Jan Hajič, Eva Hajičová

Syntactic Tagging in the Prague Tree Bank

1. An overview of the project

The most recent natural language processing project in the Czech Republic started in 1996 as a joint research project of seven university departments and institutes (of Charles University in Prague and Masaryk University in Brno); the principal investigator is the Institute of Formal and Applied Linguistics. The project is scheduled for six years and its aim is to create a solid base for a versatile computerized processing of Czech language serving both as a multifacetted source of material for empirical and theoretical linguistic research as well as for multifarious applications in the domain of text processing and information retrieval. The project is conceived of as consisting of three branches, running in parallel but in close interrelations (for a more detailed description of the project, see Hajičová 1996): (i) the buildup of Czech National Corpus as the largest complex and representative Czech language data base, (ii) the research of present-day Czech based on contemporary methods and techniques of computational linguistics and lexicography, which includes the development of a tagging system for the corpus and formulation and implementation of tagging procedures, and (iii) continuation in the theoretical research in the domain of sentence and text structure, also with the perspective of possible applications in computerized system employing natural language processing.

2. The three layers of tagging of Czech

The development of tagging systems for Czech is not a trivial task, in view of the fact that most of the existing tagging systems have been developed for languages typologically different from Czech: the highly inflectional character of Czech and its communicatively rather than grammatically based word order present specific problems most of which have not yet been tackled in a sufficient and satisfactory way.

The tagging system for the Czech corpus is being conceived of as a system consisting of three interrelated layers (see Hajič, Hajičová and Rosen, 1996; let us note that in accordance with the grammatical tradition we distinguish ´morphological' and (two layers of) 'syntactic' tagging, not using the term ´morphosyntactic' because of the possibility to interpret it either in the ´coordinating' sense, that is synonymously with 'grammatical', or as referring to a domain capturing more of morphology than of syntax):

(a) morphological (augmented POS) tagging enriched by morphological data about the particular word forms; see Hladká (1996) on a stochastic POS tagger for Czech, working with more than 1100 different tags as combinations of the values of morphological categories of different POS's, and Hajič and Hladká (1997) on the experiments with a stochastic tagger based on the full-fledged automatic morphological analysis of Czech (Hajič 1994);

(b) the so-called analytic syntactic tagging the result of which are dependency trees the nodes of which are labelled by the word forms of the sentence (i.e. each word form constitutes a node of the tree) together with tags representing the syntactic relations between the governing and the dependent node (such as subject, object, adjunct, adverbial etc.);

(c) syntactico-semantic tagging resulting in dependency trees the nodes of which are labelled by the autosemantic word forms of the sentence with tags representing the syntactico-semantic (tectogrammatical, in the sense of the Functional Generative Description, see e.g. Sgall, Hajičová and Panevová 1986) relations such as Agent/Bearer, Patient, Addressee, Effect, Origin, and circumstantial modifications of different kinds. Inquiries will be carried out to make first steps also towards marking the intersentential relations including topic/focus articulation, anaphoric relations etc.

3. The analytic layer and the repertory of its relations

In the present contribution we concentrate on characterizing the second layer, namely the analytic syntactic tagging, with some hints how the third level, the tectogrammatical tagging should look like.

3.1 Tools for tagging

In the first phase of the research we have proposed the repertory of the tags for the analytic syntactic tagger and we have developed some tools contributing to a systematic and consistent tagging of the first sample; this sample has been tagged manually, only with the help of a software (MS Windows based) tool for tree manipulation. However, one of the intermediate aims is to develop a semi-automatic procedure based on the parser derived from the error-checking syntactic procedures of protype of the grammar checker for Czech (see e.g. Kuboň and Plátek, 1993). Possibilities will also be investigated to induce some idiosyncratic syntactic properties of particular words such as valency of verbs from the texts (annotated as well as raw) of the corpus (Böhmová, in prep.).

