

Tagging of very large corpora: Topic-Focus Articulation

Eva Buráňová and Eva Hajičová and Petr Sgall

Institut of Formal and Applied Linguistics,

Faculty of Mathematics and Physics

Charles University, Prague, Czech Republic

Abstract

After a brief characterization of the theory of the topic-focus articulation of the sentence (TFA), rules are formulated that determine the assignment of appropriate values of the TFA attribute in the process of syntactico-semantic tagging of a very large corpus of Czech.

1 Introduction: The Prague Dependency Treebank (PDT)

PDT is a corpus (a part from the Czech National Corpus), tagged on the following levels:

1. morphemic (POS and annotations using a very large number of tags, as required by the language with rich inflection; cf. (Hajič and Illadká, 1997));
2. ‘analytic’ (dependency syntax, with nodes for all word occurrences, also for punctuation marks etc., and with the tags for morphemic units and for basic kinds of surface syntactic relations (Subject, Object, Adverbial, Adjunct), cf. (Hajič,))
3. tectogrammatical (underlying) syntax, with a much more detailed classification of syntactic relations and with nodes for autosemantic lexical occurrences only (rather than function words), with indices corresponding to the syntactic relations, such as Actor, Addressee, Objective (Patient), Locative, Manner, Means, etc., and to morphological values such as Preterite (Anterior), Conditional, Plural, etc., and also as the prototypical values of ‘in’, ‘into’, ‘on’, ‘from’, etc.; correlates of functional words (and morphemes) on this level have the form of indices of lexical node labels.¹

¹An exception concerns coordinating conjunctions, which, in PDT, are treated as head nodes of the co-

2 Representing Topic-Focus Articulation (TFA) in TGTSs

2.1 A brief characterization of TFA

The tectogrammatical tree structures (TGTSs) should capture not only the syntactic (dependency) relations, but also the TFA of the utterances in the corpus, since TFA is expressed by grammatical means and is relevant for the meaning of the sentence (even for its truth conditions), i.e. it constitutes one of the basic aspects of underlying structures. The semantic relevance of TFA can be illustrated by examples such as (1), which is a translation of the Czech ex. (1') (the capitals denote the placement of the intonation centre, i.e. the focus proper):²

- (1) (a) *English is spoken in the SHETLANDS.*
(b) *In the Shetlands, ENGLISH is spoken.*
(1') (a) *Anglicky se mluví na Shetlandských OSTROVECH.*

ordinted groups. This makes it possible to represent the tectogrammatical structures of all sentences as trees (rather than using more-dimensional networks); in this point, PDT differs from the theoretical assumptions of the Pragmian Functional Generative Description (now discussed in (Hajičová et al., 1998)).

²In the prototypical case the intonation centre is characterized by falling (or rising-falling) stress, but there are also cases in which (similarly as in questions, to a certain degree) the centre has a rising stress. This concerns utterances displaying a feature of hesitation or incompleteness, cf. (M.,); often also with greetings (such as Czech *Dobré jitro* [Good morning]) a difference of this kind marks the ‘starting’ token, connected with the expectation of an answering token, which exhibits a falling stress. Although in a sentence containing occurrences of both a rising and a falling stress the former expresses a contrastive (part of) topic, we prefer to analyze it as the focus in a sentence without an occurrence of the latter; in such a position, the rising stress regularly is carried by an item referring to ‘new’ information. In written texts, some occurrences of the rising stress are marked by a semicolon or by ‘...’.

- (b) *Na Shetlandských ostrovech se mluví ANGLICKY.*

The communicative function of the sentence can basically be rendered by understanding its topic (T) as ‘what is the sentence about’, and its focus (F) as the information that is asserted about the topic, i.e., schematically, the interpretation of the sentence S can be understood as

$$S = F(T)$$

Thus, (1)(a) asserts, on its preferred reading (with just the locative modification constituting its focus) about where English is spoken that it is in the Shetlands, which hardly can be accepted as true w.r.t. what we know of the actual world, if no specific context is present. (1)(b) is understood as true, stating about E. that it is spoken in the S.

