# Topic-Focus Articulation in Corpus Annotation

**Petr Sgall**
sgall@ufal.mff.cuni.cz

The research group of theoretical and computational linguistics at Charles University, Prague, which owes so much to the interest and support of Walther von Hahn in the difficult decades, now works on the Prague Dependency Treebank (PDT), a collection of annotated sentences from the Czech National Corpus, based on the descriptive framework of Functional Generative Description (see Sgall et al. 1986). Up to now, about 20 000 sentences have been annotated at the level of (underlying) syntax, out of which 2000 have been analyzed also as for their Topic-Focus articulation (TFA). On the background of dependency based syntax, TFA has been understood to constitute one of the main aspects of the underlying structure, analyzed already by Weil (1844), later by G. von der Gabelentz, P. Wegener, V. Mathesius and others; now see Hajičová, Partee and Sgall (1998), where also issues of a formal treatment of the interpretation of this articulation, based on the 'aboutness' relation, are discussed. As reflecting the ‘given – new’ strategy in discourse, TFA has been considered to belong to the main objects of linguistic study. The explicit descriptive framework allows to describe the TFA not only as concerning the intrinsic dynamics of the process of communication, patterned in the utterance (sentence occurrence), but also as constituting the structure of the sentence itself.

TFA is semantically relevant, as the following examples show:

(1) a. I work on my dissertation on Sundays.
b. On Sundays, I work on my dissertation.
(2) a. We went by car to a lake.
b. We went to a lake by car.
(3) a. They moved from Chicago to Boston.
b. They moved to Boston from Chicago.

In its preferred reading (and with the normal intonation, i.e., with the intonation center at the end of the sentence), sentence (1) a. asserts about the speaker‘s work on her/his dissertation that this takes place on Sundays, while (1) b. asserts about Sundays that the speaker spends them working at her/his dissertation. In (2) a. there are (at least) two possibilities: either it is asserted about a group including the speaker that they went by car to a lake, or it is asserted about their trip by car that its goal was a lake. In (2) b. *to a lake* is included in the Topic on all readings, and it is asserted that the trip was made by car. In a similar way, also with (3) a. two readings are present, which is not the case with (3) b.

The assignment of the TFA features is based on operational criteria such as the question test, according to which e.g. in (3) b. *from Chicago* is understood as the Focus, since this part of the sentence is the counterpart of the interrogative element in the question (3) c., which can be answered by (3) b. The rest of the sentence, the content of which is “known“ from the question, is its Topic.

(3) c. From where did they move to Boston?

The Functional Generative Description has been elaborated as a formal framework in which the syntactic tectogrammatical representations (TRs) are viewed

as the interface level of the language system and the layers of cognition (in which also the specification of reference, the inferencing based on contextual and other knowledge and a truth-conditional or other basis of semantics are relevant, cf. Sgall 1994). The TRs contrast with the morphemic („surface“) representations, i.e. strings of closely and loosely connected morphemes, which are directly expressed by phonemic strings.

The primary shape of the TR (in which no coordination constructions occur) is a dependency tree, with its root labelled by the underlying counterpart of the verb, which occupies the position of PRED(icate) and displays in its valency frame the functors, characterizing types of its dependents, i.e. arguments and adjuncts (either of which can be obligatory or optional with the given head). A formal specification of the TRs can be found in Plátek et al. (1984) and in Petkevič (1995). A discussion of the computational treatment of TRs can be found in Sgall and Böhmová (in prep.).

The surface (morphemic) word order corresponds, in the unmarked case, to the left-to-right order of the nodes in the TRs, i.e. to the scale of Communicative Dynamism (determined by 'systemic ordering' in the Focus, see Sgall et al. 1995), in which Topic precedes Focus, or, more exactly, the contextually bound (CB) nodes precede their non-bound sister nodes and heads.