3.2 Characterization of the analytical level

The analytic level can be characterized as follows:

(a) each word and a punctuation mark is represented by a single node,

(b) no nodes are added except for the root of the tree (and some clearly specified exceptions, such as particles *-ň, -s* orthographically attached to the preceding word without a blank, as an ending, eg. *tys = ty* (you) + *s (-´jsi´,* are);

(c) the resulting representation is a dependency tree, in which edges (links) are explicitly labeled (by means of the tags attached to the dependent member of the dependency pair)

(d) non-projectivity is allowed.

The tags with each node of the resulting tree on the analytic syntactic level consist of three parts:

(i) the lexical part (word form),

(ii) the morphological tag (see the works quoted above for the shape of these tags),

(iii)the syntactic tag (name of the dependency link, see point (d) above).

Note: For the reasons of transparency, in our illustrations throughout this paper we do not include into the tags the morphological parts of them.

3.3 Illustration

The above-given characteristics can be illustrated by the tree (Fig. 1) for the Czech sentence (1) (taken from our corpus); we give a literal English translation under each example.

(1) Že bude zle, bylo jasné hned.

that it-will-go wrong was clear immediately

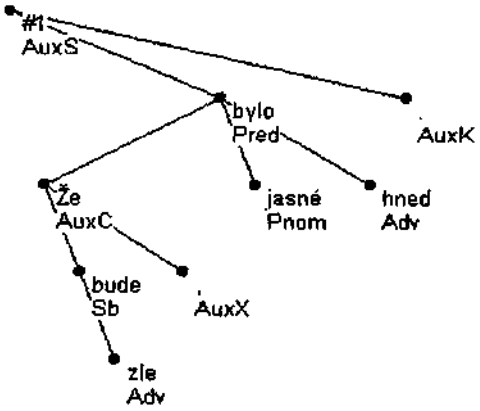

Fig. 1

3.4 The repertoire of analytic tags

At this stage of tagging, the following classification of analytic functions has been made (the abbreviations given below are also used in the examples of trees in Sect. 3.5):

Pred - the predicate of the main clause (also the copula), see Fig. 2

Sb - subject, see Fig. 3

Obj - object, without distinguishing a direct and indirect object and the so-called 'second' object *(zvolili ho předsedou* [lit.: they elected him chairman])

Atr - attribute, see Fig. 4

Adv - different kinds of adverbials (for some specific cases see below)

Atv - complement; the following convention is used: the Atv is represented as depending on the verb (with the label AtvV, see Fig. 5) if the Atv is related to a deleted member (eg. *přišel bos* '(he) came barefooted'); in other cases the Atv is represented as depending on another member of the sentence (on subject, see Fig. 6, on a direct object *with přivedli ho raněného* '(they) brought him wounded', on an indirect object with *dali mu to ještě spícímu* '(they) gave to-him it still sleeping', on an adverbial with *psal tím perem už rozbitým* '(he) wrote (with) that pen already broken') because its syntactic dependency on the verb of the given clause can be easily derived later, during the transition to the tectogrammatical level

Pnom - a predicate noun *(Jirka je šachista* 'George is chess-player') or adjective *(Jirka je nemocný* 'George is ill), see Fig. 7;

ExD - a node the governor of which is deleted (the governor is not present in the surface shape of the sentence, eg. a noun in a non-verbal sentence in a headline, or in a coordination with deletion, see Fig.9); in the tree, such a node "hangs" on the governing node of the deleted node, i.e. one step higher than it would be in the corresponding TR.

For cases, in which the syntactic relation is ambiguous to such an extent that it cannot be decided upon even if a broader context is taken into account, the annotators have at their disposal the following "double" labels (they can make a decision as for which relation they would prefer and denote this by the order of the component of the label; the preferred relation is denoted by the first part):

ObjAtr, AtrObj (the node is represented as depending on the noun, since the verb can be easily identified), eg. the notorious cases of the so-called PP-attachment such as *koupil boty pro kluka* '(he) bought shoes for boy'
AdvAtr, AtrAdv (with a similar choice possible as above)
AtrAtr (with a dependency relation to the leftmost of the competing nouns)

Functional (auxiliary) words and punctuation marks are labeled in the following way:

AuxV - the auxiliary verb *být* 'to-be' in the future tense forms, in passive and in conditional (see Fig. 8)

AuxT - the particle *se* with reflexiva tantum, eg. *smát se* 'to-smile'

AuxR - a reflexive particle that cannot be classified as an object (eg. with reflexive passives as *tento dům se stavěl pět let* 'this house was-built five years'

AuxC - a subordinate conjunction

AuxO - a 'demonstrative' word

AuxP - a preposition

AuxZ - a focalizer

AuxY - a secondary part of a complex connecting expression, eg. *buď* 'either' in cooccurrence with the primary *(a)nebo* 'or'

AuxX - a comma

AuxG - a dash or a bracket

AuxK - the punctuation mark at the end of the sentence (fullstop, question mark, exclamation mark, semicolon); this symbol is represented as depending on the root of the tree (i.e. in prototypical cases, as a right-hand sister of the main predicate of the sentence)

AuxS - the added root of the tree

Coordination (see Fig. 9) is incorporated into the dependency structure in a specific way: the node for the coordinating conjunction (representing the whole coordinated structure) has the syntactic function Coord and the individual coordinated members are labeled according to the syntactic position of the structure in the sentence with the suffix _Co attached to the label of that syntactic function (eg. Sb_Co, Pred_Co, etc.). Similar conventions concern apposition and parenthesis.

Phraseological complexes do not carry specific labels; their inner structure is analyzed in the same way as with regular syntagms.

Other conventions concern the way in which we represent the relations between nouns and prepositions, the auxiliaries within verbal complex forms, the attachment of dependent clauses etc. The main principle we have tried to observe was not to loose any piece of information that would be relevant for the (perspective) transition to the third layer of tagging and for the use of the annotated corpus in general.

3.5 Examples

To illustrate the classification of analytic functions and the conventions used for handcrafted tagging, we present here some examples of analytic trees. We give first the original Czech sentence with English glosses, and then the analytic tree with nodes labeled with lexical units and punctuation marks and analytic functions.

(2) Pekař peče housky.

Baker bakes bagels

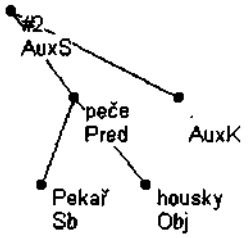

Fig. 2

(3) Kniha byla přeložena.

book was translated

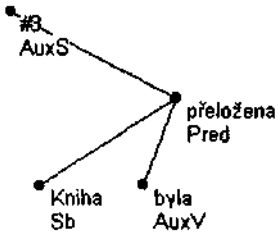

Fig. 3

(4) Dům, který je drahý, si nekoupíme.

House, which is expensive, ourselves we-will-not-buy

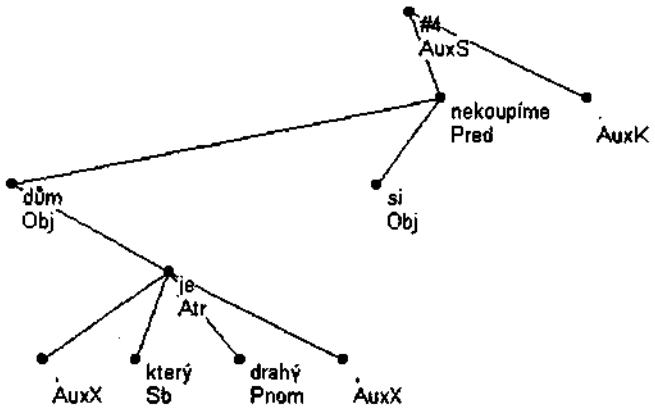

Fig. 4

(5) Má uvařeno.