In the TGTSs the order of nodes is such that all parts of T precede all parts of F. Moreover, the order of nodes corresponds to the scale of communicative dynamism (CD, see Section 3 below); a less dynamic node prototypically has the broader scope than a more dynamic one (if the nodes correspond to operators). F proper is then the most dynamic (the rightmost) node.

TFA is relevant also for the semantics of negation:

- (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).

With the paraphrase (a), the negated verb ‘come’ is included in T, i.e. the fact that John's being ill is the cause of an event is asserted about the event that he did not come. With (b), the main verb ‘come’ also belongs to T, but what is negated, is the relation between T and F: John came, but what is asserted about his coming is that the cause of this event was not his illness (he might have been ill, though).

Every node in a TGTS is either contextually bound (CB) or non-bound (NB); this opposition is a linguistic counterpart of the cognitive

dichotomy of ‘given’ vs. ‘new’, where also an item, if corresponding to a ‘given’ referent presented as occupying a newly characterized specific position (often in relation to one or more ‘given’ items), has the feature NB, cf.:

- (3) *Give this to YOUR mother. (My parents don't like such gifts.)*
- (4) *(Mary knows both Peter and Jane.) However, this time she only invited HER.*

The indexical pronoun ‘your’ in (3) and the anaphoric pronoun ‘her’ in (4) can only refer to items that in a sense are ‘known’ in the given situation. However, in these examples, both of them occur as NB; their stress indicates their function as F proper of the respective sentence.

Prototypically, an NB node belongs to F and a CB node is in T; however, a node not dependent immediately on a finite verb (esp. an adjunct) need not meet this condition. Thus, in (5), ‘my’ as a shifter, directly determined by the conditions of the discourse, is CB, although belonging to F, since it depends on a part of F (see (Hajičová et al., 1998) for a definition of T and F on the basis of contextual boundness and of syntactic dependency, as well as for other details of the given descriptive framework).

2.2 The attribute TFA in PDT

Three values of the attribute TFA are distinguished with every node in a TGTS:

1. T a non-contrastive CB node, which always has a lower degree of CD than its governor, if any;
2. F an NB node (if different from the main verb, then following after its head word in the TGTS)
3. C a contrastive CB node

Examples:

- (5) *(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).

(6) *Sportovec(C) on(T) je(F) dobrý(F), ale jako politik(C) nevyniká(F)*.

(As a) Sportsman(C) he(T) is(F) good(F), but as a politician(C) he does not excel(F).

The instructions for the assignment of the values of TFA can be briefly specified as follows, if the surface word order and the position of the intonation center (IC, see footnote 2 above) is taken into account, as well as the ‘systemic’ (canonical) ordering of the kinds of dependents (which, in fact, can differ with different head words; SO is specified either in the valency frames in the individual lexical entries, or, if possible, for whole lexical classes and subclasses):

1. the bearer of IC \implies F typically = the rightmost dependent of the verb
2. if the IC is placed on a node other than the rightmost one, the complementations placed after IC \implies T
3. a left side dependent of the verb \implies T or C, except for cases in which it clearly carries IC
4. the verb and those of its dependents that stand between the verb and the F-node (see 1) and that are ordered (without an intervening sister node) according to SO \implies F; among sister nodes, all those carrying T follow after all those with C, and all those carrying F follow after all those with T; there are two sets of exceptions:
 - (a) a focus sensitive particle can carry F even when preceding its governing node that carries C, cf. Section 3.2 below
 - (b) a node M carrying T or C can follow after its mother node if a node with F is present among the nodes subordinate to M, but is absent both among the sisters of M and among its superordinate nodes (here the relation of ‘superordinate’ and ‘subordinate’ is the transitive closure of ‘governing’ and ‘dependent’); cf. the notion of ‘proxy focus’, characterized in (Hajičová et al., 1998), and examples such as (*Kterého učitele jsi tam*

viděl?) Viděl jsem tam učitele chemie [lit. (Which teacher.Accus have-you there seen?) I saw there (the) teacher of-chemistry], with which the Patient *učitele* follows after the verb in the underlying tree, although it carries T

