Building an electronic language database nowadays:

The Prague Dependency Treebank

Jarmila Panevová (Charles University, Prague)

1. I am sure that the domain of the creation of the large language corpora, where linguistic annotations are assigned to the input data, still belongs to the interests of the Festschrift's owner. I met Prof. Ferenc Papp for the first time in 1964 in Prague at the Colloquium on Mathematical Linguistics and we immediately have found a common basis of interest: how to store linguistic data about raw texts and how to deal with them using the contemporary technical equipment. At that time the most advanced technique for natural language processing available for linguists was represented by punch-card machines. In October, 1964, Ferenc Papp organized a small conference in Budapest, where the participants (I had the honour to be one of them) exchanged their opinions on what type of data can be stored on punch-cards, how they can be classified and evaluated from the point of view of their linguistic nature as well as from the point of view of the efficiency and role of the punch-card machine set. Reminding this in 2000, in the year of F. Papp's anniversary, the whole issue sounds as a kind of a crazy nostalgia. The punch-card machines disappeared very soon and the punch-cards, storing the rich inventory of linguistic data, become unreadable. However, the idea remains alive: In the 1980s corpus linguistics was born and in the 1990s (syntactically) annotated corpora started to develop.

I believe that a comparison of the possibilities offered by contemporary information technologies with the situation of 35 years ago as well as a description of some recent results achieved in this domain in Prague will bring for the Festschriftee some enjoyment.

2. The capacity of present-day computers, the availability of PC's, their speed as well as the software tools for 'user-friendly' processing of textual data create quite new conditions for empirical studies in linguistics. On the other side, it is necessary to choose the linguistic framework for the procedure of the annotation. In the case of the Prague Dependency Treebank (PDT in sequel), which will be characterized in the following sections, the dependency approach to syntax, close to the structuralist approach, as well as to the Functional Generative Description (systematicaly described the firts time in Sgall, 1967), is used. PDT is a part of Czech National Corpus (about Czech National Corpus see Čermák (1995)) containing annotations of particular input sentences on three levels: The morphological tags are a necessary precondition for the syntactic annotations, for which two steps are used. At the 'analytical' level (AL in sequence) the tree structure, corresponding approximately to the surface syntax, ¨

represents an auxiliary step for the tectogrammatics, i.e. for the disambiguated representation of the syntactic structure (its underlying representation). The tasks of manual, automatic and semiautomatic procedures will be described below (in Sect. 5).

The Prague Dependency Treebank, designed and elaborated in the Institute of Formal and Applied Linguistics at Charles University in Prague since 1996, represents an electronic corpus annotated on the three mentioned levels. Its creation was inspired by the Penn Treebank (see e.g. Marcus et al (1993)).

The part of Czech National Corpus chosen for PDT consists of 40% of general newspaper article, 20% of economic news and analysis, 20% of popular science magazine and 20% of information technology texts.

3. I leave aside the morphological tagging, which is described in Hajič, Hladká (1997) and Hajič (1998). I restrict myself here only to the fact that we have based the syntactic tagging on the knowledge of the disambiguated morphological tag for every word-form; it does not matter here if it was obtained by a stochastic procedure on the basis of manually tagged corpus, serving as the training data for machine learning, or by a rule based symbolic automatic procedure.

In the first stage (where the training data for stochastic machine learning procedure were obtained), word forms were submitted to a 'user-friendly' software tool helping the human annotators to build the tree structure on the AL: the governor were found for every word-form and its analytical function (AF in the sequel) was intellectually assigned. The annotator provided his/her choice in an interactive way from a 'menu' containing the whole set of AF proposed by the software tool.

The shape of the tree on AL is determined by the output from the morphological analysis. Every string between two blanks must have a node of its own. It means that some redundant nodes (from the point of view of the syntactic structure) are present here (e.g. punctuation and other graphic marks, synsemantic 'function' words etc.). On the other side, the number of the nodes is predetermined by the input of the analytical procedure (it is equal to the number of word-forms in the sense defined above, with one additional node - as a root of the whole tree - as an identification mark for the sentence in the file). In consequence, no node absent on the surface (e.g. because of deletions) can be added on AL.

The tag set (AF) on the AL comprises about 60 basic tags multiplied by three (because almost all AFs, corresponding to the dependency relations, can stand in coordination, apposition or within parenthesis; the full set of AFs is given in Hajič (1998)). AF Obj(ect) and Adv(erbial) are not further classified on the AL; they are converted into valency slots (inner participants and free modifiers) only on the

tectogrammatical level, where tectogrammatical tree structure (TGTS in sequel) is created.