He-has cooked

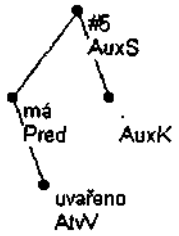

Fig. 5

(6) My jsme přišli tři.

we are came three

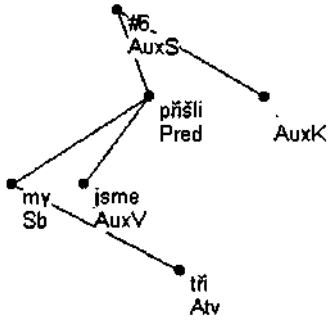

Fig. 6

(7) Bazén byl již napuštěn.

pool was already filled

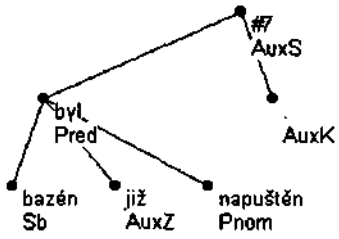

Fig. 7

(8) Směl by být zapsán.

he-allowed would-be to-be enrolled

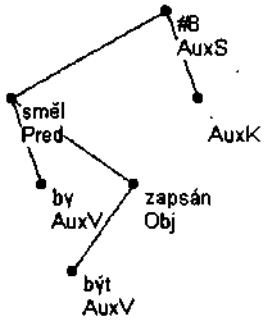

Fig. 8

(9) Petr pracuje dobře, ale Pavel špatně.

Peter works well, but Paul badly.

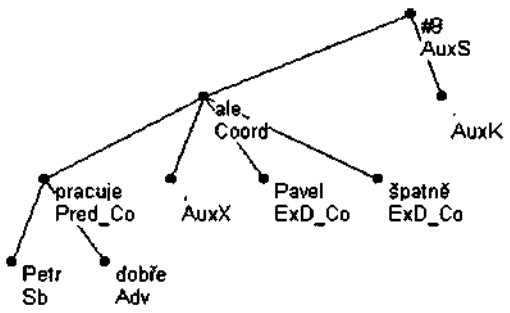

Fig. 9

4. Transition to the tectogrammatical level

4.1 Introductory remarks

As we have remarked above, the repertory of the analytic syntactic tags and the tagging procedure is conceived in view of the transition from this layer to the third layer of tagging resulting in tectogrammatical representations (TR's). The form of these representations is conceived of in accordance with the theoretical assumptions of FGD (for a detailed treatment, see Sgall et al. 1986); the formulation of the transition from the analytic to the tectogrammatical level can also use with a great advantage the material prepared by the members of the Institute for their previous project of syntactic analysis of Czech (Panevová et al. 1976).

4.2 Characterization of the third level of tagging

The structural annotations of the third layer of tagging are distinguished from the second, analytical level, by the following points:

(a) auxiliary nodes of the analytical level are deleted, only meaningful words are represented by a node of their own

(b) nodes can be added (to „restore“ surface deletions)

(c) analytic functions are replaced by tectogrammatical functions (such as Actor/Bearer, Patient, Addressee, Origin, Effectum, different kinds of Circumstantials)

(d) at least some basic features of the information structure of the sentences (Topic-Focus Articulation) will be added

(e) coreference relations will be specified.

4.3 Illustrations

In order to demonstrate some of the features of TR's in comparison with the analytic trees, we adduce in Figures 2' through 8' the tectogrammatical counterparts of the trees given above for sentences (2) through (8); since in the Functional Generative Grammar the relation of coordination is understood as a third dimension of the representation, different from the dependency relation, we give the TR of (9) in the form of a bracketted linear representation in (9'). The nodes in the TR's are labeled with (simplified) complex symbols consisting of the lexical meaning (represented in our examples by an (underlined) English word corresponding to the meaning of the Czech original), of a set of grammatical meanings (such as meanings of tenses and modalities with verbs, determination and number with nouns), and the name of the given tectogrammatical relation.

