Topic-Focus Articulation and degrees of salience in the Prague Dependency Treebank

Petr Sgall, Eva Hajičová and Eva Buráňová
{sgall,hajicova,buranova}@ufal.mff.cuni.cz

Since Eloise Jelinek has been interested in the issues of negation, focus and information structure, to the research of which she has contributed substantially, we want to use this nice occasion and present here partial results of an analysis of the Topic-Focus articulation (TFA) of Czech and of the impact of these results on inquiries into coreferrence in coherent discourse.

In Czech linguistics, TFA has been systematically explored thanks to the classical Prague School of functional and structural linguistics. As reflecting the 'given – new' strategy in discourse, TFA has been considered to belong to the main objects of linguistic study. Continuing the results gained by V. Mathesius, J. Firbas and others since the 1920s, the explicit linguistic descriptive framework characterized in Sgall et al. (1986), Hajičová (1993), Hajičová E., Partee B. and P. Sgall (1998) includes a possibility to describe 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, i.e. grammar. Within this framework, TFA is understood as one of the basic aspects of (underlying) sentence structure, which characterizes the sentence as a unit of the interactive system of language; TFA thus is seen as a manifestation of the sentence being anchored in the context.

To put it quite briefly, we may characterize the Praguian framework as based on the relation of syntactic dependency; the framework does not work with the concept of 'constituent'. This makes it easier to account for the fact that, as our examples below document, most different combinations of sentence parts may belong either to Topic or to Focus. In the underlying representations of sentences, which prorotypically are dependency trees, the left-to-right order of nodes, i.e. the underlying word order (the scale of 'communicative dynamism') starts with Topic proper and proceeds to Focus proper (the most dynamic part of the sentence). The interplay of word order and of specific features of sentence prosody corresponds to the underlying word order as its expression means.

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.

The semantic basis of TFA may be seen in the relation of aboutness: a prototypical declarative sentence asserts that its Focus holds about its Topic. The Topic primarily consists of 'contextually bound' items (referring to what presystemically is called „given information", and

the Focus contains ‘contextually non-bound’ („new“) elements.

After having been discussed and theoretically elaborated for several decades, these issues now are analyzed computationally in the context of the Prague Dependency Treebank (PDT), based on Czech National Corpus (CNC, containing hundreds of millions of word occurrences in journalistic fiction and other texts). The PDT scenario, which thus serves for checking and enriching the chosen description, comprises three layers of annotation:

(i) the morphemic (POS) layer with about 2000 tags for the highly inflectional Czech language, assigned by a stochastic automatic tagger (Hajič and Hladká 1997, Böhmová and Hajičová 1999), with a success rate of more than 95%;

(ii) a layer of 'analytic' ("surface") syntax (Hajič 1998): cca 100 000 Czech sentences have been annotated;

(iii) the underlying syntactic level: tectogrammatical tree structures (TGTSs): up to now, in an experimental phase, running texts of 200 sentences each have been annotated in full detail (the so-called ‚model collection‘), and 2000 sentences in what concerns syntactic structure itself (‘large collection‘).

TGTSs are much simpler than constituency based structure layers of annotation. They do not contain any nonterminal symbols; each of their nodes is labelled by a complex symbol composed of a lexical and a morphological part (values of morphological categories such as number, tense, modalities), and each edge is labelled by the symbol indicating a syntactic relation (i.e. the type of the dependent node, see point (d) below). The main properties of TGTSs can be summarized as follows:

(a) only autosemantic (lexical) words have nodes of their own; function words, as far as semantically relevant, are reflected by parts of complex node labels (with the exception of coordinating conjunctions);

(b) nodes are added in case of deletions on the surface level;

(c) the condition of projectivity is met (i.e. no crossing of edges is allowed)

(d) tectogrammatical functions ('functors') such as Actor/Bearer, Patient, Addressee, Origin, Effect, different kinds of Circumstantials are assigned;

(e) basic features of TFA are introduced.

