Prague Dependency Treebank: Restoration of Deletions *

Eva Hajičová, Ivana Kruijff-Korbayová, and Petr Sgall

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic,
{hajicova,korbay,sgall}@ufal.mff.cuni.cz,
WWW home page: http://ufal.mff.cuni.cz

Abstract. The use of the treebank as a resource for linguistic research has led us to look for an annotation scheme representing not only surface syntactic information (in 'analytic trees', ATS) but also the underlying syntactic structure of sentences and at least some aspects of intersentential links (in 'tectogrammatical tree structures', TGTS). We focus in this paper on some of the issues of the transduction of ATSs into TGTSs.

1 Two steps of syntactic tagging in PDT

In the Prague Dependency Treebank (PDT) project, the structure of sentences is made explicit by means of two steps of syntactic tagging resulting in:

(i) 'analytic' tree structures (ATSs), in which every word form and punctuation mark is represented as a node of the tree, and the edges of the tree correspond to (surface) syntactic dependency relations; and,

(ii) tectogrammatical tree structures (TGTSs) corresponding to underlying sentence representations and having the shape of dependency trees with the verb as the root of the tree.[1] In TGTSs the functional (synsemantic) words (such as prepositions, auxiliaries, subordinating conjunctions) as well as punctuation marks are principally not represented by nodes of their own; their functions are captured as parts of complex tags of the nodes standing for autosemantic (content) words. Surface deletions are 'restored' in TGTSs.

The syntactic information which is absent in the surface (morphemic) shape of the sentence is introduced - at least for the time being - in the manual phase of the transduction procedure ([Hajičová et al. 1998]), translating (in a 'user-friendly' environment) ATSs to TGTSs. Every added (restored) node gets the index ELEX (if its antecedent is an expanded head node) or ELID (if this is not so). The added nodes always depend on their governors from the left-hand side, except for certain cases in coordinated constructions (cf. (2) below).

* The work reported on in this paper has been supported by the grant of the Czech Ministry of Education VS 96/151 and by the Czech Grant Agency GAČR 405/96/K214.

[1] With the exception of TGTSs for coordinated constructions, see below.

A specific case concerns coordinating conjunctions: although they belong to function words, they retain their status as nodes (labeled as CONJ, DISJ, etc.) in the TGTSs, which in this point differ from the theoretically substantiated form of tectogrammatical representations. This exception makes it technically possible to work with rooted trees, rather than with networks of more dimensions. One-to-one linearization of ATSs and TGTSs has been defined, which will be applied below, when presenting our examples of TGTSs.

2 Types of lexical labels of the added nodes

Two cases of node restoration according to the character of the lexical labels of the restored nodes can be distinguished: (a) restoration of full lexical information (i.e. adding a node with a particular lexeme in its label), and (b) restoration of a pronominal (anaphoric) element.

2.1 Restoration of full lexical information

The lexical part of the complex label of the 'restored' (added) node consists in a particular lexeme, including a lexeme with a 'general' meaning, in the following situations:

(i) In coordination: The restored node (included in square brackets in our examples) can be either a dependent node, as in (1), or a governor, as in (2).[2]

(1) nové knihy a časopisy ⇒ nové knihy a [nové] časopisy
new books and journals ⇒ new books and [new] journals

(2) červená a modrá barva ⇒ červená [barva] a modrá barva
red and blue paint ⇒ red [paint] and blue paint

We give precedence to a "constituent" coordination before a "sentential" one, whenever possible. Thus in the TGTS for (3) neither the Actor *Jirka* nor the Objective *Marii* will be 'doubled' because the coordination of the two verbs *potkal* and *pozdravil* will be treated as a coordination of two verbs that have a single Actor and a single Objective in common.

(3) Jirka potkal a pozdravil Marii.
George met and greeted Mary.

The complex labels for the coordinated nodes include a special symbol CO to distinguish them from nodes that modify the coordination as a whole. Thus, a simplified linearized representation (only with the lexical labels representing the respective nodes and with every dependent enclosed in a pair of parentheses) for (3) is given in (3′).

[2] It should be noted that we give here only one of the possible interpretations of (1); (1) can be also understood as '(nové knihy) a (časopisy)', where no restoration occurs.

