Chapter 1

Automatic procedures in Tectogrammatical Tagging

1.1 Abstract

Many research teams are currently aiming at building syntactically annotated corpora, called "treebanks" in the current terminology. The task of building a treebank involves the formulation of an appropriate annotation scheme, development of the annotation tools, as well as incorporation of methods of linguistic analysis which may serve both for POS tagging and syntactic parsing.

The build-up of a treebank is a time-consuming and labour-intensive task, usually performed by human annotators with support of various software tools and utilities. The process consists of automatic procedures and manual corrections of their results, in the form of post-editing or interactive cooperation of the humans with the program.

The purpose of this study is to discuss and illustrate several procedures involved in the parts of the Prague Dependency Treebank (PDT) which are devoted to underlying syntactic analysis. The author has formulated an automatic part of the transition from analytic' (surface) syntactic trees to tectogrammatical (underlying) ones (based on the descriptive framework of the Praguian Functional Generative Description), and has constructed tens of software tools, macros and other utilities which make it possible for the corresponding steps of manual parts of the tagging procedure to proceed more quickly and efficiently, without unnecessary effort on the side of the annotators. The task contains many specific points due to the fact that Czech is a highly inflectional language, the basic typological properties are different from what is known from English, French, Italian and other languages nowadays having a large background for their analysis.

1.2 Prague Dependency Treebank

As it is obvious that the annotated data provide a large scale of possibilities in computer processing of natural language, it has been decided at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and physics, Charles University Prague, to develop an annotated corpus of Czech language. This system, called Prague Dependency Treebank (PDT), uses a rich annotation scheme. The long term research of the formal language description led by Petr Sgall since the 1960's at the university created the basis for the

annotation scheme and formalisms. The size of the corpus and data formats used were inspired by the Penn Tree Bank [8]. The Prague Dependency Treebank is being built as an annotated version of a part of the Czech National Corpus [9].

A semi-automatic syntactic annotation of a part of the Czech National Corpus in the PDT has among its aims the possibility to check the theoretical approach chosen (Functional Generative Description, see [2]). We also expect that with the growth of the amount of annotated data we will be able to improve the automatic annotation process on the tectogrammatical level. In the process of creating the guidelines for annotation of the corpus and formulating the procedure itself, we are gaining information about the grammar and morphology characteristics that could be processed automatically after they are extensively explored by linguists. This will lead to a new level of the empirical studies. Another aim is to offer the language material for different further studies, in the fields such as grammar, literary studies, information retrieval or stylistics.

A language with rich inflection and with a high degree of "free" word order, such as Czech, cannot be handled by primarily using cues based on cooccurrence with neighboring items, but requires specific procedures for the disambiguation of morphemic units (prepositional and simple case forms, agreement forms, etc.), which hardly could be fully automated. The work on such procedures has led to our conviction that many insights of classical structural linguistics may still be highly useful, although they have not been duly reflected in theories using an approach based on constituency (that originated with Bloomfieldian descriptivism). Considering syntactic dependency (which is being developed on the basis of the work of L. Tesnière) to constitute the primary layer of sentence patterns, we work with a structure that corresponds to extremely flat constituency patterns, and we use no nonterminals in the dependency trees. Instead of notions such as NP or AP, the dependency approach shows just items dependent (immediately or not) on a noun or an adjective, respectively.

1.3 The annotation process

The annotations of texts are done in 4 separate steps: morphemic analysis [4], morphological tagging [5], syntactic tagging on the intermediate analytical level and tectogrammatical level tagging, which leads to full underlying representations of the corpus sentences.

The analytical level was developed as the first stage of annotations, to bridge the gap between the linear sentence representation and the underlying dependency tree. In the analytical tree (ATS) every word of the sentence is represented by a single node, including the punctuation marks. The rules for building the dependency tree based on this condition have been formulated [6] and approximately 100000 sentences have been annotated at the analytical level.

The process is illustrated by figure 1.1.

