The Prague Dependency Treebank: Crossing the Sentence Boundary

Eva Hajičová

Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
E-mail: hajicova@ufal.mff.cuni.cz

Abstract. The units processed by tagging procedures - both automatic and manual - are sentences (as occurring in the texts in the corpus), but the human annotators are instructed to assign (disambiguated) structures according to the meaning of the sentence in its environment, taking contextual (and factual) information into account. We focus in the paper on two issues: how to capture (i) the topic-focus articulation as one of the fundamental properties of sentence structure, which is related to the use of the sentence in a broader context, be it a suprasentential or a situational one, and (ii) the coreferential links in the text.

1 Introduction: The Prague Dependency Treebank

The Prague Dependency Treebank (PDT) project is conceived of as a collection of tree structures representing sentences of (a part of) the Czech National Corpus (CNC) in the shape of syntactic trees (tagged both on the analytical and the tectogrammatical levels, in addition to the morphological tags). The tagging on the tectogrammatical layer is based on the theoretical framework of Functional Generative Description (FGD, see [3]). The units processed by tagging procedures - both automatic and manual - are sentences (as occurring in the texts in the corpus) but the human annotators are instructed to assign (disambiguated) structures according to the meaning of the sentence in its environment, taking contextual (and factual) information into account. Another aspect that has led us to think about the context in which the given sentence occurs, is the regard to the use of the PDT as a resource for linguistic research not only within the limits of the sentence. These two considerations have their consequences for several points in the specification of the tectogrammatical tree structures (TGTSs), of which we would like to concentrate in our paper on the attribute TFA and on the order of nodes (Sect. 2), adding some preliminary remarks on some special attributes reflecting the linking of sentences in the text (Sect. 3).

2 Representing Topic-Focus Articulation in TGTSs

2.1.0 One of the basic claims of the theoretical framework of Functional Generative Description (FGD) concerns the relevance of topic-focus articulation (TFA) for the meaning of the sentence; the representations of meaning (tectogrammatical representations, TR's) thus capture both the syntactic (dependency) relations and the TFA. TFA reflects the communicative function of the sentence: topic can be informally paraphrased as what the sentence is about' and focus as the information that is 'asserted' about the topic. As a matter of fact, this dichotomy is derived from the primary distinction of contextually bound and non-bound nodes in the syntactic tree and from the underlying order of nodes (corresponding to the so-called communicative dynamism). A detailed empirical analysis and a formal account of TFA is given in the writings quoted below;

2.1.1 The relevance of TFA for the meaning of the sentence, and thus also for annotations on the underlying level, can be illustrated by the following cases:

(i) The semantic relevance of TFA has already been pointed out by [2] and is exemplified in (1); in Czech no special constructions to change the TFA are needed in this case. (The capitals denote the placement of the intonation centre.)

(1) a. English is spoken in the SHETLANDS.
b. In the Shetlands, one speaks ENGLISH.

The sentence (1)(b) is certainly true (at least in the actual world); this is not the case with (a), which in a prototypical situation implies that the Shetlands are the only (or, maybe, the most important, prevailing etc.) country where *English* is spoken. In (a), we speak 'about' English (*English* is in the topic), and inform the hearer in which countries it is spoken; in (b), we speak 'about' the Shetlands (*the Shetlands* is in the topic) and state which language is spoken there.

(ii) The semantic relevance of TFA is also supported by the semantics of negation, which shows a close relation to the position of negation in topic or focus:

(2) John didn't come because he was ill.
a. The reason for John's not-coming was his illness.
b. The reason for John's coming (e.g. to the doctor) was not his illness but something else (e.g. he wanted to invite the doctor for a party).

If (2) is uttered with the meaning of (b), John might have been ill but not necessarily so, and it is implied that John did come, while (a) implies that John was ill and he did not come. This difference in meaning is a result of different TFA's of (a) and (b): (a) is 'about' John's not-coming (*John didn't come* is in the topic part of the sentence); when the operator of negation is in the topic, the end of the scope of negation coincides with the boundary between topic and focus and the elements in the focus trigger, in a prototypical case, a presupposition. In (b), the sentence is about John's coming, and what is negated is that the reason was not his illness; the operator of negation together with the *because*-clause is in the focus and as such triggers allegation rather than a presupposition.

(iii) Similar situation obtains in case of operators called in the recent linguistic literature focalizers; there belong such particles as *only, also, even*, etc. A detailed analysis of the meaning of constructions with such focalizers is given in [1]; in the present paper, we will reserve ourselves to one example:

(3) John only introduced Sue to BILL.