(3′) (Jirka) (potkal.CO) CONJ (pozdravil.CO) (Marii)

Sentence (4) is an example of the addition of a node that stands for a whole structure; in such a case this 'restored' node carries the label ELEX (for an expanded deleted item), see (4′):

(4) Jirka potkal Marii včera a já dnes.
George met Mary yesterday and I today.

(4′) ((Jirka) potkal.CO (Marii) (včera)) CONJ ((já) potkal.ELEX.CO (dnes))

(ii) In cases of so-called 'general participants': Among the items that are often deleted in the surface, there is the case of an Actor or another argument (inner participant) of a verb with the meaning of 'general' (coming close to the English *one* or German *man*, as for the subject). This argument is represented in the TGTSs as a node with the lexical value 'Gen'; cf. the following examples, for which we adduce linearized representations:

(5) Ten dům byl postaven ve dvacátých letech.
That house was built in the-twenties years.

(5′) ((ten.Restr) dům.Pat) (Gen.ELID.Act) postavit ((rok.Temp (dvacátý.Restr))

(6) Ta trouba dobře peče.
That oven well bakes.

(6′) ((ta.Restr) trouba.Act) (Gen.ELID.Pat) péct (dobře.Mann)

(7) Dědeček dobře vypravuje pohádky.
Grandfather well tells fairy-tales.

(7′) (dědeček.Act) (Gen.ELID.Addr) vypravuje (dobře.Mann) (pohádky.Pat)

The General Actor can also be expressed by the so-called reflexive passive; in that case the node corresponding to the particle *se* occurring in ATS gets the lexical label Gen with the functor Act (without ELID).

(8) Domy se stavějí z cihel.
Houses Refl built from bricks.
(Houses are built from bricks.)

(8′) (dům.Pat) (Gen.Act) stavět (cihla.Orig)

(iii) In case of zero subject with infinitive: The so-called verbs of control take an infinitive as their Object (Patient) and their Actor or Addressee is referentially identical to the (deleted) 'subject' of the infinitive. Thus, the Actor of the main clause is such a 'controller' in (9), and the Addressee in (10):

(9) Jirka slíbil matce přijít domů včas.
Jirka promised mother to-come home in-time.

(9′) (Jirka.Act) slíbit (matka.Addr) ((Jirka.ELID.Act) přijít.Pat (domů.Dir) (včas-Temp)

(10) Rodiče žádali Jirku nechodit tam.
Parents asked George not-to-go there.

(10′) (rodiče.Act) žádat (Jirka.Addr) ((tam.Dir) (Jirka.ELID.Act) nechodit.Pat)

A similar structure is present if the infinitive is passivized:

(11) Richard se bál být spatřen.
Richard Refl. was-afraid to-be seen.

(11′) (Richard.Act) bát-se ((Richard.ELID.Pat) (Gen.Act) spatřit)

(iv) Cases of a deleted "non-omissible" obligatory participant: With certain verbs, an argument can only be deleted if it is given in the immediately preceding co-text, cf. (12):

(12) (Potkal Milan Jirku?)
Potkal. (Has-met Milan George?) Met-Masc.

(12′) (Milan.Act.ELID) potkat (Jirka.Pat.ELID)

In cases (i) through (iv), full lexical items can be identified as antecedents by the annotator, and thus they are placed into the positions of the deleted tokens. With the exception of (iv), the possibility (or necessity) for the relevant item to be deleted is determined by the grammatical structure of the sentence. In (iv), the specific lexical value of the restored item reproduces that of the overt item present in a structurally corresponding position in the immediately preceding utterance.

2.2 Restoration of a pronominal (anaphoric) element

A prototypical context in which a pronominal rather than a lexically fully specified element is added to the tree structure, is that of zero subjects with finite verbs (Czech is a so called pro-drop language):

(13) Přišel pozdě.
Came-masc. late

(He came late.)

(13′) (on.ELID.Masc.Act) přijít (pozdě.Temp)

(14) Přišla pozdě.
Came-fem. late

(She came late.)

(14′) (on.ELID.Fem.Act) přijít (pozdě.Temp)

If we compare example (9) above with (15), the respective TGTSs in (9′) and (15′) reflect the difference between two kinds of coreference: one given grammatically by the properties of Czech verbs of control, and the other determined by the context, which may even go beyond the sentence boundary (*he* is not necessarily coreferential with *Jirka*).