The annotations of the selected text are carried out in four separate steps. The first three steps have been automated to a high degree, using (i) a morphemic analyzer, which yields all possible values of the word forms present in the outer form of the sentence, (ii) a morphemic tagger, which chooses one of the values (Hajič and Hladká 1997), (iii) the 'analytical level', which has been developed as a technical device that has no immediate theoretical significance, but constitutes the first stage of syntactic annotations, bridging the gap between the linear sentence representation and the underlying dependency tree. In the analytical tree structure (ATS) every word of the sentence including the punctuation marks is represented by a single node. The output of Collins' dependency parser, which yields the ATSs, is manually corrected by human annotators (Hajič 1998); approx. 100 000 Czech sentences have been annotated by a semi-automatic procedure; the resulting ATSs can be schematically characterized as the dependency tree in Fig. 1, a simplified ATS of the Czech equivalent of sentence (1) b., i.e.: *O nedělích pracuju na své disertaci.*

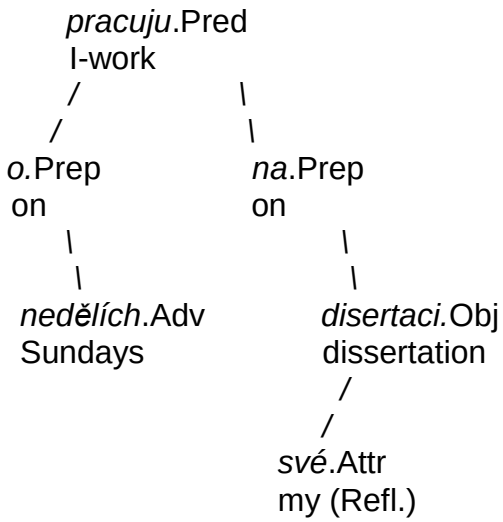


Fig. 1.

The final step is the annotation on the underlying level, on which only the autosemantic words consitute nodes of the dependency tree and the condition of projectivity is met, i.e., no crossing of edges is allowed; the underlying word order often differs from the surface order (e.g. in the order of an adjective and its head noun). All the auxiliary words and punctuation marks are captured as indices of the nodes (grammatemes). The relations between the nodes are marked with a fine-grained set of functors, and nodes are added in case of deletions in the surface form of sentences. The trees are handled in the shape of tectogrammatical tree structures (TGTSs), which differ from the theoretically postulated TRs in that they contain specific nodes for coordinating conjunctions, instead of displaying more than two dimensions. A simplified TGTS of (1) b. is given in Fig. 2.

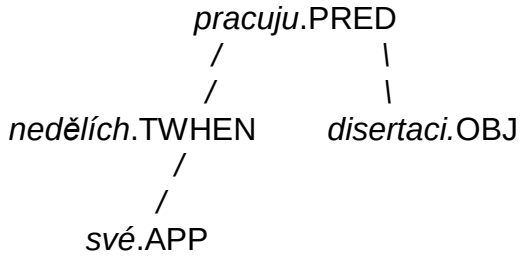


Fig. 2.

The procedure of transition from ATS to TGTS is partly automated, and the result of the automatic procedure is manually edited by humans. Along with the automatic treatment of large sets of prototypical phenomena, another set of automatic steps has been prepared, which completes some of the manual operations in cases in which it has not been difficult to formulate general rules. The TGTSs include an indication of the position of every node in the topic-focus articulation (TFA) with respect to the scale of Communicative Dynamism, represented as the left-to-right order of the nodes. Note that the left-to-right order of coordinated nodes in the TGTSs does not reflect Communicative Dynamism.

Every lexical (autosemantic) occurrence is assigned one of three values of a specific TFA attribute:

- t for 'contextually bound', CB (prototypically in Topic, T),
- c for 'contrastive (part of) Topic',
- f (‘non-bound’, NB, typically in Focus, F)

Sentence (4) is a typical example, known from older discussions (with *he* bearing a rising contrastive stress and *her* carrying the typical sentence final falling stress):

(4) (She called him a republican.) Then.t he.c insulted.f her.f.

In unmarked cases, the main verb (V) and its direct dependents following it belong to Focus, they carry index f; the items preceding V carry t or c. In marked cases, the verb can be CB, i.e. in the Topic, or the Focus may precede the verb; usually the intonation centre (sentence stress) then marks the Focus, occupying a marked

position. Dependents of nouns primarily are NB, i.e. they carry index f, even if they belong to the Topic of the sentence together with their head noun. In the underlying order, NB dependents follow and CB dependents precede their heads.