The meaning of the focalizer *only* indicates that an alternative is being chosen from a set of alternatives: the statement can be understood as being 'about' John's introducing Sue (topic), who could have chosen several people to whom he could introduce her; it is said that it was Bill (focus) and no other person, to whom Sue was introduced by John. The focus of the focalizer *only* prototypically coincides with the focus of the sentence.

It should be noticed in this connection that an important role is played by the position of the intonation center ; a change of the position of the intonation centre indicates a different TFA of the sentence:

(4) John only introduced SUE to Bill.

Now the statement can be understood as being 'about' John's introducing to Bill (topic), and it is said that it was Sue (focus) and no other person, who was introduced to Bill by John. Here again the focus of the focalizer *only* is equal to the focus of the sentence.

2.1.2 Paying due respect to TFA offers a good support for the assignment of reference, as illustrated by examples (5) and (6).

(5) a. The chair stood in front of a TABLE. This was old and shabby.
b. The chair stood in front of a TABLE. It was old and shabby.

(6) The chair stood in front of a TABLE. It was small, round, with three legs.

The strong pronoun *this* in the second sentence in (5)(a) refers to the item displaying the highest activation, i.e. *a table*; prototypically, it is the item constituting the focus proper and as such carrying the intonation center of the sentence. The reference by a strong pronoun in such cases is unambiguous, though it should be kept in mind that the reference to the item with the highest activation can be overshadowed by inferencing, based on world knowledge.With a weak pronoun in the second sentence in (5)(b), the preferable reference is to the subject of the preceding sentence, i.e. *the chair*. The reference by a weak pronoun is ambiguous, though there is a preferred reading keeping the syntactic symmetry (there is a tendency to preserve the subjects in the successive sentences, if possible).

2.2 Three values of the TFA attribute are distinguished:
(i) T (a non-contrastive contextually bound node, with a lower degree of communicative dynamism, CD, than its governor),
(ii) F (a contextually non-bound node, "new" piece of information),
(iii) C (a contrastive (part of) topic; in the present stage, this value is assigned only in cases in which the node concerned is in a non-projective position).

It is assumed that an F node is always more dynamic and a C node is less dynamic than a sister or parent T node.

The following examples illustrate these three values:

(7) (Nadpis: Volby v Izraeli.)
Po volbách(T) si Izraelci(T) zvykají(F) na nového(F) premiéra(F).
(Headline in the newspapers: Elections in Israel.)
After the elections(T), the Israelis(T) get used(F) to a new(F) Prime Minister(F).

(8) Sportovec(C) on(T) je(F) dobrý(F), ale jako politik(C) nevyniká(F).
Sportsman(C) he(T) is(F) good(F), but as a politician(C) he does not excel(F).

2.3 The instructions for the assignment of these values are formulated as follows:

2.3.1 In prototypical cases, i.e. cases of projective ATSs:

(i) left side dependents on the verb get T (except for cases in which this dependent would clearly carry the intonation center, IC),

(ii) the rightmost dependent of the verb gets F (under the assumption that it carries the IC; if the IC is placed more to the left, then every item dependent on the verb and placed after IC gets T),

(iii) the verb and those of its dependents that stand between the verb and the node assigned F and are ordered (without an intervening sister node) according to the systemic ordering (for Czech the systemic ordering (SO) of the main types of dependency is Actor < Temporal < Location < Instrument < Addressee < Patient < Effect; for the notion of SO see [3]), get F, unless they are repeated (perhaps coreferential, associated with or included in the meaning of their antecedent) from the previous sentence or context; the nodes between the verb and the node assigned F and the repeated nodes get T, as well as those placed more to the left than what would correspond to SO,

(iv) embedded attributes get F, unless they are only repeated or restored in the TGTS,

(v) indexical expressions such as 'já' [I], 'ty' [you], 'teď' [now], 'tady' [here], weak forms of pronouns, as well as pronominal expressions with a general meaning ('někdo' [somebody], 'jednou' [once upon a time] get T, except in clear cases of contrast or as bearers of IC,

(vi) strong forms of pronouns get the value F; after prepositions, the assignment of T or F these forms is guided by the general rules (i) - (iii),

(vii) restored nodes (i.e. those that are absent in ATSs but are added in the corresponding TGTSs) are always assigned T (and as such depend on their governors from the left.

2.3.2 An application of the above instructions leads to the following assignments of the values of the TFA attribute in sentences (9) through (14).

(9) Některé(T) ekologické(F) iniciativy(T) označily(F) informaci(F) o chystaném(F) teroristickém(F) útoku(F) za provokaci(F).
Some(T) ecological(F) initiatives(T) denoted(F) the information(F) about a prepared(F) terroristic(F) attack(F) as a provocation(F).

