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

Abstract

This paper describes a specific part of the Prague Dependency Treebank annotation, the step from the surface dependency structure towards the underlying representation of the sentence. The first section explains the theoretical basis of the project. In Section 2 all the procedure of conversion to the tectogrammatical structure is summarized and Section 3 presents in detail the present stage of the automated part of the conversion procedure.

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]), as well as to establish a basis that could serve as a suitable starting point for a large-scale monographic analysis of the numerous problems of the sentence structure in general and of the grammar of Czech in paritcular which still require empirical research. Such an analysis is expected to be helpful at least in three respects:
(i) to make a relatively complete description of Czech into a realistic task,
(ii) to fill lacunes present in the prelimnary form of the annotation procedures formulated up to now, and
(iii) to proceed towards a procedure that would be automatized to a maximally high degree.
We do hope that our paper may be useful in attracting more attention not only to the need of an annotation reaching the underlying sentence structure (rather than just the usual 'surface-structure' parsers, which may help in natural language processing, although not that much in achieving the aims of a theoretical linguistic description), but also to a deep-reaching comparison of different approaches to syntax. We are convinced that dependency-based syntax with its maximally economical tree structures may be of particular interest for the aims of contemporary computational linguistics. This holds especially of an approach in which function words are classed together with inflectional morphemes as corresponding to indices in complex node labels, rather than to specific nodes, and in which also other aspects of the underlying, tectogrammatical, structure are established. Moreover, a comparison of the problems concerned in syntactic annotation procedures for languages of different types certainly can be important if general theories of language description are looked for and compared with each other; such a theory should show the core of linguistic structure to be economical enough both to help explain the easiness of mother tongue acquisition and to be implementable in computers.
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. A detailed discussion of tectogrammatics, which cannot be included into the present paper, can be found in [2], [4].

The following strategy of annotation has been found useful, and this may hold also for many other languages: The first phases of the annotation of PDT are (i) the morphemic representations and (ii) the dependency trees on an intermediate ('surface') analytic level, i.e. analytic tree structures (ATSs, see [1]), where (i) has to use a combination of statistical and structural methods to obtain a reliable automatic treatment, and (ii) has to be carried out manually. While (i) and (ii) have been discussed elsewhere, the present paper is devoted to a subsequent phase (iii), the transduction (conversion) from ATSs 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 (i.e. autosemantic, rather than auxiliary) 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, i.e. of kinds of valency slots, e.g. PAT (patient or objective), ADDRessee, LOCative, MANNer) 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 (see [2], [3] for a discussion of TFA).

2 Automatic parts of transduction:

The transduction from ATSs 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) a manual part, which changes the analytic functions (esp. Subject, Object, Adverbial, Attribute), into corresponding functors (only the most basic cases are changed automatically); 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 are added;

(b) certain 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 elaboration 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 new attributes are 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 assigned in the ATS. 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 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 coreferential node of a controlee, or 'Neg' for negation.

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

Next to the morphological grammatemes there are attributes describing the position of 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 resolution 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 tables below show what assignments are made in the automatic procedure for the verbal node. Table 1 contains the rules applied to the nodes for autosemantic verbs, the rules are captured in the table rows in the sequence they are being used. If all the conditions are fulfilled for some node, the rule is applied. E.g. the second row of the table reads as follows: If the verb daughter node is labelled either with the lemma "být" or"by", disregarding the possible presence of "se" (which was already handled by rule 1), and the morphological tag of the verb begins "VR" (symbol for preterite tense), then assign the verb attribute `tense` the value ANT.

	Presence of dependent node with lemma			Morph. tag of the verb	Assigned attributes
no	**být (to be)**	**by (cond.)**	**se, f=AuxT**		
1	-	-	yes	-	trlemma => attach '_se' to the trlemma of the verb
2	no	no	-	VR	tense => ANT
3	no	no	-	VU	tense => POST
4	no	no	-	other	tense => SIM
5	no	yes	-	-	tense => SIM verbmod => CDN
6	yes	yes	-	-	tense => ANT verbmod => CDN
7	yes	no	-	-	tense => ANT

Table 1. *Verbs*

Examples:

(i) **otevřel**.VR **se**.AuxT =>
trlemma=**otevřít_se** (rule 1)
tense**=ANT** (rule 2)
E: *(it) opened*

(ii) **učil**.VR **by**.AuxV **se**.AuxT =>
trlemma=**učit_se** (rule 1)
tense**=SIM,** verbmod=**CDN** (rule 5)
E: *(he) would learn*

(iii) **byl**.AuxV **by**.AuxV **spal**.VR =>
trlemma**=spát**
tense=**ANT**, verbmod=**CDN** (rule 6)
E: *(he) would have slept*

(iv) mohla jsem být (já) spatřena
trlemma**=spatřit**
tense=**ANT** (rule 7)
deontmod=**POSS**

E: *(I) could have been seen*

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 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 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) determination 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.

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 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 double 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 the 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 for the following two cases:

- the node for a comma placed in the position directly following a noun is left in the tree to enable the annotators to decide about the type of the adjunct (restrictive or descriptive),
- a comma node that is a bearer of coordination or apposition is not deleted, as far as this function can be recognized from the ATS.

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 the 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 complex sentence and of all coordinated clauses in compound sentences. The sentence modality attribute value is determined 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
	.	-	-	-	ENUNC
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/bearer, underlying subject) is assigned to every subject of an active verb. If there is a single object depending on an active verb, its node is assigned functor PAT (patient, objective). The head verbs of the sentences are assigned the functor PRED (predicate).

Example:

(i) **Sestra**.Sb **spatřila**.A **souseda**.Obj.
ACT PAT
E: *sister spotted the neighbour*

3.2.9 „Default" values

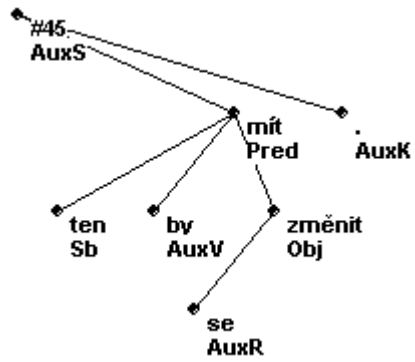
Unresolved syntactic and morphological grammatemes are assigned their default value by the procedure. By the default value we understand 'NIL' value for attributes that cannot be assigned any value for the given node (e.g. case for verbal nodes), or it is chosen to express

the uncertainty for the annotators (e.g. value "???" for unresolved `func` attribute).

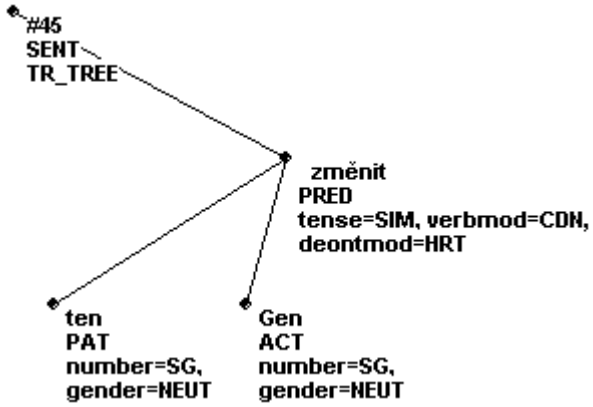
3.3 Example of input and output

Sentence: *To by se mělo změnit.*
That should (itself) change.
Meaning: That should be changed.

ATS:



TGTS:



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] Hajičová E., B. H. Partee and P. Sgall (1998) *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Dordrecht:Kluwer.

[4] 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.