Three values of a specific TFA attribute assigned to every lexical (autosemantic) occurrence:

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

A typical example, known from older discussions (with *he* bearing a rising contrastive stress and *her* being stressed with the typical sentence final falling pitch contour):

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

Prototypically, 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.

Let us present the description of TFA in PDT by the tectogrammatical analysis of a sample of sentences contained there, to illustrate how this approach makes it possible to analyze also complex sentences as for their TFA patterns, with neither Topic nor Focus corresponding to a single constitutent (argument or adjunct). 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, i.e. Temp and Necess(ity); Fig. 1. 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 (their) debt (to the) TV viewers.
E.: This year, Czech Radiocommunications have quickly to pay their debt to the TV viewers.

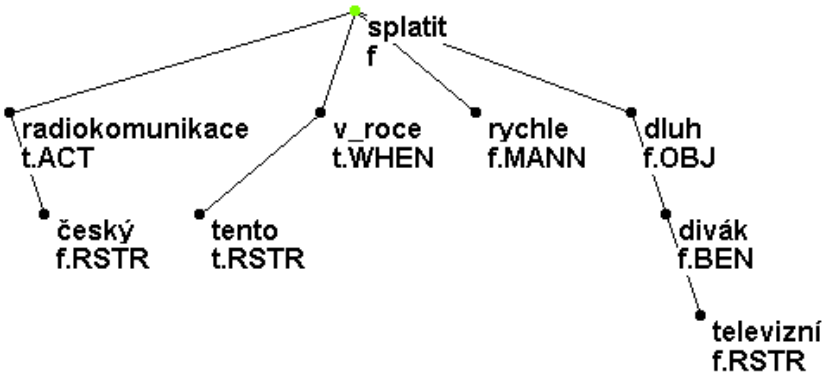

Fig. 1.

The (highly simplified) linearized form of the TGTS:

(5') ((České.f) radiokomunikace.t) ((tomto.t) roce.Temp.t) musí splatit.Necess.f (rychle.f (dluh.f ((televizním.f) divákům.f))

The possibility of such a one-to-one linearization of the dependency tree in the form of a well parenthesized string of complex symbols 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 relationshhips between dependency and coordination (as well as apposition or parenthesis) lead to the possibility of the one-to-one linearization, the parenthesized strings of which come close to proposition calculus. This points to the possibility to describe 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 properties. Thus, also the internalization of the core of the mother tongue could be explained on the basis of such common properties, without postulating a complex framework of innate features.

To illustrate the issues of tectogrammatical annotation, let us add further examples of sentences from PDT:

A focus sensitive particle in the 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.

The linearized TGTS (with many simplifications):

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

Czech differs from English in that the position of the focus sensitive particle *již* ‚already' in (6) directly reflects the boundary between Topic and Focus. With the order *již má* (‚already has'), the main verb would be included in the Focus. However, in the present form, the verb *má* ‚has' is understood as contextually bound (which is well possible with such a semantically poor verb).

Thus, (6) is an example of the primary position of an operator such as *already*, which covers the whole Focus of a sentence, as may be illustrated by semantically more specific examples, cf. (7) and (8):

(7)(a) Jim was looking only for a swimming pool.
(b) Jim only was looking for a swimming pool.

(8)(a) Jerry came to the seminar not to listen to the lecture.
(b) Jerry did not come to the seminar to listen to the lecture.

It is possible to paraphrase (7)(a) by ‚Jim was looking for nothing else than for a swimming pool' and (8)(a) by ‚Jerry came to the seminar for another aim than to listen to the lecture.' Only in the (b) examples the verb itself can be understood as being negated, so that a paraphrase with ‚J. did nothing else than...' is possible.

Our next example from PDT illustrates the (non-prototypical) presence of contextually bound (CB) items (contrastive or not) within Focus:

(9) Přiznám se, že já osobně to dost prožívám.
I-admit that I personally it intensively live-through.
I amdit that I personally live this through quite intensively.

