# Tectogrammatics in Corpus Tagging

Eva Hajičová, Jarmila Panevová and Petr Sgall

A semi-automatic syntactic annotation of a part of the Czech National Corpus in the Prague Dependency Treebank is being carried out, based on the Praguian Functional Generative Description, the core of which is a dependency based account of underlying sentence structures.

The first phase of the tagging procedure (see Hajič 1998) consists of morphemic and "surface" annotations, during which the intermediate 'analytic level' (AL) is achieved; the analytic 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.

The aim of the present paper is to characterize the second phase, i.e. the main step of syntactic tagging, the procedure prepared for transducing from AL to syntax itself, i.e. to underlying, tectogrammatical representations (TRs, which are technically modified for the given purpose, as will be seen in the sequel).

## 1. Main differences:

Dependency trees are present both on AL 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 TRs 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), in ATSs syntactic relations are classified only on a "surface" layer, and without more subtle differences, such as those between types of objects or of adverbials.

One aspect of the TRs is their topic-focus articulation with a scale of communicative dynamism, represented as underlying word order; e.g. an adjective is prototypically more dynamic than its head, even if preceding it in 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).

## 2. Coordinating constructions:

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 TRs (a formal treatment of TRs 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 TRs 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.

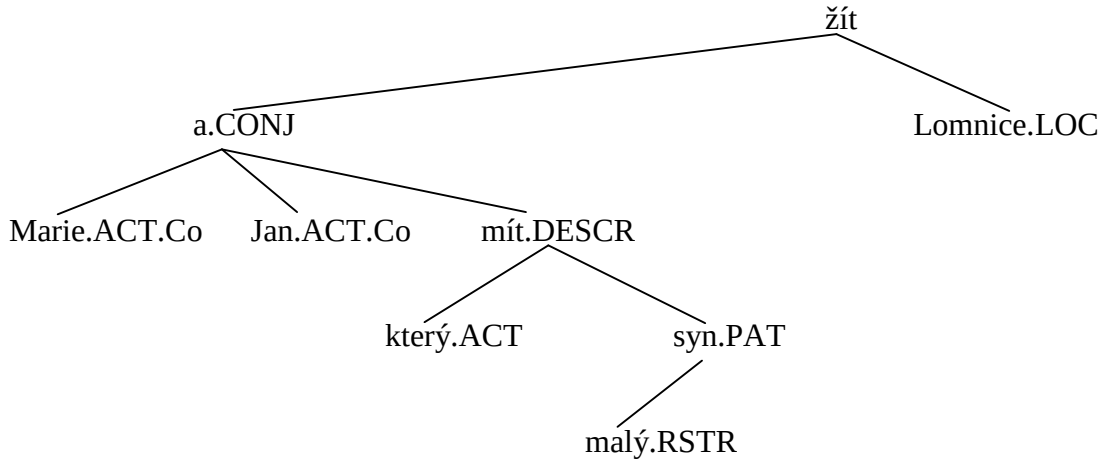


**Fig. 1.**

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

**3. Linearized underlying representations:**

TRs can be unambiguously linearized; e.g. the primary TR of (1) can be written as (1'), with each dependent item and each coordinated construction 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))))$_{Actor}$ žít ($_{Loc.in}$ Lomnice)

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

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) ($_{Restr}$ 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. 4 (ii)(c) below).

**4. The automatic part of the transduction of ATSs to TGTSs:**

A part of the transduction procedure (see Hajičová 1998) can be formulated as general steps, carried out automatically:

(i) In an automatic 'pre-processing' module, the input of which are the ATSs, 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. During this pre-processing, most of the analytical 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 will be 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 analytical tree.

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

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) After the 'manual' handling of TGTSs (see Section 5 below), another automatic module is being prepared, which will serve to add information that can be 'retrieved' automatically now 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 it does not depend on a noun (e.g. a superlative);

(b) the sentence modality value with 'content' clauses (indirect speech and similar) 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.:

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

- 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

*Jirka pozval Marii a Karel Milenu*

(lit. Jirka invited Mary and Karel Milena)

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;

- the secondary values of syntactic grammatemes (cf. Section 5 below) are added there, where a preposition allows for a reliable choice: ACCOMPANIMENT.WITHOUT (*bez*), BENEFACTIVE.NEG (*proti*), DIR3.IN (*do*), etc.;

- 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.

## 5. Intellectual part of underlying tagging

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

(i) the analytic functions (such as Subject, Object, Adverbial, Attribute), exressed by case endings, subordinating conjunctions and prepositions, are changed into corresponding functors; e.g. Dative with an LA object primarily yields ADDRESSEE, with an adverbial it yields BENEFACTIVE, Cz. *aby* or *na* yields Objective with LA 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 4 (ii)(c) as for the subsequent automatic step);

(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 pure denominations, such as nominal headings) or the attribute 'lemma' is left vacant for further treatment (e.g. in coordinations, see Section 4 (ii) (c) 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;

(iv) with possessive adjectives and pronouns dependent on nouns, the number and gender values of their basis 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 the help of a 'user-friendly' software that enables them to work with diagrammatic shapes of trees.

## 6. Concluding remarks:

Almost 100 000 sentences from the Czech National Corpus have obtained their 'analytical' annotations, and we expect to get about 10 000 sentences annotated by their TRs before the end of the year 2000.

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 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.

Whenever possible, also statistical methods will be used; specific combined procedures are being tested, based on statistical and structural approaches.

In this way a theoretically substantiated labelling of the TRs can be gained, 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 will be checked, but disambiguated representations of sentences will be achieved, which (although underspecified in the points in which the sentence structure is not fully specific - indistinctness, "systematic ambiguity", scopes of quantifiers) would constitute an appropriate input for a procedure of semantic - (pragmatic) interpretation.

## 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.

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: *Text, speech, dialogue. Proceedings of the Conference TSD* 98 (ed. by P. Sojka et al.), Brno: Masaryk University, 45-50.

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, 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, s. 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. *Wiener Slawistischer Almanach.* Sonderband 33. Ed. by T. Reuther, 273-282.

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