(15) Jirka slíbil matce, že přijde domů včas.
Jirka promised mother that he-would-come home in-time

(15′) (Jirka.Act) slíbit (matka.Addr) ((on.ELID.Act) přijít.Pat (domů.Dir) (včas-Temp))

2.3 Borderline examples

Cases in which an omissible obligatory complementation is deleted constitute a special group of deletions. These cases differ from (12) quoted in Section 2.1(iv) in that they concern a deletion licensed by the valency frame of the given head word: the frame includes the respective complementation (be it a participant or an adverbial modification) as semantically obligatory, but omissible on the surface. In case of its deletion in the surface shape of the sentence, its lexical value is chosen according to the context: e.g., with the verbs přijít 'to come' or odejít 'to leave' the choice is between sem/odsud 'here/from here' and *tam//odtamtud* 'there/from there'. In the TGTSs, this ambiguity is to be resolved, which is possible on the basis of the context (not grammatically); for a characterization of intersentential coreference see [Hajičová 1999].

2.4 Special cases

Among the special cases of adding some information that is not present (or is only implicitly present) in ATSs, there are two that deserve a special mentioning:

Case of sentence negation In Czech, negation of verbs is expressed by a negative prefix *ne-* attached to the affirmative form of the verb. In ATSs, the negative verb is thus treated as a single node. However, the semantics of negation and its relationship to the topic-focus articulation of the sentence makes it necessary to introduce into the TGTSs a special node for the operator of negation derived from the negative prefix of the verb and having the lexical value Neg. The Neg node depends on the verb; if the verb has the value F (contextually non-bound, in the focus) in its TFA attribute, Neg is placed to the left of the verb and has also the value F in the TFA attribute (this is the interpretation of negation in (16)). If the verb has the value T (contextually bound, in the topic) in its TFA attribute, Neg is placed either to the left of the verb and has also the value T in the TFA attribute (situation exemplified by (17)), or to the right with the value F (exemplified by (18)).

(16) (Co je s Honzou? Proč pláče?) Honza nespí únavou.
(What is the matter with Honza? Why is he crying?) Honza doesn't sleep due to fatique.

(17) (Proč Honza nespí?) Honza nespí, protože je unaven.
(Why doesn't Honza sleep?) Honza doesn't sleep, because he is tired.

(18) (Myslíš, že Honza spí, protože je unaven?) Honza nespí, protože je unaven, ale protože si vzal silný prášek na spaní.
(Do you think that Honza sleeps because he is tired?) Honza doesn't sleep, because he is tired, but because he took a strong sleeping pill.

Restoring grammatical values rather than entire nodes In some cases it is necessary to add some values of attributes to existing nodes. This occurs e.g. when the grammatical information is to be derived from function words or from morphemic forms; in the automatic module of the procedure translating ATSs to TGTSs, this grammatical information would only be added to one of the nodes standing in the coordination relation, see (19).

(19) Vláda musela odložit pravidelnou schůzi a svolat zasedání zvláštní komise pro bezpečnost.
The government had to adjourn the regular meeting and to convene a meeting of a special committee for security.

The modality expressed by the (function) modal verb *musela* is attached as a value of the attribute of modality with the verb *odložit*; it is necessary, however, to fill in the same attribute with the same value also with the (coordinated) verb *svolat*.

3 Summary

We have outlined one aspect of the difference between ATSs and TGTSs, namely the situation when the ATSs do not contain all the information that belongs to the tectogrammatical structure of the sentence. The restoration of the syntactic information absent in the surface (morphemic) shape of the sentence is done in the manual phase of the transduction procedure; however, the 'user-friendly' environment developed for transduction of ATSs to TGTSs is designed in such a way that it will be possible to include there automatic procedures that will fulfil some of the transduction tasks.

References

[Hajičová et al. 1998] Hajičová E.: Prague Dependency Treebank: From analytic to tectogrammatical annotations In: *Text, Speech, Dialogue* (eds. P. Sojka, V. Matoušek, K. Pala and I. Kopeček), Brno: Masarykova univerzita. (1998) 45-50.

[Hajičová 1999] Hajičová E.: The Prague Dependency Treebank: Crossing the sentence boundary. (this volume)