At the end of 1999 the number of sentences annotated on the AL reached 100 thousand. At that time point the stage of analytical annotation was stopped. The annotated files are compared and unified and at the same time the stage of tectogrammatical annotations was started. Since the tectogrammatical analysis is much more pretentious than the AL annotating, two approaches are applied: Smaller samples of files are analyzed with all details (concerning coreference relations, semantics of morphological categories, subtle analysis of all syntactic relations etc.) by two very well-trained annotators, this sample being called 'model corpus'. No rapid progress is expected here, while just a bit simplified tectogrammatical annotating called the 'large corpus' is provided by a team of 5 linguistically educated members. The large corpus is supposed to include about a couple of thousand sentences at the end of 2000.

4. Both methods of TGTS annotations represent a great challenge for the staff-members formulating a manual with instructions of the annotating procedure, for the staff of annotators as well as for the future users of the tagged corpus. The former are facing many not yet discussed and not yet anywhere described syntactic phenomena. It can be said that every sixth sentence forces the annotator to some kind of deliberation and to nontrivial decision making.

Some questions concerning the relations between the formal theoretical shape of FGD and the limitation given by the possibilities of the annotating procedure had to be solved in general:

(a) The coordination and apposition relations are described in the theoretical description as a 'third dimension' of the tree structure (see Petkevič, 1995), while in the TGTS we work with a two-dimensional tree structure. Certain 'artificial' nodes for the coordinated group (group in apposition) are introduced; these nodes represent type of connection (conjunction, disjunction, adversative, apposition etc.), see below in Fig. 1 (for apposition) and Figures 2 and 4 (for coordination).

(b) Some issues connected with 'textual' coreference (see Hajičová, Panevová, Sgall (1985-87)) are added (this concerns esp. 'model corpus'), though some of these attributes do not belong to the proper underlying structure. Intertextual relationships between the given sentence and the previous text are reflected by the attribute value PREV. All anaphoric pronouns that have been deleted are restored in TGTS (see e.g. Fig. 2). Though some of the coreferentiality charactestics are not directly relevant for the underlying (syntactic) structure, we do not want to loose such pieces of information, which are transparent for the human annotators and will be certainly

useful for the future users interested in discourse analysis. On the other side, the information about topic-comment articulation is assigned to any sentence (every node bears the information, whether it is a contrastive or non-contrastive part of the topic, or of the focus).

5. After the first stage of the manual building of the training data on AL the second stage of annotating represents new approach. Automatic preprocessing of the input is used for the analytical tagging (trained on the manually tagged Czech data, see the Section 3 above); it uses the parsing procedure proposed by Collins (1996). Though the original version of the parser was developed for English and for phrase-structure based syntactic tagging, it gives (after the absolving machine-learning procedure using Czech data) surprisingly successful results for Czech (80% edges were established correctly, while for English 91% accuracy was reached). Automatically preprocessed structure submitted at the second stage to the human annotators is on one hand a great help for the quick and smooth manual phase of annotating: the annotators check the tree structure and assign the AFs. On the other hand the annotators are often influenced by the 'preprocessed' structure including some errors and they pass them on the output.

The first stage of TGTS annotations is also an automatic procedure; it consists first of all of automatic 'pruning' of the tree structure (the auxiliary symbols as e.g. synsemantic words are merged with their autosemantic ones; this concerns e.g. auxiliary verbs in analytical verb forms, prepositions, subordinated conjunctions, reflexive particles etc.). The direction of dependency is automatically changed in numerative constructions (e. g. in the Cz. construction *pět chlapců* [five boys]) the counted noun switches in TGTS automatically into the head position and the numeral is its 'restrictive' modifier, while on AL the numeral was evaluated as a head modified by the genitive of the counted nouns. A modal verb is merged with its autosemantic verb depending on it in AL (evaluated there as its Object); according to the lexical value of the modal verb the morphological grammateme (called Verbmod) is filled in TGTS automatically by the values such as necessity, possibility, permission etc. Attribute Sentmod (sentence modality) is automatically filled in simple cases according to the final punctuation mark or according the presence of particles expressing a wish (Cz. *kéž, nechť, ať* [let]). For coordination of clauses with different modalities as e.g. Cz. *Já půjdu ven a ty tady ukliď!* [I shall go for a walk and you clean here!], more complicated rules are formulated. The value of the attribute Sentmod with complex clauses embedding the content clause (as e.g. in Cz. *Řekl mu, ať přijde; Řekni mu, zda přijdeš; Řekni, kdy přijdeš* - [He told him, let him come; Tell him, whether you will come; Tell, when you will come] will be automatically processed at the stage following the manual stage of annotating. The values will be