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

In the TGTS (9') the deleted subject pronoun has been restored, and another node has been added for the General Addressee of *prožívám* 'live through'. The subject is expressed, on the morphemic level, by the personal ending of the verb; it is CB and functions as the Topic proper. The main verb together with the embedded clause constitute the Focus, within which the two verbs are NB, as well as the adverb dost, which is the Focus proper. 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 adverb *osobně* ‚personnally', which is contextually nonbound, 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 (10):

(10) Dnes už si však bez něho svoji práci nedovedou představit.
Today already Refl however without him their work they-cannot imagine.
Nowadays, however, they cannot IMAGINE their work without him.
(10') (dnes.t) (však.t) (oni.t) (bez-něho.t) (práci.t (svoji.t)) (už.f) (Neg.f) dovedou-představit-si.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 (10), 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* all occurring as „accompanying members of the Topic". It is necessary to acknowledge that 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 we understand as one of the focus sensitive operators, has the form of the verb prefix *ne-* in Czech.

This way how to describe Topic and Focus of a sentence in a perspicuous manner, without using means which would be unnecessarily complex, may be useful also for describing certain aspects of the foundations of discourse patterns, namely of coreferrence in a connected text.

The basic question in this domain may be looked for in the factors relevant for the addressee's identification of the referents of expressions such as definite noun groups or pronouns. If the discourse contains items such as *the boy*, *the table*, or *he*, how can the hearer/reader find out which boy, which table, or who is meant?

An examination of this question, i.e. of how the referent is identified, has been the object of a longer discussion, see esp. Hajičová et al. (1982; 1998), the results of which might serve to enrich the theories of discourse structure formulated by H. Kamp and others. Especially the following two points are relevant:

(i) it is certain that the ‚iota inversum' operator does not offer the proper ground for specifying the referent of a definite noun group, since in the prototypical case more than one boy or table, etc., are included even in the narrow part of the ‚universe of discourse' (its part that the speaker assumes to be easily accessible for the hearer, the ‚scene'), or in what often is called the stock of knowledge (information) shared by the speaker and the hearer;

(ii) also the often accepted assumption that a definite noun always has an antecedent in the preceding verbal co-text is not fully substantiated.

Let us present an example of a simple discourse segment:

(11) In the library he entered the reading room, took some books from the shelf and put them on the single desk that was free, not knowing that it was reserved for you.

It may be assumed that the antecedents of *he* and of *the library* are present in the preceding co-text (although not necessarily in the preceding sentence token). The referent of *the reading room* may be determined on the basis of the associative links to the word *library* (a kind of accommodation). The presence of the presupposition that the (every?) library has a single reading room probably is not necessary (although it belongs to the pragmatic background of this utterance that the possible existence of other reading rooms at the library is backgrounded). This is similar with *the shelf*, which also exhibits an associative link to the library, without a presupposition that the room contains a single shelf (although there is also an associative link between *the shelf* and *some books*). The group *some books*, being indefinite (‚specifying'), introduces a new referent, not identified, although serving as a starting point for further coreference; in this way, a new member of the scene is established. The pronouns *them* and *it* do have their antecedents in the co-text, and the group *the single desk that was free* constitutes an explicit individual description, again with an associative link to the library. The pronoun *you* is an example of expressions referring to entities which can be mentioned in a discourse without any specific co-textual antecedent, since they either are directly connected to the pragmatic background of the utterance

(as *I, you, now, here*) or belong to the set of entities easily accessible to speakers sharing a certain cultural background (*Shakespeare, Paris, Churchill*) or technical domain.

Our question is, however, what is the finite mechanism the addressee may use to identify the referents in individual utterances of a discourse. The concept used as the basis for answering this question is that of the hierarchy of salience, introduced by D. Lewis and discussed in the publications quoted above; recently see also Krahmer (1998), Krahmer and Theune (1999). The degrees of salinece are understood as relevant for the reference potential of referring expressions in the subsequent utterances in a discourse, and thus also for the connectedness of the discourse. The hierarchy (partial ordering) of salience degrees is being modified by the flow of discourse in a certain way, which we want to capture in annotating the utterances included in PDT. Before we discuss the attributes of the contextual anchoring of word tokens used to this purpose, let us just remark that we believe there is a possibility to enrich H. Kamp's concept of discourse referent so that the referents display degrees of salience. Thus, in (11) above, the situational context of the utterance makes a certain library (and similarly the referents of *he* and *them*) much more salient than other 'competing' referents. Items such as *you* (*Shakespeare*, and so on) refer to entities enjoying a high degree of salience (in the given group of speakers, or generally) in general, without strict temporal limitations.