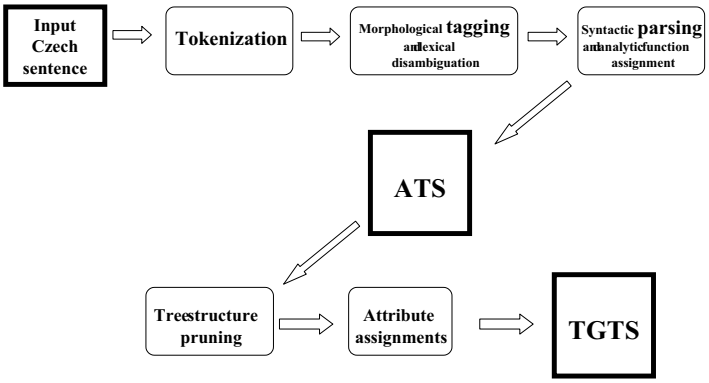

Figure 1.1: Text annotation process in PDT

1.4 The task of our research

Our task was to automatize the transduction from the ATSs towards the tectogrammatical tree structures (TGTS, see Section 1.5) to the possible extent and to explore the possibilities of further automatization of the process. Our procedure is divided into two steps. The first step takes the analytical trees as its input and produces trees (pre-processed tree structures, PPTS) with the structure close to the tectogrammatical tree and some tectogrammatical level attribute values assigned. Further changes of the tree structure and attribute assignments are done manually.

Then, the second step of the automatic procedure resolves values of grammatemes for the newly added nodes and reduces the redundant information (e.g. agreement).

1.5 Tectogrammatical Tree Structure (TGTS)

By the tectogrammatical tree structure (TGTS) we understand a representation of a sentence deep structure as defined in the Functional Generative Description. TGTS is a dependency tree. Every node N of the tree is a structure $[w, A, G, T, p]$, where w is a canonical form of the represented word (its lemma), A is a set of analytical attributes of the original node coming from the ATS, G is a set of grammatemes, T is a set of technical attributes and p is a pointer to the governor of the node N in the tree. A node represents either an autosemantic word, or a special symbol for coordination. For the tectogrammatical representation of the sentence the set A is not relevant.

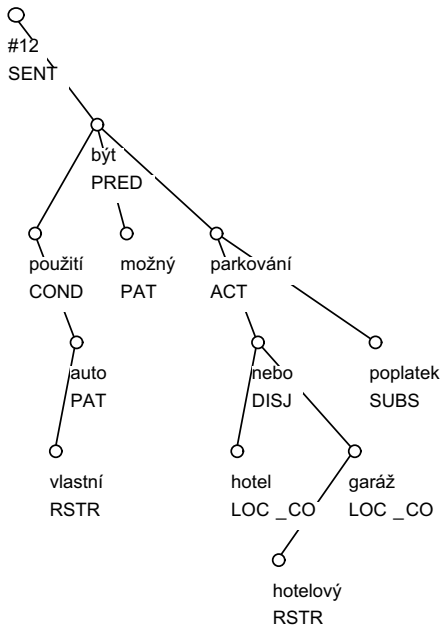

Figure 1.2: Example of a simple TGTS

For the purposes of the transduction from the analytical level to the tectogrammatical level annotation, the tree structure is enriched, no information from the previous levels of annotation is lost. Every node of the TGTS contains all the information inherited from the ATS and there are new attributes added. In the set G in TGTS we can distinguish two types of information contained in every node.

(G1) The tree structure and word order information. The tree structure is based on the ATS and it changes according to the TGTS definition. The information about surface word order (the value of the attribute 'ord') is preserved, while the TGTS is ordered by the 'deep order (dord)' attribute.

(G2) Tectogrammatical attributes. See sections 1.7.2 through 1.7.12 for the description.

The original analytical data A are kept in the original attributes and they are accessible for both the annotators and automatic procedures. However, they are not considered to be a part of the tectogrammatical annotation.

Technical attributes T are used during the process of annotation for storing annotators' comments or results of automatic sub-processes. In the attribute 'fw' the information about the function words, which were conjoined with the auto-semantic node and are marked as hidden in TGTS, is stored.

Table 1.1 lists the attributes of the tectogrammatical tree structure.

(G1)	(A)	(G2)	(T)
ordorig	tagauto	verbmod	warning
ord	tag	trlemma	reserve2
dord	semPOS	TR	reserve1
sentord	origf	tfa	ordtf
	origap	tense	gap3
	lemma	sentmod	gap2
	lemauto	memberof	gap1
	form	quoted	fw
	afunprev	phraseme	err2
	afun	number	err1
		iterativeness	commentTR
		gram	commentAR
		gender	
		func	
		dsp	
		deontmod	
		del	
		degcmp	
		corsnt	
		cornum	
		coref	
		aspect	
		antec	