Let us characterize the description of TFA in PDT by a sample of sentences contained there, to illustrate how this approach makes it possible to analyze also sentences with neither Topic nor Focus corresponding to a single constitutent in a phrase-structure based description. In (5'), which is a highly simplified linearized TGTS of (5), every dependent item is enclosed in a pair of parentheses; syntactic subscripts of the parentheses are left out here, for the sake of transparency, as well as subscripts indicating morphological values, with the exception of the two which correspond to function words; Fig. 3 presents the respective tree structure, in which three parts of each node label are specified, namely the lexical value, the syntactic function (with ACT for Actor/Bearer, RSTR for Restrictive, MANN for Manner, and OBJ for Objective), and the TFA value.

(5) České radiokomunikace musí v tomto roce rychle splatit dluh televizním divákům.
lit.: Czech Radiocommunications have in this year quickly to-pay debt (to) TV viewers.
E.: This year, Czech Radiocommunications have to quickly pay their debt to the TV viewers.

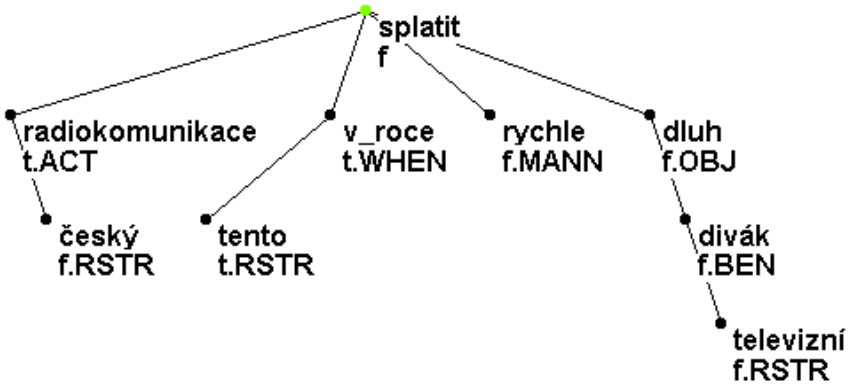


**Fig. 3.**

The (highly simplified) linearized form of the TGTS:

(5') (radiokomunikace.t (české.f)) ((tomto.Temp.t) rok.t) splatit.Necess.f (rychle.f) (dluh.f (divákům.f (televizním.f)))

The next example contains a focus sensitive particle in the primary prototypical position:

(6) Pražská matějská pouť má již čtyřsetletou tradici.
(The) Prague Matthew Fair has already (a) 400-year tradition.
The Prague St. Matthew Fair has already a tradition of 400 years.

(6') (pouť.t (pražská.f) (matějská.f)) má.t (již.f) (tradici.f (čtyřsetletou.f))

An example of the marked presence of CB items (contrastive or not) within Focus:

(7) Přiznám se, že já osobně to dost prožívám.
Lit.: I-admit Refl that I personally it quite live-through.
I admit that I personally live this through quite intensively.

(7') (já.t) (Gen.t) přiznám-se.f ((já.c (osobně.f)) (to.t) prožívám.f (dost.f))

In the TGTS (7') the deleted subject pronoun has been restored; it is CB and belongs to the Topic (the values of its grammatemes are expressed, on z morphemic level, by the agreeing personal ending of the verb). Another node has been added for the General Addressee of *přiznám se* 'I admit'. The main verb together with the embedded clause constitute the Focus, within which the two verbs are NB, as well as the adverb *dost***.** The subject of this clause, expressed by the pronoun in its strong form, is a contrastive CB item, and together with the CB pronoun *to* 'it' it belongs to the Focus, since both the pronouns depend on an item in Focus different from the main verb (namely to the embedded verb). The NB adverb *osobně* 'personally' is understood in PDT to depend on *já* ('I'). It is a general rule in Czech that the weak pronominal forms (such as *ho* 'him.Accus.', *mu* 'him.Dat', *tě* 'you.Dat', *ti* 'you.Accus.', or the zero form of the Nominative, 'pro-drop') always are CB.