The node for *iniciativy* gets T according to (i), the node standing for *za provokaci* gets F according to (ii), the node for the verb *označily* and the node for *informaci*, which carries the functor Patient and as such stands in the hierarchy of systemic ordering before Effectum (i.e. the order is in accordance with the systemic ordering) get F according to (iii), and the nodes representing the attributes *některé, ekologické, chystaném, teroristickém, útoku* receive the value F according to (iv). According to the definition of topic and focus in the Functional Generative Description, this assignment of TFA values results in the following bipartition of the sentence into topic and focus:

(9') topic: některé ekologické iniciativy
focus: označily informaci o chystaném teroristickém útoku za provokaci

Even though we work with written texts, it is sometimes evident that the author of the text assumed the sentence to be 'read' with a non-prototypical placement of the intonation centre, see (10):

(10) (Většina ministrů Stěpašinovy nové vlády patří k věrným druhům nejznámějšího ruského intrikána Berezovského.) I Aksjoněnko(F) udržuje(T) s Berezovským(T) blízké(F) styky(T).
(The majority of the ministers of Stěpašinov's new government belongs to faithful friends of the best known Russian intriguer Berezovskij.) Even(F) Aksjoněnko(F) keeps(T) with Berezovskij(T) close(F) contacts(T).

The value (F) with the node for *Aksjoněnko* is assigned according to (i) because in the given context this word would be a bearer of the intonation centre; the node for *contacts* gets T inspite of the fact that *contacts* is the last word of the sentence; this is in accordance to the instruction (ii).
In a prototypical case, the embedded attributes are more dynamic than their head words and thus receive F; in specific cases of repetitions or restoration of the respective node (as in (11)) they get T (the restored nodes in (11') are enclosed in square brackets):

(11) (Tento týden se opět sešla poslanecká sněmovna.) Včera zasedaly parlamentní komise pro bezpečnost a pro zahraniční styky.
(This week again the parliament is in session.) Yesterday there was a meeting of the parliament committee for security and for international relations.

(11') Včera(T) zasedaly(F) parlamentní(F) komise(F) pro bezpečnost(F) a [parlamentní(T)] [komise(T)] pro zahraniční(F) styky(F).

The instructions (v) and (vi) hold for the nodes for *tady* (here) and for the strong form of pronoun *jeho* (him) in (12), respectively:

(12) (Pro českou hudbu je Charles Mackerras jedinečnou osobností.) Tady(T) je(F) doma(F), a proto si organizátoři(T) Pražského(F) jara(F) pro interpretaci(T) Smetany(F) vybrali(F) právě(F) jeho(F).
(For Czech music Charles Mackerras is an unequalled personality.) Here(T) he-is(F) at home(F), and therefore the organizers(T) of the Prague(F) Spring(F) for the interpretation(T) of-Smetana(F) have-chosen(F) just(F) him(F).

The value of the TFA attribute with nodes that are added in the TGTSs (i.e. those that are deleted in the ATSs and restored in TGTSs) is T; this concerns e.g. all nodes with the lexical value Gen(eral), as indicated in (13'), or contextually licensed deletions as in (14'):

(13) V Českém Krumlově byl zahájen kulturní program seznamující se středověkými zvyky.
Lit.: In Český Krumlov (there) was opened a cultural programme acquainting with medieval customs.

(13') V Českém(F) Krumlově(T) [Gen.Actor(T)] byl zahájen(F) kulturní(F) program(F) seznamující(F) se středověkými(F) zvyky(F).

(14) (Kam uprchlíci nejčastěji směřují?) Do Makedonie.
(Where the refugees most frequently head for?) To Macedonia.

(14') [uprchlíci(T)] [nejčastěji(T)] [směřují(T)] do Makedonie(F)

2.3.3 For non-projective ATSs specific rules are formulated; a node N dependent to the left in a way not meeting the condition of projectivity will be assigned C and will be placed more to the right, to meet that condition. The nodes depending on N (directly or indirectly) will move together with N and will get the value T or F according to 2.3.1 above. Thus, e.g. the sentence (15) will have a TGTS in (15'), in which *jásot* depends on *důvod*, has the index C and is placed to the right of the verb.

(15) K jásotu(C) není(F) nejmenší(F) důvod(F).
lit. For triumphing(C) is-not(F) the-least(F) reason(F)

(15') (neg.F) být.F (důvod.F (jásot.C) (nejmenší.F))
(neg.F) be.F (reason.F (triumphing.C) (least.F))

3 Attributes capturing coreferential and other links

In Sect. 2 we have presented an outline of a possibility how to capture in a annotated corpus (the PDT, in our concrete case) a fundamental property of the sentence structure that is related to the use of the sentence in a broader context, be it a suprasentential or a situational one, namely its topic-focus articulation. Another important property of sentences that links them to each other and to

the context of situation are the coreferential links. This issue actually reaches beyond the system of language, but we are believe that its treatment, even if a rather preliminary and tentative way, is a necessary ingredient of annotation schemata.
Let us illustrate the matter on ex. (16) with two successive sentences (a) and (b).

