Syntactic Tagging: Procedure for the Transition from the Analytic to the Tectogrammatical Tree Structures

Alena Böhmová, Jarmila Panevová, and Petr Sgall

Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
E-mail: {bohmova,panevova,sgall}@ufal.mff.cuni.cz

Abstract. The syntactic tagging of the Prague Dependency Treebank (PDT) is divide into two steps, the first resulting in analytic tree structures (ATS) and the second in tectogrammatical tree structures (TGTS). The present paper describes the transition procedures, automatic and manual, from ATS to TGTS and illustrates these procedures on two Czech sentences.

Syntactic tagging in The Prague Dependency Treebank Project is conceived of in two steps: (i) analytic tree structures (ATS), in which every word form and punctuation mark is explicitly represented as a node of a rooted tree, with no additional nodes added (except for the root of the tree of every sentence) and with edges of the tree corresponding to (surface) syntactic dependency relations, (ii) tectogrammatical tree structures (TGTS) corresponding to the underlying sentence representations; TGTSs have the shape of dependency trees 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 contrast to the ATSs, functional words (such as prepositions, auxiliaries, subordinating conjunctions etc.) as well as punctuation marks principally are not represented by nodes of their own; their functions are captured as parts of the labels (tags) of the nodes standing for autosemantic words. For technical reasons, the coordinating conjuntions are represented as specific nodes, which have the positions of the head nodes of coordinated constructions.

The transition from the ATSs to the TGTSs is conceived of as a transduction procedure (see [1]), consisting of two phases: (A) an automatic 'pre-processing' module, and (B) a manual tagging with the help of a 'user-friendly' software.

We want to illustrate here the automatic module, the input of which are the ATSs (with the accessibility of both the morphological and the analytical syntactic tags). The task of the module is then to process the ATSs in view of two aspects:

(a) to prune the tree structures, i.e. to devoid them of nodes that are counterparts to auxiliary forms in the surface structure of the sentence, without losing any important pieces of information these auxiliary forms carry;

(b) to translate (by means of linguistically substantiated transduction rules) the semantically relevant information given in the ATSs into the terms of the underlying structure.

The task under (a) concerns e.g. cancellation of the auxiliary node for the sentence and other "technical" nodes, transduction of the nodes standing for the final sentence boundary to the modality grammatemes with the governing verb, putting analytical forms together (and placing them in the position of the 'highest' of their parts), and adding the information they convey in the form of indices, grammatemes and other parts of the TGTS complex tags. The part (b) includes first of all the assignment of 'grammatemes' (i.e. for the values of morphological categories such as number, tense, modality etc.) in those cases in which they can be derived from ATS.

The procedure under (b) mainly concerns transduction of the analytic functions (such as Subject, Object, Adverbial, Attribute) into their tectogrammatical counterparts, i.e. Actor, Patient, Addressee, different kinds of Free Modification using the information on their form and their immediate context (e.g. the information encoded in the prepositions).

Example (1)

Iniciátoři	dosud	nesehnali	potřebných	třicet	podpisů	poslanců
The initiators	not yet	collected	needed	thirty	signatures	deputies

z	každé	sněmovny,	aby	schůze	obou
from	each	Chamber,	for	the sessions	two

komor	FS	mohly	být	předčasně	svolány
Chambers	of the Federal Parliament	to	be	specially	summoned.

'The initiators have not yet collected the needed thirty signatures of deputies from each Chamber, for the sessions of the two Chambers of the Federal Parliament to be specially summoned.'

AuxS is for an auxiliary node for sentence, Pred: predicate, Sb: subject, Atr: attribute (Czech: 'atribut'), Adv: adverbial, Obj: object, AuxP is for preposition, AuxC is for auxiliary node for conjunction, AuxX is for auxiliary node for comma, AuxV: verb, AuxK is for final interpunction. For more detailed explanation of analytic functions see [2].

