Automatic procedures in tectogrammatical tagging

Alena BÖHMOVÁ
ÚFAL MFF UK
Malostranské nám. 25
118 00 Prague, Czech Rep.
bohmova@ufal.mff.cuni.cz

Petr SGALL
ÚFAL MFF UK
Malostranské nám. 25
118 00 Prague, Czech Rep.
sgall@ufal.mff.cuni.cz

1 Introduction

A semi-automatic syntactic annotation of a part of the Czech National Corpus in the Prague Dependency Treebank (PDT) has among its aims the possibility to check the theoretical approach chosen (Functional Generative Description, see [2]). While the first phases of the annotation of PDT, i.e. the morphemic representations and the dependency trees on an intermediate analytic level, i.e. analytic tree structures (ATSs, see [1]) have been discussed elsewhere, the present paper is devoted to the second, basic phase, the transduction from AL to (underlying) syntax itself, i.e. to tectogrammatical representations, which should be provided for 10 000 sentences during the year 2000 (at its start, 100 000 sentences have obtained their ATS annotations).

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) occurences; 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 (which distinguish 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.

2 Automatic parts of transduction:

The transduction from LAs to underlying trees has the following three parts, the first of which is discussed in more detail in Section 3:

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

(ii) an intellectual part, which changes the analytic functions (esp. Subject, Object, Adverbial, Attribute), into corresponding functors (only the most basic cases are changed automaticaly); nodes for the deleted items are 'restored' (mostly as pronouns); the TFA indices for focus, contrastive and non-contrastive topic are specified; a '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;

(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.

In the next months, the automatic procedure is supposed to 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.

3 The first part of the automatic transduction

3.1 TGTS description

Every node of the TGTS contains all the information inherited from the ATS and there are new attributes added.

The `trlemma` attribute contains the lemma of the node. The `trlemma` 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 (main?) verb.

Compound prepositions, conjunctions and numeratives: trlemma is composed of the lemmas of the parts of the preposition (e.g. the three nodes representing numerative 1150 ‘tisíc sto padesát’ 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 correferential node of a controlee, or ‘Neg’ for negation.

The morphological grammatemes are captured using these attributes: gender, number, degree of comparison, tense, aspect, iterativeness, verbal modality, deontic modality, sentence modality.

Next to the morphological grammatemes there are attributes describing the node at the tectogrammatical level: topic-focus articulation, functor, syntactic grammateme, type of relation (dependency, coordination, apposition), phraseme, deletion, quoted word, direct speech, coreference, antecedent and some other technical attributes. The attribute ‘function word (`fw`)’ is used for storing the preposition or conjunction of the word for the later resolving of the syntactical grammatemes. The attributes ‘deep order (`dord`)’ and ‘sentence order (`sentord`)’ are used to distinguish between the sentence surface word order and the deep word order.

3.2 The steps of the procedure

3.2.1 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 table below shows what assignments are made in the automatic procedure for the verbal node. The rules are captured in the table rows. E.g. the second row of the table says: 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 morphological tag of the verb begins “VR” (symbol for preterite tense), and assign the verb attribute `tense` the value ANT. All the rules are applied in the sequence given by Table 1.

Presence of dependent node with lemma			**Morph. tag of the verb**	**Assigned attributes**
být (to be)	**by (cond.)**	**se, f=AuxT**		
-	-	yes	-	trlemma => join '_se' to the trlemma of the verb
no	no	-	VR	tense => ANT
no	no	-	VU	tense => POST
no	no	-	other	tense => SIM

no	yes	-	-	tense => SIM
				verbmod => CDN

yes	yes	-	-	tense => ANT
				verbmod => CDN

Table 1. Verbs

Examples:

(i) **otevřel**.VR **se**.AuxT =>
otevřít_se.ANT
(it) opened
(ii) **učil**.VR **by**.AuxV **se**.AuxT =>
učit_se.SIM.CDN
(he) would learn
(iii) **byl**.AuxV **by**.AuxV **spal**.VR =>
spát.ANT.CDN
(he) would have slept

3.2.2 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 in three steps: the tree is rearranged in that the modal verb depends on the autosemantic verb, the value for the attribute deontmod of the latter verb is assigned 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 2. Modal verbs.

3.2.3 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 subordinate 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.

3.2.4 General actor

The reflexive particle 'se' has three possible analytical functions in a Czech sentence. The analytical function value AuxT is assigned to a reflexive 'se' having the function of lexical derivation (of a middle verb). As shown in Table 1, '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 function 'OBJ', it gets the functor 'PAT'.

3.2.5 Quotation marks, i.e. `quot` attribute

The sentence is searched for quotation marks. If a whole clause 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).

3.2.6 Punctuation

All punctuation nodes (which have the analytical function 'AuxX') are hidden except 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 is not deleted.

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

3.2.7 Node for negation

Every verb is checked. If its morphological tag contains the symbol for negative verb, a new node is created with lexical (trlemma) value 'Neg' and functor 'RHEM' (rhematizer, i.e. focus sensitive particle).

3.2.8 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 subordinate 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. The rules are described by Table 3.

Suppose we have a sentence composed of coordinated clauses X_i: X_1, X_2, ..., and X_n.

Position in clause Xi	**final interp.**	**verb modali-ty**	**Sentence modality of Xn**	**other conditions**	**verb modality assigned**
Xn (verb in the last or in the only clause)	**?**	-	-	-	INTER
	!	-	-	-	IMPER
	.	-	-	-	DECL
X1, ..., Xn-1 **For n>1**	-	-	INTER	-	INTER
	-	IND	-	-	ENUNC
	-	IMP	-	-	IMPER
	-	CDN	-	X_i contains 'kéž' (E:'let')	DESID
	-	CDN	-	otherwise	ENUNC

Table 3. Sentence modality assignment

As for functors, their value is resolved automatically in the following three cases. Value 'ACT' (actor) is assigned to every subject of an active verb. If there is a subject and an object depending on a passive verb, these two nodes are assigned functor 'PAT' and 'ACT', respectively. The head verbs of the sentences are assigned the functor 'PRED' (predicate).

3.2.9 „Default" values

Unresolved syntactic and semantic grammatemes are assigned their default value by the procedure. By the default value we understand either the most common value for that attribute (e.g. F as focus for topic-focus articulation, `tfa` attribute), 'NIL' value for attributes that cannot be assigned any value for the given node (e.g. case for verbal nodes), or it is chosen to express the uncertainty for the annotators (e.g. value "???" for unresolved `func` attribute).

References

[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.