based on the functions of the content clauses and the coocurrence of the verb in matrix sentence and the subordinated conjunction or wh-element in embedded sentence. The analytical function is converted into its tectogrammatical counterpart ('functor') automatically only in very simple cases (e.g. AF 'subject' in active sentence is replaced by the functor Actor in TGTS, cf. the valency theory in Panevová (1974-75, 1994)). If on the AL an active verb has two objects depending on it, one in Accusative case and the other in Dative, the former is automatically transduced into the functor Objective (Patient), and the latter into Addressee. Analytical Object in Instrumental case with passive verb is converted into the functor Actor. The rest of functors must be treated manually.

An attribute QUOT is automatically established with the head verbs of sentences, where the direct speech is included. This attribute is assigned automatically also to the words surrounded by question marks. Some of the graphical symbols are automatically deleted.

I want to stress here that we try to find effective and reasonable subdivision of tasks between automatic preprocessing and manual tagging procedure. In any case a great deal of linguistic analysis remains open for human annotators. The manual of instructions for annotating on AL and the other manual for annotating on TGTS were published as technical reports (Bémová et al, 1997, Hajičová et al, 1999). We pass now over to a stage, during which three of the annotators analyze the same sets of sentences; their results are automatically compared and the differences among them are listed. These differences are the main object of the team discussions.

As to the technical and software side, macro language was developed together with a tree structure graph editor. These are powerful tools for handling tree structure which allow e.g. in an interactive way to cut or paste a subtree, to find the parent node or a left or right sister node, to assign an attribute an appropriate value etc.

The building of PDT sketched here only briefly; the range of problems cannot be exhausted in one article. Let me only state that during the syntactic analysis needed for the AL and TGTS annotation we encounter a mess of phenomena not yet explicitly described. The requirement on a consistent analysis of raw texts open new horizons for future syntactic handbooks which would not be formulated for the annotators, but for students, teachers etc.

The PDT - a corpus tagged on three levels - is planned to be used for a stochastic parsing of unrestricted texts. However, linguists dealing with Czech language will appreciate its contribution within their respective particular monographs and studies about Czech; they must only know how PDT can be efficiently and creatively explored.

6. In the last section I want to adduce four sentences with a part of their TGTS annotation in order to illustrate how the output (a source for further linguistic studies) looks like in its core part (tree structure with nodes labeled by their respective lemmas and functors is shown in Figures 1 to 4; the value ELID for the nodes absent on surface is present here); every node has of course its own set of 23 attributes, only three of which are printed in the illustrative sentences here (the list of all attributes and their values is given in Appendix 1). In the Appendix 2 some functor's labels are interpreted.

Fig. 1: #2 Jon Woronoff, autor knihy Mýtus japonského managementu, se tento mýtus snaží postavit do reálného světa

Lit.: [#2 *Jon Woronoff author (of) book Myth Japanese management this Myth tries-Refl build up into real world*]

Eng.: #2 Jon Woronoff, the author of the book Myth of Japanese management, tries to build up this myth into a real world.