Table 1.1: Attributes in TGTS

1.6 Annotation: from ATS to TGTS

The transduction from ATSs to underlying trees has the following three parts, the first of which is discussed in more detail in section 1.7:

(i) an automatic 'pre-processing' module,

(ii) an intellectual part, where the human annotators change the analytic functions (esp. Subject, Object, Adverbial, Attribute), into corresponding functors (only the most basic cases are changed automatically). Nodes for the items deleted on the surface level are 'restored' (mostly as pronouns); the TFA indices for focus, contrastive and non-contrastive topic are specified. The 'user-friendly' software enables the annotators to work with diagrammatic shapes of trees;

(iii) a subsequent automatic module adds first of all

(a) information on the lexical values of restored nodes in unmarked cases in which the (marked) values have not been specified in (ii): esp. in coordinated constructions the values of the (symmetric) counterparts in the given construction are added; (b) the

secondary values of syntactic grammatemes (esp. where a preposition allows for a reliable choice);

(c) at the same time, the gender and number values are cancelled whenever they only indicate agreement (as with adjectives in most positions), and

(d) the remaining nodes corresponding to commas, dashes, quotes, etc. are deleted.

1.7 Automatic Pre-processing Module

Let us first introduce the basic notions of hidden and newly added nodes and of the tectogrammatical lemma. Then the following sections decribe the features of TGTS and the changes committed to the ATS during the first step of the transition procedure.

1.7.1 Hidden and added nodes

The nodes abolished in the automatic procedure or during the handcrafted annotations are marked as hidden and are not displayed on the screen in the course of the further processing, but they remain in the tree structure and can be re-introduced if needed. Thus no information gets lost for later research and analysis. The hidden node is indicated by value 'hide' in the attribute 'TR'. A node can be hidden only if all its subtree is hidden. Typically, the hidden nodes are leaves. The hidden nodes are not assigned any tectogrammatical attributes from set G.

On the other hand, nodes for the lexical occurrences that were deleted on the surface are restored and assigned the appropriate attribute values (lemma, functor, etc.). These nodes are marked by a value of "ELID" in the attribute `del`.

1.7.2 Lemma in TGTS

The 'trlemma' attribute contains so-called tectogrammatical lemma of the node. The tectogrammatical lemma of a single node (even if the node is hidden, i.e. marked as absent in the TGTS) is equal to its analytical lemma. The compound nodes that represent more than one word of the surface sentence are assigned the trlemma attribute in the following way:

- *Verbal nodes*: lemma of the autosemantic (lexical) verb.
- *Compound prepositions, conjunctions and numeratives*: trlemma is composed of the lemmas of the parts of the item (e.g. the three nodes representing the numerative 1150 'tisc sto padest' are joined into one node with trlemma = 'tisíc_sto_padesát').
- *Newly added nodes* are assigned either proper lexical values (in case of filled deletions - mostly pronouns), or technical lexical values, such as 'Gen' for the general participant, 'Cor' for the coreferential node of a controllee, or 'Neg' for negation (which in Czech language has the form of a verb prefix) .

- *Punctuation nodes* that are to be preserved at the tectogrammatical level are assigned these values:

, (comma)	Comma
: (colon)	Colon
; (semicolon)	Semicolon
- (dash)	Dash

Table 1.2: Lemmas for punctuation nodes

1.7.3 Morphological grammatemes

The morphological grammatemes are captured in the TGTS using the attributes of gender ('gender'), number ('number'), degree of comparison ('degofcmp'), tense ('tense'), aspect ('aspect'), iterativeness ('iterativeness'), verbal modality ('verbmod'), deontic modality ('deontmod'), sentence modality ('sentmod'). All these attributes belong to set G (introduced in section 1.5).

The list of all possible attribute values follows in the brackets after each attribute name. The detailed description of the values can be found in the full version of my thesis. Each of the morphological grammatemes can be assigned one of the listed values, or value "NA" (not applicable), expressing that the node cannot be assigned any value of this attribute, or value "???" expressing that the value was not determined yet and therefore it needs further attention. The following morphological grammatemes are assigned their values automatically:

1. `gender (ANIM, INAN, FEM, NEUT)`

   The morphological category of gender can be determined for nouns, adjectives, and pronouns, as well as for certain numerals and verb forms. Based on the morphological tag, the grammateme of gender is automatically assigned to all nouns, adjectives and pronouns (if possible). For all remaining word classes it is assigned value NA (not applicable). The ideal objective is for TGTSs to contain these values, and also those of number, whenever their function concerns more than just grammatical agreement.