Note: For Czech, the SO of the main types of dependency has been found (on the basis of empirical analysis of texts and of experiments with groups of speakers, see (Sgall et al., 1995)) to have (with most verbs and other heads) the following form, as for the main kinds of dependents:

Actor < Temporal < Location < Instrument < Addressee < Patient < Effect³

5. embedded attributes \implies F (unless they are only repeated or restored)
6. indexical expressions (*já* [I], *ty* [you], *teď* [now], *tady* [here], weak forms of pronouns, pronominal expressions with a general meaning (*někdo* [somebody], *jednou* [once upon a time]...) \implies T (except in cases of contrast or as bearers of IC)
7. strong forms of pronouns \implies F (after prepositions and in coordinated constructions, the assignment of T or F in Czech is guided by the general rules 1 through 4)
8. restored nodes, deleted in the surface forms of sentences \implies T; we devote Section 2.3 below to the placement of the restored nodes Note: There are special cases of coordination, both in Czech and in English, which do not meet this condition: e.g. in ‘They drank white and red wine’ the first occurrence of ‘wine’, which may be NB, is deleted in the surface (and restored in the TGTS).
9. a node N dependent to the left in a way not meeting the condition of projectivity: \implies C (this node is then placed more to the right, to meet that condition; these and

³Let us note that Directional.3 (‘where to’) follows after Patient in Czech as well as in English and also in German, according to the empirical research discussed in (M.,); thus it is not exact to characterize the canonical order of German as a “mirror image” of that of English.

other movements are discussed in Section 2.4 below)

10. the nodes subordinate to such an N move together with it and get T or F (according to the rules above)

Note: The resulting TGTSs are projective, i.e. for every pair of nodes x , y in a TGTS it holds that if x depends on y and x follows (precedes) y , then every node z following (preceding) y and preceding (following) x is subordinate to y . Thus, ‘not to meet the condition of projectivity’ concerns the ‘analytic’ trees; this means, in other words, that this condition would not be met if the positions of x and y in the left-to-right order of the nodes in the TGTS (in the ‘underlying word order’) always corresponded to their positions in the surface (morphemic and ‘analytic’) word order.

Example (with a very simplified linearized notation of the TGTS, in which every dependent is closed in its pair of parentheses):

- (7) *K jáсотu(C) není(F) nejmenší(F)*
 For triumphing is-not the-least
důvod(F).
 reason
- (7') (neg.F) *být.F* ((*jáсот.C* *důvod.F*
 (neg.F) bc.F ((triumphing.C) reason.F
 (*nejmenší.F*))
 (least.F))

A sentence with a non-prototypical placement of the IC:

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

2.3 The position of a restored node

The degree of CD of a node that is being restored (i.e. supposed to have been deleted in the surface form of the sentence), and thus also its position in the underlying word order, is determined on the basis of its relationship to its governing node. Since such a node almost always is contextually bound (with the exception of the specific case of coordinated structures, see the Note after point 8 in Section 2.2 above), it is placed to the left of its governing word; more specifically:

- (a) if the restored node RN depends on a verb, then:
- (aa) if RN is not the single item depending on the given verb token, then RN is to be added in the ‘Wackernagel position’;
- (ab) if RN has no sister nodes, then it is placed at the beginning of the clause;
- (b) if RN is restored as depending on a noun (or adjective), RN is placed as the least dynamic dependent of this governing word;
- (c) if more than one node are inserted as depending on one and the same item, then their order should conform to the systemic (‘canonical’) ordering of the valency slots (see the remark on SO in Section 2.2 above, point 4).