Let us now illustrate how the hierarchy of the degrees of salience in a discourse can be captured by the descriptive means of PDT.

In the prototypical case, a new discourse referent emerges as corresponding to a lexical occurrence that carries the value f of the attribute TFA; further items referring to the same referent without longer interruptions in the discourse carry the values t or c, with referents determined by their degrees of salience.

The relationships of individual lexical occurrences to their coreferential antecedents are indicated in PDT by specific attributes for coreferential links:

COREF - the lexical value of the antecedent,

CORNUM - the serial number of the antecedent,

CORSNT - value NIL, if the antecedent in the same sentence; otherwise, the value is PREVi (with i being a natural number) for the case in which the antecedent is in the i-th preceding sentence,

ANTEC - value equal to the functor of the antecedent with grammatical coreference: relative clauses, reflexive pronouns (and particles) such as *se* (‚-self'), the relation of control.

As was discussed since the beginning of the 1980s (see Hajičová et al. 1982, 1998), certain basic rules determining the degrees of (reduction of) salience can be assumed. In a schematic way, with n(r) indicating that the referent r has the salience of degree n (a natural number), we in fact measure the reduction of salience:[1]

(i) if r is expressed by a noun (group) or pronoun carrying f, then n(r) => 0(r);

(ii) if r is expressed by a noun (group) or pronoun carrying t or c, then n(r) => 1(r);

(iii) if n(r) => m(r) in sentence S, then m+2(q) obtains for every referent q that is not itself referred to in S, but is immediately associated with the item r present here;

(iv) if r neither is included in S, nor refers to an associated object, then n(r) => n+2(r) if r was referred to by an NB item only, in the precedign co-text; if r was referred to by a CB item, then n(r) => n+1(r).[2]

As example (12) illustrates, if the degree of salience of a referent R that can be expressed by an item r is higher (by 2 or more) than that of its possible competitors, then r is interpreted by the hearer/reader as referring to R (we add indices indicating the degrees of salience to the relevant nouns and pronouns):

(12) Bill.1 met his cousin.0 at the airport yesterday.
(13)(a) He.1 looked very worried.
(b) He.1 was just looking at the list of arrivals there.
(14) He.1 started to explain that...

If (12) is followed in a discourse by (13)(a), then *he* may be interpreted as coreferential with *cousin*; however, if (13)(b) is uttered after (12), then *he* may refer to *Bill* as well as to *cousin.* This shows that the difference between the degrees 0 and 1 is not sufficient for a safe choice of reference. On the other hand, if (14) follows after (12) and (13)(a), then *he* in (14) can only be coreferential with the referent of *he* in (13)(a), since the salience of the other referent has now been lowered to 2, according to rule (iv).

This shows that the degrees of salience are relevant for the choice of reference, although the minimal difference of one degree does not give a reliable basis for the choice and can be outweighed by infwerences based on contextualor other knowledge. This (with an impact of certain features of S. Kuno's ‚empathy', or „the speaker's" viewpoint) is the case of the choice in (13)(a), and this view is confirmed also by examples from PDT, such as the following one:

In the segment of text chosen from PDT, the utterance (5) presented above, and reproduced here as (15), is followed by (16):

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

(15') ((České.f) radiokomunikace.t) ((tomto.t) roce.Temp.t) musí splatit.Necess.f (rychle.f (dluh.f ((televizním.f) divákům.f))

(16) Jejich vysílače dosud pokrývají signálem programu ČT 2 méně než polovinu území republiky.

lit.: their transmitters up-to-now cover by-signal of-program CT 2 less than (the) half of-(the)-territory of-(the)-Republic
E.: Their transmitters hitherto only cover less than a half of the territory of the Republic.