2. `number (SG, PL)`

   Based on the morphological tag, the grammateme of number is assigned to all nouns, adjectives, pronouns and verbs. For all remaining word classes it is assigned value NA (not applicable).

3. `degcmp (POS, COMP, SUP)`

   The grammateme of degree of comparison is assigned to all nodes representing adjectives and adverbs based on the morphological tag. The remaining nodes are assigned value "NA".

4. `tense (SIM, ANT, POST)`

   The attribute of tense is assigned to verbs. Based on the morphological tag, all verbal nodes are assigned the value of tense. All the non-verbal nodes are assigned "NA" for tense.

5. `aspect (PROC, CPL, RES)` Based on the morphological tag, all verbal nodes are assigned the value of aspect. All non-verbal nodes are assigned "NA". For details see section 1.7.12.

6. `iterativeness (IT1, IT0)`

   The attribute of iterativeness is assigned to verbs. If the morphological tag indicates the verb is iterative, value IT1 is assigned to the verbal node. Otherwise the verbal node is assigned value IT0. All non-verbal nodes are assigned "NA". For details see section 1.7.12.

7. `verbmod (IND, IMP, CDN)`

   The attribute of verbal modality is assigned to verbal nodes. For details see section 1.7.5.

8. `deontmod (DECL, DEB, HRT, VOL, POSS, PERM, FAC)`

   Deontic modality is assigned to the nodes representing verbs or verbal phrases. It mostly corresponds to the meanings of purely modal verbs. In the process of joining the autosemantic verb with its auxiliary nodes into one complex verbal node, the lexical values (attribute `lemma`) of the modal verbs determine the deontic modality of the resulting node. Where the modal verb is not present, the deontic modality defaults to the value "DECL". All non-verbal nodes are assigned "NA" for deontic modality. For details see section 1.7.6.

9. `sentmod (ENUNC, EXCL, DESID, IMPER, INTER)`

   Sentence modality is determined for the head nodes of the clauses. The values are automatically assigned to the head nodes of single sentences and the main clauses of complex sentences, based on the final punctuation and presence of certain words or verb forms in the sentence. For details see section 1.7.12.

1.7.4 Other attributes

Next to the morphological grammatemes the set G contains also attributes describing the position of the node at the tectogrammatical level: topic-focus articulation ('tfa'), functor ('func'), syntactic grammateme ('gram'), type of relation such as dependency, coordination or apposition ('memberof'), phraseme ('phraseme'), deletion ('del'), quoted word ('quot'), direct speech ('dsp'), coreference ('cor'), antecedent ('antec') and some other, technical attributes.

The set of technical attributes T contains the attribute 'function word (fw)' used for storing the preposition or conjunction of the word for the later resolution of the syntactical grammatemes. The attributes of deep order ('dord') and sentence order ('sentord') are used to distinguish between the sentence surface word order and the deep word order. For the detailed description of the attribute values and their automatic assignments, please refer to the full version of my thesis.

The main points of the transduction include:

(a) deleting those nodes of the ATSs which correspond to function words and to most punctuation marks, with an indication of their functions in the form of indices of the corresponding lexical (autosemantic) occurrences; as an exception, we use nodes for coordinating conjunctions (as heads of the coordinated constructions), thus working with underlying representations in the specific form of 'tectogrammatical tree structures' (TGTSs).

(b) assigning every lexical occurrence the appropriate syntactic functors (distinguishing more than 40 kinds of syntactic relations) and morphological grammatemes (marking the values of tense, aspect, modalities, number, etc.), as well as syntactic grammatemes (values such as 'in, on, under, among' with Locative or Directional);

(c) restoring those nodes of TGTSs which are deleted in the surface form of the input sentences;

(d) indicating the position of every node in the topic-focus articulation (TFA) with a scale of communicative dynamism, represented as underlying word order.

1.7.5 Auxiliary verbs, i.e. verbmod attribute

The verb is conjoined with its auxiliary nodes into a complex value of a single node, placed in the highest position in the relevant subtree. All AuxV nodes are hidden. The verb is assigned the values of the grammatemes of tense and verb modality on the basis of the lexical values of these auxiliary nodes. The lemma of the autosemantic verb is put into the trlemma attribute of the remaining node, which is assigned the grammateme values depending on the AuxV dependent nodes. The complete set of rules can be found in my thesis.