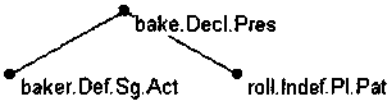

Fig. 2'

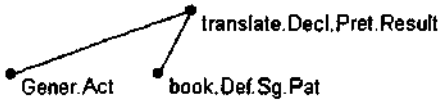

Fig. 3'

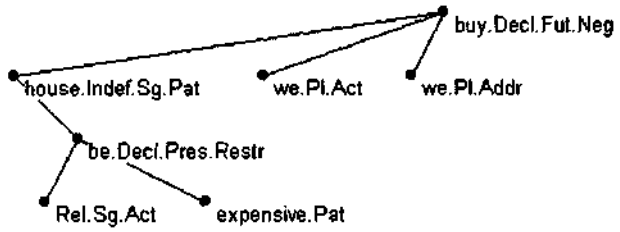

Fig. 4'

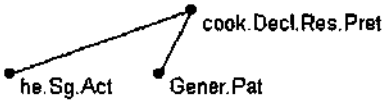

Fig. 5'

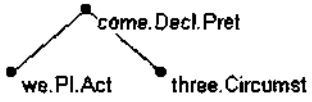

Fig. 6'

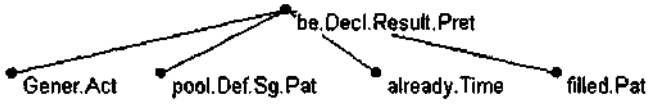

Fig. 7'

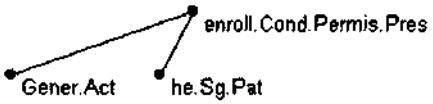

Fig. 8'

(9') ((Peter.Sg)$_{Act}$ work.Decl.Pres (Manner^{well}) (Paul.Sg)$_{Act}$ work.Decl.Pres (Manner^{badly}))$_{Advers}$

5. Conclusion

We are fully aware that the tagging procedure the output of which are annotations on the analytic level does not account fully for the proper syntactic structure of the sentence (see Sgall 1996). However, we are convinced that the results achieved provide a basic shape of the representation of a sentence structure in the shape of a dependency tree (even if with superfluous nodes for auxiliary words and punctuation marks and without nodes for deleted word forms) labeled with the basic types of syntactic functions. The analytic tree bank may thus serve for monographic studies of most different syntactic phenomena of Czech without being bound to some specific syntactic theory, and, first of all, as training data for a large-scale semiautomatic syntactic analysis of Czech as well as an input for the third (tectogrammatical) tagging procedure, the theoretical foundations and specification of the output of which are already prepared.

References:

Böhmová, Alena. In prep. *On data-oriented learning of valency frames.*

Hajič, Jan. 1994. *Unification Morphology Grammar.* PhD thesis, Faculty of Mathematics and Physics. Prague: Charles University.

Hajič, Jan, Eva Hajičová and Rosen Alexandr. 1996. „Formal representations of language structures". *TELRI Newsletter* 3,12-19.

Hajičová, Eva. 1996. *The past and present of computational linguistics at Charles University,* Tech. Report No.l, ÚFAL MFF UK, 2-10.

Hajič, Jan, Eva Hajičová, Jarmila Panevová and Sgall Petr. In prep. *Syntax v českém národním korpusu* [Syntax in Czech National Corpus], prepared for Slovo a slovesnost.

Hajič, Jan and Barbora Hladká. 1997. „Probabilistic and rule-based tagging of an inflective language - A comparison." *Proceedings of ANLP 1997,* Washington, D.C., March 1997.

Hladká, Barbora. 1994. *Počítačové vybavení pro zpracování velkých českých textových korpus.* [Software tools for processing of large Czech text corpora.] Diploma thesis. Prague: Charles University.

Kuboň, Vladislav and Martin Plátek. 1993. „Robust parsing and grammar checking of free word order languages". *Natural language parsing: Methods and formalisms.* Eds. K. Sikkel and A. Nijholt. Twente, 157-161.

Panevová, Jarmila et al. 1976, *Algoritmické zpracování syntaktické analýzy češtiny* [Algorithms for automatic syntactic analysis of Czech]. Research report. A brief survey to be published as Tech. report, ÚFAL MFF UK.

Sgall, Petr. 1996. „What linguists may expect and require from syntactic tagging". *TELRI Newsletter* 3, 9-11.

Sgall, Petr, Eva Hajičová and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects.* Prague: Academia and Dordrecht: Reidel.