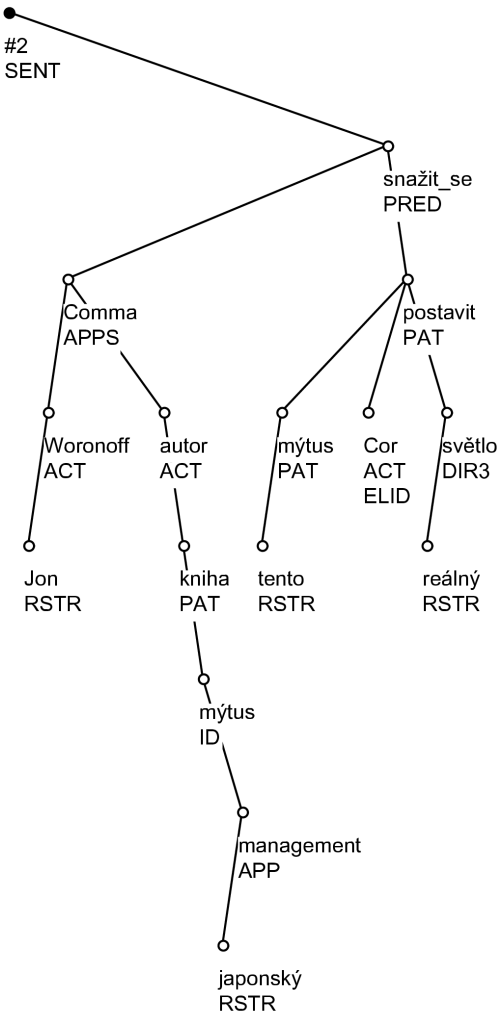

Fig. 1

Fig. 2: #3 Rozebírá přednosti i slabiny japonského managementu

Lit.: [*(He) analyzes advantages and weak points (of) Japanese management*]

Eng. #3 He analyzes advantages as well as weak points of the Japanese management.

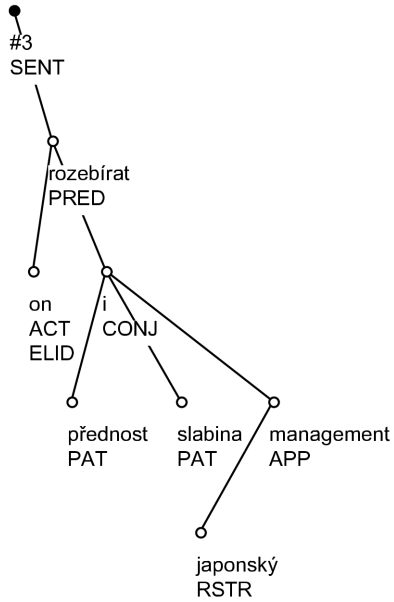

Fig. 2

Fig. 3: #36 Ještě před vlastním projednáním celé záležitosti v tripartitě jsme se zeptali prvního náměstka ministra financí ČR Jana Klaka, jaký je jeho názor na předložený návrh.

Lit.: [#36 *Still before own negotiation of whole affair in Tripartity we asked the first vice-minister (of) finance (of) CR Jan Klak, what is his opinion about submitted proposal*]

Eng.: #36 Still before the proper negotiations of the whole affair in Tripartity, we asked the Vice-Minister of Finance of the Czech Republic (CR) Jan Klak what is his opinion about the submitted proposal.

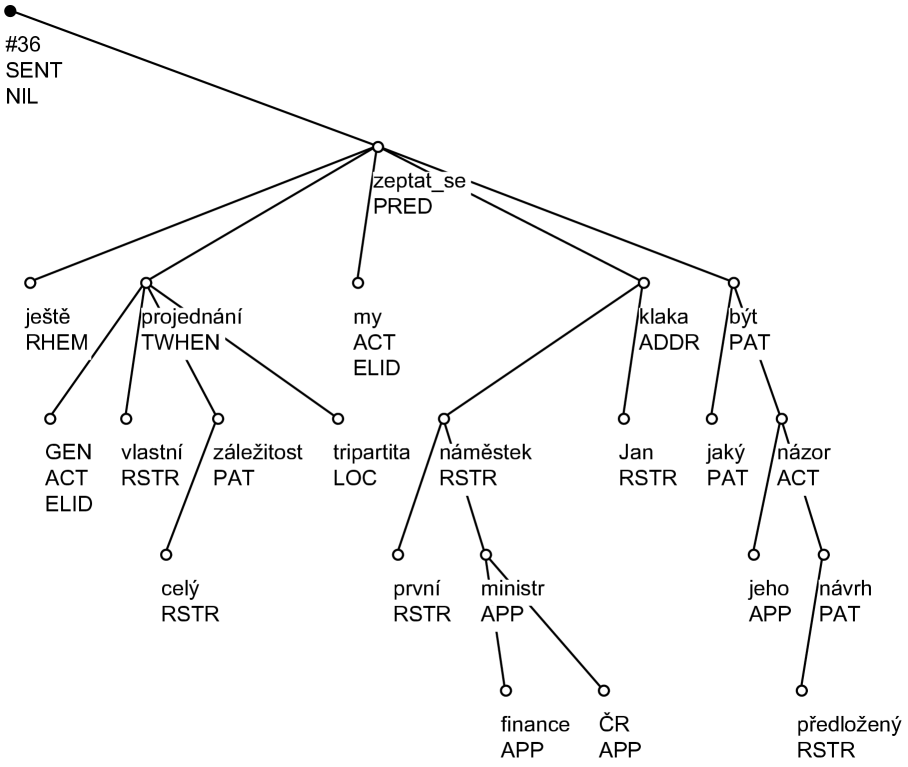

Fig. 3

Fig. 4: #43 Záměr snižovat postupně daně hlásají jak vláda, tak i podnikatelé

Lit.: [#43 *Intention to decrease gradually taxes announce both government and entrepreneurs*]

Eng.: #43 Government as well as the entrepreneurs announce their intention to decrease gradually the taxes.