Let me bring an example of a rule applied to verbal nodes here: If the verb has as its daughter neither a node with the lemma "být" nor with lemma "by", disregarding the possible presence of "se", and the verb in in preterite tense, then assign the verb attribute tense the value ANT.

Examples:

1. otevřel.VR se.AuxT => otevřít_se.ANT *(it) opened*

2. učil.VR by.AuxV se.AuxT => učit_se.SIM.CDN *(he) would learn*

3. byl.AuxV by.AuxV spal.VR => spát.ANT.CDN *(he) would have slept*

1.7.6 Modal verbs, i.e. deontmod attribute

The modal verb is merged with the autosemantic verb depending on it in the ATS. The transduction procedure consists of three steps: the tree is rearranged in that the modal verb depends on the autosemantic verb, which ensures that the autosemantic verb appears in the right position after the transformation, the modal verb node becomes a leaf node and therefore can be deleted. Then the value for the deontmod of the latter verb is assigned its value according to the lexical value of the modal verb, and the modal verb node is deleted.

Modal verb	English transl.	Auto-semantic verb form	f of the verb	deontmod assigned
Chtít	want	infinitive	object	VOL
Muset	must	-		DEB
Moci, dát_se	can	-		POSS
Smět	be allowed	-		PERM
Umět, dovést	can	infinitive	object	FAC
Mít	should	infinitive	object	HRT

Table 1.3: Modal verbs

1.7.7 Prepositions and conjunctions, i.e. fw attribute

Every preposition node is deleted and its lexical value is stored in the attribute `fw` of the noun. The preposition will be used for the future (at least partly automatized) determinantion of the value of the syntactic grammateme of the noun.

Every subordinating conjunction node is deleted. Its lexical value is stored in the `fw` attribute of the head verb of the subordinate clause. Conjunctions for coordination and apposition are used in the tectogrammatical tree as the heads of the coordinated clauses.

1.7.8 General participants

The reflexive particle 'se' has three possible analytical functions in a Czech sentence. The analytical function value 'AuxT' is assigned on the analytical level to a reflexive 'se' having the function of lexical derivation (of a middle verb). The analytical form of 'se' is conjoined with the lemma of the verb in such case. If 'se' was assigned the function 'AuxR' at the analytical level, it expresses a general actor of the verb. The node is preserved, its attribute trlemma is filled with the 'Gen' value and its functor is 'ACT'. If 'se' was assigned the analytical function 'Obj', it gets assigned the functor 'PAT'.

1.7.9 Quotation marks, i.e. quot attribute

The sentence is searched for quotation marks. If a whole clause, having the form of a sentence, is inserted into a pair of double quotes, its verb obtains the value 'DSP' (direct speech) on the attribute quot. If only one token of a quote appears in the sentence, the attribute quot of the head word(s) of the string containing the quote is assigned 'DSPP' value (direct speech part). Otherwise, the head word(s) of string enclosed in quotes is/are assigned quot = 'QUOT' (quoted word).

1.7.10 Punctuation

All punctuation nodes (which have the analytical function 'AuxX') are hidden except for the following two cases:

- A comma placed in the sentence in the position directly following a noun is left in the sentence to enable the annotators to decide about the type of the adjunct (restrictive or descriptive),
- A comma which is a bearer of coordination or apposition not indicated by a conjunction (its analytical function is 'Coord' or 'Appos') is not deleted.

The trlemma attribute of undeleted comma node is filled with Comma value.

1.7.11 Node for negation

Every verb, adjective and adverb is checked. If its morphological tag contains the symbol for negative expression (i.e. the word is derived using the prefix "ne-"), a new node is created with the lexical (trlemma) value 'Neg' and functor 'RHEM' (rhematizer, i.e. focus sensitive particle) depending on the negative word.

1.7.12 Other attribute assignments

Based on the morphological tag inherited from the analytical level of description, the values of the following morphological grammatemes are assigned: gender , number, tense, degcmp (degree of comparison), aspect.

The sentence modality is captured in the sentmod attribute of the head node of each clause. We assign the sentence modality of the head word of a simple sentence, of the main clause of a complex sentence and of all coordinated clauses in compound sentences. The sentence modality attribute value is given by the final punctuation mark of the whole sentence and by the verb modality of the main verbs of the sentence clauses. Suppose we have a sentence composed of coordinated clauses $X_i : X_1, X_2, ..., and X_n$. The rules are described by Table 1.4.

