: thus *proti* 'against' and s 'with' are transformed to the functors BEN-AGST with *útok* 'attack' and ACMP-WITH with *kryty* 'shelters', respectively;

(b) transduction of the subject, objects and adverbials (i.e. the nodes labeled by Sb, Obj and Adv in ASTs) into different kinds of functors: e.g. Adv *dosud* 'hitherto' is transduced as TTILL, Sb *domy* 'houses' as PAT (this change is triggered by the passive morphology of the verb complex);

(3) 'restoration' of nodes in cases of deletions on the surface level (the 'restored' node gets a special mark);

(a) nodes with pronominal lexical labels and corresponding functors are added in place of deleted subjects of finite forms of verbs (the *pro*-drop character of Czech);

(b) arguments of verbs having the character of 'general' participants (such as Actor, Patient, or, as the case may be, Addressee): a separate node depending on the respective verb is established with the lexical value 'gen' and with a functor corresponding to the 'missing' dependency relation; this is the case of gen-ACT in (1);

(c) obligatory arguments and adjuncts in the valency frames of the head words, but deleted in the surface shape of the sentence (and thus not occurring in the AST) are restored on the basis of context (their lexical values being first of all pronominal elements the determination of reference of which is a matter of an inference procedure leading from the tectogrammatical level to the layer of cognitive content); thus, e.g. with *come* either *here* or *there* is added; for the purpose of an envisaged procedure of reference assignment, a slot COREF is prepared with every TST node corresponding to a referential expression;

(d) addition of nodes with the so-called verbs of control: the 'restored' node receives a lexical label corresponding to the 'controller' and a functor corresponding to its dependency relation to the dependent verb (in the infinitive form in the AST); thus e.g. in the TST for *John tried to come home* the added node would have the lexical label *John* (i.e. the same label as the controller *John* with the verb of control *to try*) and would depend on *come* as its Actor; a similar treatment will be necessary for nouns derived from the control verbs and probably also for some other items;

(e) 'restoration' of nodes deleted in coordination:

(i) according to the lexical value of the coordinating conjunction, the analytic label for coordination Coord is replaced by the corresponding functor and an index co is added to all nodes coordinated by that relation: see ADVS replacing Coord merged with *ale* 'but' in Fig. 2;

(ii) in case of coordination of verbs, the lexical part, the grammatemes and, as the case may be, the functor of the label of the restored verb are copies of the respective values of the label of the verb with which the restored verb is coordinated except for the grammateme for contextual boundness, which is changed into T; cf. stavět-ENUNC-IND-DEB-co-T in Fig. 2; it still remains an open question which nodes have to be copied (and provided with the index for a contextually bound node, namely T) together with the governing verb; in our example (1), only the obligatory participants of the verb *stavět* 'build' (ACT and PAT) and the 'missing' governor kryty-ACMP-WITH of the node labeled ExD in AST (Fig. 1) are added;

(4) the topic-focus articulation is reflected in that the contextually bound ("given") lexical occurrences are marked by the index T (denoting nodes less dynamic than their governor) and the non-bound ones by the index F (denoting nodes more dynamic than their governor), and the surface word order is transduced to the underlying one (based on the hierarchy of communicative dynamism), cf. the indices T and F and the order of nodes in Fig 2.

The tectogrammatical annotation scheme and the transducing procedure sketched above represent a first attempt at a complex and systematic tagging on an underlying syntactic level; there are many questions open for further discussion and for a broader empirical study. However, the hitherto achieved analytic syntactic annotations of 30000 sentences of the corpus (together with the POS tags for all the occurrences in these sentences) offer us a rich material that already now can serve for monographic studies of most different syntactic phenomena of Czech without being bound to some specific syntactic theory, and, at the same time, as training data for a large-scale semi-automatic syntactic analysis of Czech.

Acknowledgements. The research reported on in this paper was carried out thanks to the financial support of the grants of the Czech Grant Agency GAČR 405/96/K214 and 405/96/0198 and the grant VS 96/151.

References

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

Hajič J. and Hladká B. (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., 111-118.

Petkevič V. (1995). A new formal specification of underlying representations. Theoretical Linguistics 21:7-61

Sgall P. (1967). Generativní popis jazyka a česká deklinace. [Generative Description of Czech and Czech Declension.] Prague: Academia.

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