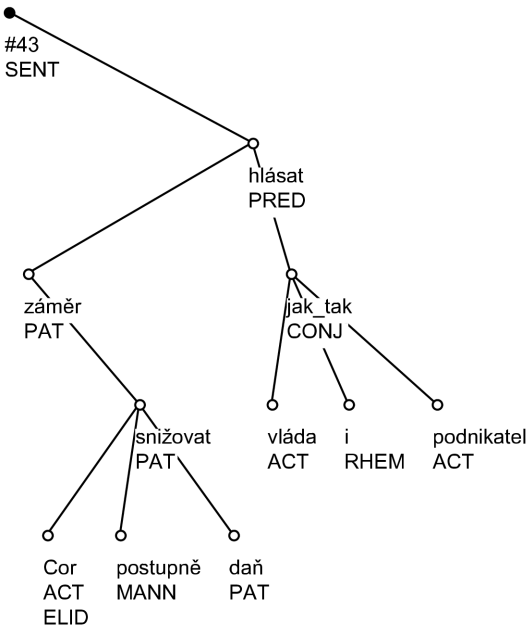

Fig. 4

In the TGTS of all examples the values of other attributes are hidden in the outer hape. The user of the graph editor can open them at any moment of his/her work.

In Fig. 1 the convention about the representation of apposition is applied (Comma as a value of lemma is used here as a connector unifying two parts of apposition). The grammatical coreference is present here: a verb of control *(snažit se* [to try]) has the Actor (expressed by both parts of apposition) as its controller and the Actor (Subject) of the embedded infinitive as its controllee (designed here as a node with lemma Cor), absent (ELID) in the surface shape.

Fig. 2 represents a sentence immediately following in the text the sentence from Fig. 1. Though this sentence is very simple in its structure, it illustrates here kind of textual coreferentiality: the node for Actor with the lexical value of the anaphoric pronoun *on* [he] is introduced in the TGTS as ELID (elision). The coordination structure of two Patients connected by the conjunction *i* [as well as] is illustrated here; this way of reflecting coordination allows us to show that the construction labeled by APP(urtanance) - *japonského managementu* [of Japanese management] - modifies both nouns in coordination. An other type of construction, e.g. *předseda a ministr vnitra odešli* [a chair and the minister of inner affairs left], will receive a different structure, where the noun in Genitive *vnitra* [of inner affairs] modifies only its preceding noun (*ministr* [minister]).

Fig. 3 represents a richer structure as to the number of the nodes. Here the omitted subject (Actor) *my* [we] is added, the adverb *ještě* [still] is determined as a Rhematizer (having in its scope the rest of sentence, i.e. all nodes on its right-hand side). The deverbal noun *projednání* [negotiation] has its own valency structure, i.e. Actor, which is generalized here (having lemma GEN), and Patient; also the noun *názor* [opinion], though it is not transparently postverbal, has its valency (argument) structure (its modifier *na návrh* [on proposal] is analyzed as Patient). With nonverbal nouns arguments which are not present on surface are not (according to the agreed convention) filled in. There is an error present in the position of the lemma: the morphological tagger finding 'unknown' proper name *(Klak)* analyzed it as a form of the common noun *klaka* [company].

In Fig. 4 coordination is again present (its parts are connected by the double part conjunction *jak - tak* [Lit.: as - thus, *E.: and*]). The controlled infinitive construction modifies the noun as its Patient, the control (called often as arbitrary) is reflected here by adding the node with lemma Cor and functor Actor (about the description of control phenomena in Czech see Panevová (1996)).

APPENDIX 1

List of attributes in TGTSs and their values

The values used here are mostly self-explanatory; NA means that this attribute is not applicable in the given case; ??? is a value, which is not yet resolved.