As for functors, their value is resolved automatically in the following three cases. Value 'ACT' (actor) is assigned to every node indicated in the ATS as a subject of an active verb. If there is a subject and an 'object' (i.e. a correlate depending on a passive verb,

Position in clause X_i	final punctuation	verb modality	Sentence modality of X_n	other conditions	verb modality assigned
X_n (verb	?	-	-	-	INTER
in the last or in	!	-	-	-	IMPER
the only clause)	.	-	-	-	DECL
$X_i : X_1, X_2, \ldots$ $\ldots, and X_{n-1}$	-	-	INTER	-	INTER
For n¿1	-	IND	-	-	ENUNC
	-	IMP	-	-	IMPER
	-	CDN	-	X_i contains 'kéž' (E:'let')	DESID
	-	CDN	-	otherwise	ENUNC

Table 1.4: Sentence modality assignment

these two nodes are assigned functor 'PAT' and 'ACT', respectively. The head verbs of the sentences are assigned the functor 'PRED' (predicate).

1.7.13 "Default" values

Unresolved syntactic and morphological grammatemes are assigned their default value in the automatic procedure. By the default value we understand 'NA' value for attributes that cannot be assigned any value for the given node (e.g. aspect for non-verbal nodes), or another value is chosen to express the uncertainty for the annotators (e.g. value "???" for unresolved func attribute). For the following attributes the more probable value is assigned, to simplify the work for the annotators: `tfa` = 'F'.

1.8 The manual part of the annotation

The annotators' main task in the manual phase of the annotation is to assign the functors (attribute '`func`') to all the nodes in the TGTS. The preprocessed structure can be also changed, if necessary. Therefore we have developed interactive macros for the annotators, that simplify editing the trees. The macros used during the annotation are of the following two types: functor assignments and structure changes (usually accompanied by certain attribute assignments). The interactive functions are called either by selecting the macro from the list of macros, or they have a hotkey sequence assigned.

Attribute	Condition	Value
verbmod	non-verbal node	NA
tfa		F
tense	non-verbal node	NA
sentmod	other node than main verb	NA
memberof	unresolved	NA
quoted	unresolved	NA
phraseme		NA
number	non-verbal node	NA
iterativeness	non-verbal node	NA
gram	unresolved	???
gender	unresolved	NA
func	unresolved	???
dsp	unresolved	NA
deontmod	unresolved	NA
del		NIL
degcmp	other than adjective node	NA
corsnt		???
cornum		???
coref		???
aspect	non-verbal node	NA
antec		???

Table 1.5: Default values

1.9 Enrichment of the automatic module - near future prospects

A richer version of automatic analysis is being prepared, based on the experience from the present stage of the tagging. The new version will take over some further tasks handled by the manual procedure up to now.

The automatic procedure will be enriched in various respects, such as the build-up of the lexicon (with entries including the valency frames), word derivation, and the degrees of activation of the 'stock of shared knowledge,' as far as derivable from the use of nouns and pronouns in subsequent utterances. Several types of grammatical information, e.g., the disambiguated values of prepositions and conjunctions, can only be specified after further empirical investigations, in which, whenever possible, also statistical methods will be used. In any case, the annotated corpus will offer a suitable starting point for monographic analysis of the problems concerned. The weight of the automatized part of the procedure

will increase step by step, thanks to a series of improvement iterations.

Bibliography

[1] 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.

[2] Hajičová E. (1993) *Issues of sentence structure and discourse patterns*. Charles University.

[3] 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.

[4] J. Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Charles University Press - Karolinum, in press *Disertace*

[5] Hladká B. (2000) *Disertace*

[6] Sgall P., E. Hajičová and J. Panevová (2000) *Manuál pro tektogramatické značkování*

[7] Béemová A., Buráňová E., Hajič J., Kárník J., Pajas P., Panevová J., Štepánek J., Urešová Z. (1997) Anotace na analytické rovině - příručka pro anotátory, Technical Report #4, LJD UFAL MFF UK, Prague, Czech Republic.

[8] M. M. Marcus, B. Santorini and M.-A. Marcinkiewicz. Building A Large Annotated Corpus of English. The Penn TreeBank. *Computational Lingistics, 20(2)*, pp. 313-330, 1993

[9] J. Hajič, E. Hajičová, J. Panevová and P. Sgall. Syntax v Českém národním korpusu [Syntax in the Czech National Corpus]. In *Slovo a slovesnost*, 3, LIX, pp. 168-177, 1998.