The order of items within Topic can be illustrated by (8):

(8) Dnes už si však bez něho svoji práci nedovedou představit.
Lit.: Today already Refl however without him their-Refl work they-cannot imagine.
Nowadays, however, they cannot IMAGINE their work without him.

(8') (dnes.t) (však.t) (oni.t) (bez-něho.t) ((svoji.t) práci.t) (už.f) (Neg.f) představit-si.Possib.f

In Czech, the word order is "free" enough (i.e., is flexible enough to reflect the scale of communicative dynamism, the underlying word order, without many movement rules) to be understood as the main means expressing the underlying order of the items within the Topic of a sentence. If, following V. Mathesius, we speak of 'Topic proper' and 'Focus proper' as the two extreme parts of the sentence (i.e., of its underlying representation), with other parts of Topic and Focus occupying intermediate positions, we may see Topic proper as the least dynamic part of the sentence (referring to "what the sentence is about", and Focus proper as the most dynamic one. In (8), then, we would say that *dnes* 'today' is the Topic proper, with the zero subject (Actor, the strong form of which is *oni* 'they'), the group *bez něho* 'without him', and the object *svoji práci* 'their-Refl work' all occurring as "accompanying members of the Topic". The specific positions of *už* 'already', *si* (a reflexive particle lexically belonging to the verb) and *však* 'however' are determined by the character of these words as clitics. The operator of negation, which is one of the focus sensitive operators, has the form of the verb prefix *ne-* in Czech.

As illustrated with the examples (5) – (8), a one-to-one linearization of the dependency tree is possibile, having the form of a well parenthesized string of complex symbols. This possibility is of fundamental importance. On the one hand, it may be maintained that a relatively natural image of the sentence structure, as internalized by speakers, comes close to the pattern based on rooted trees; in fact, sentence structure is more complex, since the combinations of the relations of dependency and of coordination require more dimensions than the two that are proper to the dependency tree. On the other hand, the strong restrictions of 'projectivity' (with no two edges crossing each other) and of a similarly limited repertoire of relationships between dependency and coordination (as well as apposition or parenthesis) allow for such a linearization, the parenthesized strings of which come close to proposition calculus. This points to the possibility of describing the core of sentence structure (without non-prototypical features and subsystems such as coordination, secondary positions of focus sensitive operators, movements concerning *wh*- items, irregularities of morphemics) as not substantially surpassing what often is understood by logicians as common human mental abilities. Thus, it appears worth a further discussion whether also the internalization of the core of the mother tongue could not be explained on the basis of such common abilities, without postulating a complex framework of innate features.

**References***:*


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, 106-132. Prague: Karolinum.

Hajič J. and B. Hladká (1997) Probabilistic and Rule-Based Tagger of an Inflective Language - A Comparison. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., 111-118.

Hajičová E., Partee B. H. and P. Sgall (1998) *Topic-focus articulation, tripartite structures, and semantic content*. Dordrecht:Kluwer.

Petkevič V. (1995) A new formal specification of underlying structures. *Theoretical Linguistics* 21:7-61.

Plátek M., Sgall J. and P. Sgall (1984) A dependency base for a linguistic description. In Sgall P., ed.: *Contributions to functional syntax, semantics and language comprehension.* Amsterdam: Benjamins – Prague: Academia, 1984, 63-97.

Sgall P. (1994) Meaning, reference and discourse patterns. In: Ph. Luelsdorff (ed.). *The Prague School of Structural and Functional Linguistics.* Amsterdam/Philadelphia: J. Benjamins, 277-309.

Sgall P. and A. Böhmová (in prep.) The simple core and the complex periphery of natural language. Submitted for COLING 2002.

Sgall P., Hajičová E. and J. Panevová (1986) *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht:Reidel - Prague: Academia.

Sgall P., O. Pfeiffer, W. U. Dressler and M. Půček (1995) Experimental research on Systemic Ordering. *Theoretical Linguistics* 21:197-239.

Weil H. (1844) *De l'ordre des mots dans les langues anciennes comparées aux langues modernes*. Paris. Translated as *The order of words in the ancient languages compared with that of the modern languages*. Boston 1887, reedited in Amsterdam:Benjamins 1978.