Point (a) appears to be substantiated by the fact that e.g. the subject pronoun appears in the zero form in Czech under similar conditions as the weak, clitic pronouns, for which the position immediately to the left of the verb is typical, cf. sentences such as *Včera (on) přišel pozdě* [Yesterday (he) came here late], *Janu (oni) neviděli* [lit.: Jane-Accus they have-not-seen], or *(On) spal* [He was-sleeping]. This concerns also such deletable items as e.g. the Directional with *příjet* [arrive], cf. *Jan dnes (sem/tam) nepřijel* [lit. John to-day (here/there) has-not-arrived].

The appropriateness of these preliminary rules is being checked during the tagging procedure, the results of which will be of importance for a more exact (and more complete) formulation of the relevant parts of the description of the sentence structure of Czech. This aspect

of the usefulness of the corpus tagging concerns also many other points of grammar.

2.4 Underlying and surface word order

Within the tagging procedure, the differences between the two levels of the left-to-right order can be described by movement rules, a preliminary form of which can be briefly characterized as follows:

1. if a node M1 carries C and a node M2 depending on M1 is placed to the right of a node M3 superordinate to M1 in the surface word order, then M1 is placed immediately to the left of M2 in the resulting tree; cf. e.g. *Sportovec* (M1) *on je* (M3) *dobrý* (M2) [lit. (As a) sportsman he is good], see ex. (6) in Section 2.2
2. if the positions of the nodes M1, M2 and M3 differ from point 1 only in that M1 depends on M2, then again M1 is placed immediately to the left of M2 in the resulting tree; cf. example (7) in Section 2.2 above, in which *jásot* occupies the position of M1, *důvod* that of M2, and *není* that of M3, or:

- (9) *Jirku* (M1) *jsme plánovali*(M3)
poslat(M2) *do Francie*
 [lit. George.Accus (M1) we-planned
 (M3) to-send (M2) to France]

3. a comparative of an adjective that precedes its governing noun in the surface is moved to the right of this noun in examples such as *větší město než Boston* [a larger town than Boston]; this surface order probably should be limited (by a rule of grammar) to cases in which the two nouns belong to a single semantic subclass.
4. in sentences exhibiting a secondary placement of IC, the bearer of IC occupies the rightmost position in the resulting tree; cf. example (1)(b) in Section 2.1 above, in which 'English' is the focus proper; the assumption underlying the placement of IC in a written text is that a written form of a sentence may correspond to different (spoken) sentences, according to the differences of the placement of IC in the appropriate way of pronouncing the sentence.

3 The special case of focus sensitive particles

Since the focus sensitive particles are identified (by the functor value RHEM for 'rhematizer' or 'focalizer'), it is possible to use PDT also for a specification of their occurrences in different positions both in the dependency structure of the sentence and in its TFA. The starting hypotheses, which might be checked on the basis of PDT, are as follows (cf. (Hajčková et al., 1998)):

3.1 Focus sensitive particles in prototypical positions

The prototypical syntactic position of a focalizer can be understood as that of a dependent of a verb node; thus, in examples like (10) or (11), it is possible to specify the scope of the focalizer as the whole subtree subordinated to the verb (where 'subordinated' is understood as the transitive closure of 'dependent' in the reflexive sense, so that the verb itself is included); the scope is divided into background and focus of the focalizer (ff), as will be specified in 3.2. Thus, in the interpretation of (10) on the reading represented (with many simplifications) by (10') it is included that (according to what P. knows) among those whom T. saw there was noone else than M (i.e. while 'T. saw' constitutes the background of 'only', its ff is 'Mary'). Similarly, if in (11) the negation (although expressed by a prefix in Czech) is handled as a dependent of the verb, its background is the subject and ff includes both the verb and the object.