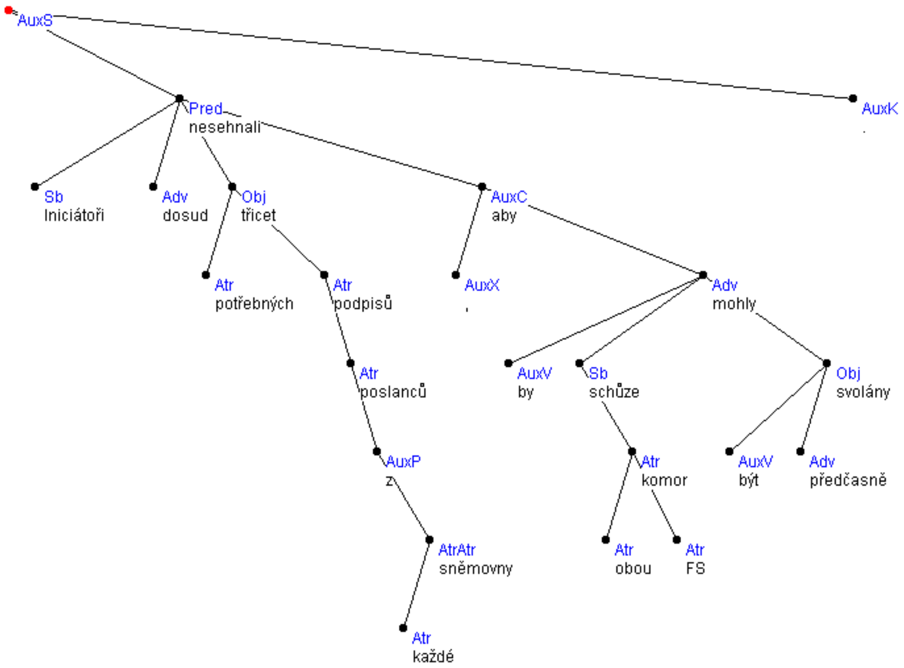

Fig. 1 ATS of sentence (1)

Example (1')

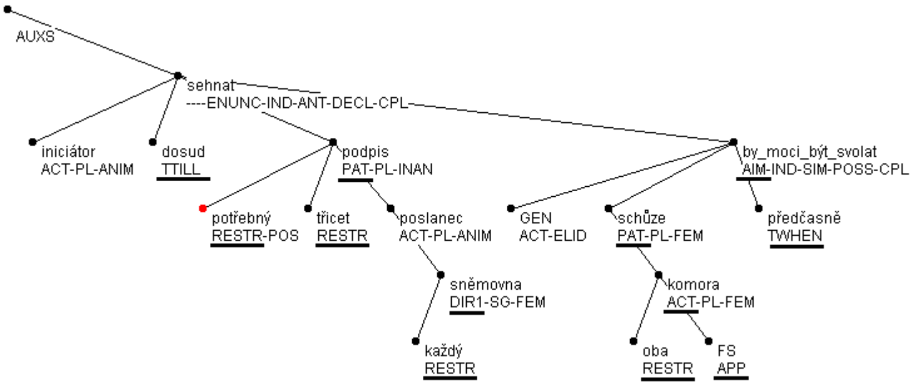

Fig. 2 TGTS of sentence (1)

ACT is for actor, PAT is for patient, TTILL: time adverbial 'till', TWHEN: time adverbial 'when', DIR1 is for direction, RESTR: a restrictive adjunct, APP: appurtenance, AIM: aim adverbial. ENUNC: ennunciation, IND is for the indicative verb mode, ANT and SIM are for verb tense (anterior and simultaneous, respectively), POSS and DECL are for deontic mode (possibilitive and declarative), CPL is for complex aspect.

Example (2)

Slovo "elita" se ovšem v Československu stále ještě chápe trochu
word "élite", however, in Czechoslovakia still is understood a little
pejorativně,
pejoratively

jako podezřelá kategorie samozvaně privilegovaných.
as a suspicious category of self-appointed privileged.

'The word "élite", however, in Czechoslovakia still is understood a little pejoratively, as a suspicious category of self-appointed privileged people.

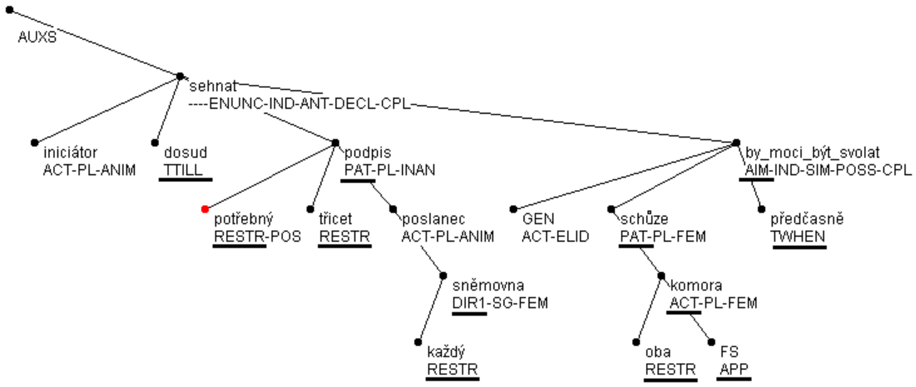

Fig. 3 ATS of sentence (2)

AuxG is for graphical symbols, AuxR is for a reflexive particle.