(16') ((jejich.t) vysílače.t) (dosud.t) pokrývají.f (signálem.f (programu.f (ČT.t (2.f)))) ((méně.f (než-polovinu.f)) území.f (republiky.t))

The reference of the pronoun *jejich* (their) in (16) as following (15) by itself is indistinct; the rules (i) - (iv) allow the pronoun to refer either to Czech Radiocommunications or to the TV viewers, since the referent of the former expression exhibiots salience of degree 1 after (15), in which it occurred as CB (t), and the rferent of the latter expression has degree 0 (occurring in (15) as NB, with f). Only inferencing based on knowledge helps then to establish that Czech Radiocommunications, rather than their viewers, are referred to, since they possess tramsmitters, while viewers normally do not have transmitters at their diposal.

If whole texts are examined on this basis, then it may be possible to study much more systematically issues such as those of "topics of texts" or of their individual segments; text segmentation itself may be established with the help of the method outlined. Such inquiries can also bring results advantageous for information retrieval or data mining (distinguishing between texts in which a given topic is actually treated and those in which it is just occasionally mentioned), for question answering, and so on. Let us just remark that an analysis of this kind devoted to a longer text segment that starts, in the PDT, by the sentences (15) and (16) will be included in a paper by E. Hajičová, J. Havelka and P. Sgall, prepared for the fifth volume of *Prague Linguistic Circle Papers* (*Travaux du Cercle linguistique de Prague, n.s.*), to be published by John Benjamins Publ. House, Amsterdam/Philadelphia.

We are well aware that the cotextual factors discussed here are not the only ones relevant for the salience degrees, and that also other factors than salience degrees are to be investigated as possibly being of impact for the reference potential in discourses of different kinds. Among these factors, there are different sources of inferencing based on contextual and other knowledge, such as the situation of the discourse, domain knowledge and cultural background, as well as specific textual patterns (with episodes, poetic effects, and so on). Factors of these and further kinds can be studied on the basis of the salience degrees that are typical for basic discourse situations. In any case, we may conclude that it is useful for a theory of discourse semantics to reflect the degrees of salience. This makes it possible to distinguish the reference potential of referring expressions and thus the connectedness of the discourse. An explicit, formal representation of the semantics of discourse can be enriched by taking into account the degrees of salience of individual discourse referents in different time points of the discourse.

Notes

* Acknowledgement. The work reported on in this paper has been carried out under the projects GACR 405/96/K214 and MSMT LN00A063.

1 This appears to be necessary, since the only fixed degree of salience that can be observed is that of maximal salience, which is exhibited by an item occurring in the Focus of the utterance under examination. It would not be appropriate to work with an arbitrary value for this maximal degree, from which the rate of salience reduction would be abstracted.

2 These tentative rules, formulated here with a slight modification, have been presented at several occasions for the aims of a further discussion. However, they still wait for systematic testing and evaluation, as well as for enrichments and more precise formulations. Such issues may find new opportunities now, when e.g. a comparison with the centering theory is possible and when a large set of annotated examples from continuous texts in PDT is available.

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.

Hajičová and Hladká B. (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. and P. Sgall (1998). *Topic-focus articulation, tripartite structures, and semantic content*. Amsterdam:Kluwer

Hajičová E. and J. Vrbová (1982). On the role of the hierarchy of activation in the process of natural language understanding. In: *COLING* 82. Ed. by J. Horecký. Amsterdam: North Holland, 107-113.

Krahmer E. (1998). Presupposition and anaphora. *CSLI Lecture Notes* 89. CSLI, Stanford, CA.

Krahmer E. and M. Theune (1999). Efficient generation of descriptions in context. In: R. Kibble and K. van Deemter (eds.), *Proceedings of the workshop The Generation of Nominal Expression,* associated with the 11th European Summer School in Logic, Language and Information.

Sgall P., Hajičová E. and J. Panevová (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, ed. by J. L. Mey, Dordrecht:Reidel - Prague: Academia.