(16) a. Rakouská vláda se rozhodla, že bude vyvíjet tlak na Prahu ve věci stavby jaderné elektrárny v Temelíně.
The Austrian government decided that it will execute a pressure on Prague in the matter of building the nuclear power station in Temelin.

b. Rakouští představitelé dali jasně najevo, že otázku Temelína spojují s přijetím Česka do Unie.
Austrian representatives have made it clear that they connect the issue of Temelin with the acceptance of Czechia to the Union.

The expression *představitelé* (representatives) in (b) refers back to *vláda* (government) in (a); in other words, the expression *vláda* is an antecedent of *představitelé*. The relation between the two expressions can be captured by a special attribute attached to each expression, the value of which would be the lexical value of its antecedent. However, it can be easily shown that this is not enough, see a slight modification of (16) in (17):

(17) a. Rakouská vláda se rozhodla, že bude vyvíjet tlak na pražskou vládu ve věci stavby jaderné elektrárny v Temelíně.
The Austrian government decided that it will execute a pressure on Prague government in the matter of building the nuclear power station in Temelin.

b. Rakouští představitelé dali jasně najevo, že otázku Temelína spojují s přijetím eska do Unie.
Austrian representatives have made it clear that they connect the issue of Temelin with the acceptance of Czechia to the Union.

In (17)(b) the expression *představitelé* (representatives) again refers back to *vláda* (government) in (a), but to the first rather than to the most recent occurrence of this expression. Therefore, in addition to the attribute capturing the lexical value of the antecedent we need also to register which occurrence of the antecedent is referred to; this can be ensured e.g. by putting the serial number of the antecedent as the value of another attribute attached to each node.
Sentence (16) brings about a still another problem: in TGTSs, new nodes are added in case of deletion of elements in the surface shapes of sentences . Thus, in the TGTS of (16)(a) the node [on.Fem] has to be restored as the Actor of the second clause (a similar situation obtains for the Actor of the second clause in (16)(b)), as indicated in (16'):

(16') a. Rakouská vláda se rozhodla, že [on.ELID.Fem.Sg.Actor] bude vyvíjet tlak na pražskou vládu ve věci stavby jaderné elektrárny v Temelíně.

b. Rakouští představitelé dali jasně najevo, že otázku
Temelína [on.ELID.Anim.Pl.Actor] spojují s přijetím Česka do Unie.

To indicate whether the antecedent is in the same sentence or in the preceding context, we have added a third attribute, with the value 0 for the former case and the value PREV for the latter.

To sum up, three attributes are introduced in the TGTSs to account for the three ingredients sketched above: the attribute COREF with a value equal to the lexical value of the antecedent, the attribute CORNUM with a value equal to the (serial) number of the antecedent, and the attribute CORSTC with two values, namely PREV (obtained if the antecedent is in the previous sentence(s)) and 0 (in case the antecedent is in the same sentence). Thus, the node *představitelé* in (16)(b) and the two restored nodes in (16)(a) and (b) will have the following values in these three attributes:

(16") a. on.ELID.Fem.Sg.Actor: COREF [vláda]
CORNUM [2]
CORSTC [0]

b. představitelé: COREF [vláda]
CORNUM [2]
CORSTC [PREV]

on.ELID.Anim.Pl.Actor: COREF [představitel]
CORNUM [2]
CORSTC [0]

Anaphoric relations crossing sentence boundaries are captured only in the so-called 'exemplary' set of TGTSs and they will be registered in the further stages of the project, in which also the distribution of degrees of salience in the stock of shared knowledge will be taken due account of.

Acknowledgements Research for this paper was supported by a grant of the Czech Ministry of Education VS 96/151 and by Czech Grant Agency GAČR 405/96/K214.

References

1. Hajičová E., Partee B. H., Sgall P.: Topic-Focus Articulation, tripartite Structures, and Semantic Content. Dordrecht: Kluwer Academic Publishers (1998)
2. Sgall P.: Functional sentence perspective in a generative description. In: Prague Studies in Mathematical Linguistics 2, Prague: Academia (1967) 203-225
3. Sgall P., Hajičová E., Panevová J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht:Reidel (1986)

This article was processed using the LaTeX macro package with LLNCS style