Example (2')

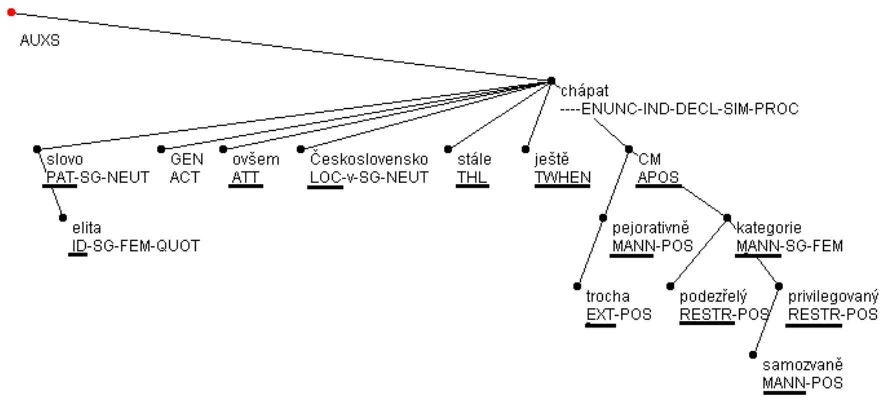

Fig. 4 TGTS of sentence (2)

LOC is for location, THL: time adverbial 'how long', APOS: apposition, MANN: manner.

By the ex. (1), (1') and (2), (2') we exemplify the subdivisions of tasks between the procedures (A)(a),(b) and (B); examples (1) and (2) are outputs from the ATS, their counterparts (1') and (2') correspond to the (simplified) output from TGTS. The tags left out for processing in the manual (B) procedure are underlined.

In ex. (1') we recieve as a result of the application of the automatic procedure (A) the following information:

(i) the orientation of the relation between head and its modifier in the construction *třicet podpisů* gets changed,

(ii) the nodes in ATS corresponding to the analytical forms of the verb and the modal verb with its infinitive complement (with the analytical function Obj) are combined in a single node (*by mohly být svolány*); the preposition (AuxP - 'z') is deleted as a node in the tree and it is stored in the attribute for the "future" value of the syntactic grammateme of the noun *sněmovna*),

(iii) all morphological grammatemes expressing the meaning of verbal categories (Verbmod, Deontmod, Tense and Aspect), gender and number with nouns and degrees of comparision with adjectives and adverbs are filled on the basis of their morphological tags (some asymmetries between forms and their respective functions will be solved later during the manual procedure),

(iv) the grammatemes of Sentmod with the root of the tree is specified automatically (this attribute is assigned to all heads of main clauses on the basis of the data present in the analytical tree),

(v) the analytical function Subject with the verb in active voice is converted into the tectogrammatical functor ACT (actor). The rest of functors will be determined on the basis of the "user-friendly" software manually by the procedure (B), which also includes the addition of a new node for a general actor (in the embedded clause with the verb in passive voice).

In ex. (2'), the same steps as in (1'), i.e. (ii), (iv) were applied; in this example an automatic procedure adds

(vi) the "special" grammateme QUOT for the quote word in quotation marks,

(vii) the analytical function AuxR denoting the reflexive passive is converted into a node with lexical value "general" (actor).

In the ex. (2) and (2') the representation of the apposition (which is analogical to the coordination) in ATS and in TGTS are illustrated.

Neither the automatic nor the manual part of the tagging can achieve a complete formulation of tectogrammatical representations. Several types of grammatical information will be specified only after further empirical investigations. Thus, e.g., the disambiguation of the functions of prepositions and conjunctios 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. Whenever possible, also statistical methods will be used.

References

1. Hajičová E.: Prague Dependency Treebank: From analytic to tectogrammatical annotations. Proceedings of the Conference TSD 98, Brno (1998)
2. Hajič J.: 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), Prague: Karolinum (1998) 106-132