1.	trlemma	
2.	gender	ANIM INAN FEM NEUT NA ???
3.	number	SG PL NA ???
4.	degcmp	POS COMP SUP NA ???
5.	tense	SIM ANT POST NA ???
6.	aspect	PROC CPL RES NA ???
7.	iterativeness	IT1 IT0 NA ???
8.	verbmod	IND IMP CDN NA ???
9.	deontmod	DECL DEB HRT VOL POSS PERM FAC NA ???
10.	sentmod	ENUNC EXCL DESID IMPER INTER NA ???
11.	tfa	T F C NA ???
12.	func	ACT PAT ADDR EFF ORIG ACMP ADVS AIM APP APPS ATT BEN CAUS CNCS COMPL COND CONJ CPR CRIT CSQ CTERF DENOM DES DIFF DIR1 DIR2 DIR3 DISJ DPHR ETHD EXT EV GRAD HER INTF INTT ID LOC MANN MAT MEANS MOD NORM PAR PARTL PREC PRED REAS REG RESL RESTR RHEM RSTR SUBS TFHL TFRWH THL THO TOWH TPAR TSIN TTILL TWHEN VOC VOCAT NA SENT ???
13.	gram	0 GNEG DISTR APPX GPART GMULT VCT PNREL DFR ON BEF AFT JAFT INTV WITH WOUT FOR AGST NIL ???
14.	reltype	CO PA NA ???
15.	fw	
16.	phraseme	
17.	del	ELID ELEX EXPN NIL ???
18.	quoted	QUOT NIL ???
19.	dsp	DSP DSPP NIL ???
20.	coref	

21.	cornum	
22.	corstn	PREV NIL ???
23.	antec	as func

APPENDIX 2

The interpretation of the values used in Figures 1 t,o 4

ACT Actor/Bearer

ADDR Addressee

APP Appurtenance

APPS Apposition

CONJ Conjunction (type of coordination)

DIR3 Direction - Where to

ELID Elision (the node is absent on surface)

ID Identity

LOC Locative

MANN Manner

PAT Patiens

PRED Predicate

RHEM Rhematizer

RSTR Restrictive (modification)

SENT Sentence (the artificial root of every tree with a sequential number of the tree in the file)

TWHEN Time - When

REFERENCES

Bémová, A. et al. (1997), Anotace na analytické rovinì: návod pro anotátory [Annotations on the analytical level: instructions for the annotators], *Technical Report UFAL TR -1997-03. Charles University.*

Collins, M. (1996): A New Statistical Parser Based on Bigram Lexical Dependencies. *In*: Proceedings of the 34th Annual Meeting of the ACL ‘96, Santa Cruz, CA, USA, June 24-27, 184-191.

Čermák, F. (1995): Jazykový korpus: Prostøedek a zdroj poznání [Language Corpus: An Instrument and a source of understanding], *Slovo a slovesnost 56*, 119-140.

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. E. Hajičová). Prague: Karolinum, 106-132.

Hajič, J. - 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.

Hajičová, E. - Panevová, J. - Sgall, P. (1985-87): Corefernce in the Grammar and in the Text. Part I: *Prague Bulletin of Mathematical Linguistics 44 (PBML*), 3-22; Part II: *PBML 46*, 1986, 1-11; Part III: *PBML 48*, 1987, 3-12.

Hajičová, E. - Panevová, J. - Sgall, P. (1999): Manuál pro tektogramatické značkování [Manual for the Tectogrammatical Tagging]. *Technical Report UFAL TR-1999-07. Charles University*.

Marcus, M. P. - Santorini, B. - Marcinkiewicz, M. A. (1993): Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, 19 (2)*, 313-330.

Panevová, J. (1974-75): On verbal frames in Functional Generative Description. Part I: *Prague Bulletin of Mathematical Linguistics 22 (PBML*), 3-40; Part II: *PBML 23*, 1975, 17-52.

Panevová, J. (1994): Valency Frames and the Meaning of the Sentence. *In*: Prague School of Structural and Functional Linguistics (ed. Ph. L. Luelsdorff), Linguistic and Literary Studies in Eastern Europe 41, Amsterdam-Philadelphia: John Benjamins, 223-243.

Panevová, J. (1996): More Remarks on Control. *In:* Prague Linguistic Circle Papers, Vol. 2 (eds. E. Hajičová, O. Leška, P. Sgall, Z. Skoumalová), Amsterdam-Philadelphia: John Benjamins, 101-120.

Petkevič, Vl. (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 Language and the Czech Declinsion], Praha: Academia.