- (10) *Pavel ví, že Tomáš*
 'Paul knows that Thomas
viděl jen MARIÍ.
 saw only MARY.'

- (10') (Paul) knows ((Thomas) saw (only)
 (Mary))

- (11) *Martin nečte NOVINY.*
 'Martin not-reads NEWSPAPERS.'

In (12) only the adjective constitutes the ff of 'only', its background consisting of 'car' (among all cars, P. only wants a blue one); thus, the focalizer can best be described here as depending on 'car'.

- (12) *Petr chce jen MODRÉ auto.*
 'Petr wants only (a) BLUE car.'

3.2 Focus sensitive particles in the hierarchy of communicative dynamism

The primary position of a focalizer in a TR is at the boundary between the topic and the focus of the verb clause and the focus of the clause is then identical to the focus of the focalizer. If a focalizer is included in the topic, then its focus contains those items which in the TR are placed between this focalizer and the next item marked as C to the right and are more dynamic than the focalizer).

It should be noted that CD is understood here as a partial ordering defined so that:

- (i) in every set of a head and its daughter nodes, every daughter node placed to the right of its head is more dynamic than every daughter node placed to the left of its head;
- (ii) the relation ‘more dynamic’ is determined by the irreflexive transitive closure of (i).

Thus, e.g. in the TR (10’), ‘knows’ is more dynamic than ‘Paul’ and less dynamic than ‘saw’ according to the point (i), and both ‘only’ and ‘Mary’, being more dynamic than ‘saw’, are more dynamic than ‘knows’ according to the point (ii); however, ‘Thomas’ is neither more nor less dynamic than ‘knows’. If (10) is embedded into a more complex sentence as (a part of) its topic, then ‘Mary’ is more dynamic than ‘only’ and has the feature C; thus, e.g. with ‘Since Paul knows that Thomas saw only Mary, he is not afraid’, ‘Mary’ constitutes the whole of ‘only’, similarly as in (10’).

The underlying word order *W* (a linear ordering) is then defined on the basis of CD, with (iii) and (iv) holding for every two nodes *x* and *y* in a tree:

- (iii) if node *x* is more dynamic than node *y*, then *x* follows *y* under *W*;
- (iv) if node *x* follows node *y* under *W*, node *u* is subordinated to *x* and node *z* is subordinate to *y*, then *u* follows both *y* and *z*, and *x* follows *z* under *W*.

Among the non-prototypical, secondary positions of focalizers, there are also the cases of their clustering (e.g. ‘not only’), as well as the

sentences in which a focalizer itself constitutes the whole focus of the sentence (‘He DID realize this’).

4 Summary

After a brief characterization of the Prague Dependency Treebank and of the Praguian theory of Topic-Focus Articulation we have presented a proposal how the main aspects of the information structure of the sentence (i.e. of its topic-focus articulation) can be integrated into the tagging system that captures the underlying structures. The present form of the system makes it possible to check our hypotheses on a large text corpus, and thus perhaps to achieve a higher degree of automation (and reliability) of the proposed procedure. The last section exemplifies how the proposed approach makes it possible to analyze structures with the so-called focus sensitive operators.

References

- Jan Hajič. Building a syntactically annotated corpus: The Prague dependency treebank. In E. Hajičová, editor, *Issues of Valency and Meaning*, Studies in Honour of Jarmila Panevová, pages 106–132. Karolinum, Prague.
- Jan Hajič and Barbora Hladká. 1997. Probabilistic and rule-based tagger of an inflective language - a comparison. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 111–118, Washington, D.C.
- Eva Hajičová, B. Partee, and Petr Sgall. 1998. *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer, Amsterdam.
- Steedman M. Information structure and the syntax-phonology interface. unpublished manuscript.
- Petr Sgall, O. Pfeiffer, W. U. Dressler, and M. Půček. 1995. Experimental research on systemic ordering. *Theoretical Linguistics*.