Topic-focus and salience[*]

Eva Hajičová
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Praha, Czech Republic
hajicova@ufal.mff.cuni.cz

Petr Sgall
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Praha, Czech Republic
sgall@ufal.mff.cuni.cz

1 Objectives and Motivation

Most of the current work on corpus annotation is concentrated on morphemics, lexical semantics and sentence structure. However, it becomes more and more obvious that attention should and can be also paid to phenomena that reflect the links between a sentence and its context, i.e. the discourse anchoring of utterances. If conceived in this way, an annotated corpus can be used as a resource for linguistic research not only within the limits of the sentence, but also with regard to discourse patterns. Thus, the applications of the research to issues of information retrieval and extraction may be made more effective; also applications in new domains become feasible, be it to serve for inner linguistic (and literary) aims, such as text segmentation, specification of topics of parts of a discourse, or for other disciplines.

These considerations have been a motivation for the tectogrammatical (i.e. underlying, see below) tagging done within the Prague Dependency Treebank (PDT) to contain also attributes concerning certain contextual features, i.e. the contextual anchoring of word tokens and their relationships to their coreferential antecedents.

Along with this enrichment in the intersentential aspect, we do not neglect to pay attention to intrasentential issues, i.e. to sentence structure, which displays its own features oriented towards the contextual potential of the sentence, namely its topic-focus articulation (TFA).

In the present paper, we give first an outline of the annotation scenario of the PDT (Section 2), concentrating then on the use of one of the PDT attributes for the specification of the Topic and the Focus (the 'information structure') of the sentence (Section 3). In Section 4. we present certain heuristics that partly are based on TFA and that allow for the specification of the degrees of salience in a discourse. The application of these heuristics is illustrated in Section 5.

2 Outline of the Prague Dependency Treebank

The Prague Dependency Treebank (PDT) is being built on the basis of the Czech National Corpus (CNC), which grows rapidly in the range of hundreds of millions of word occurrences in journalistic and fiction texts. The PDT scenario comprises three layers of annotation:

(i) the morphemic (POS) layer with about 2000 tags for the highly inflectional Czech language; the whole CNC has been tagged by a stochastic tagger (Hajič and Hladká 1997;1998, Böhmová and Hajičová 1999, Hladká 2000) with a success rate of 95%; the tagger is based on a fully automatic morphemic analysis of Czech (Hajič in press);

(ii) a layer of 'analytic' ("surface") syntax (see Hajič 1998): cca 100 000 Czech sentences, i.e. samples of texts (each randomly chosen sample consisting of 50 sentences of a coherent text), taken from CNC, have been assigned dependency tree structures; every word (as well as every punctuation mark) has a node of its own, the label of which specifies its analytic function, i.e. Subj, Pred, Obj, Adv, different kinds of function words, etc. (total of 40 values);

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

no nodes are added that are not in the surface shape of the sentence (except for the root of the tree, carrying the identification number of the sentence); the sentences from CNC are preprocessed by a dependency-based modification of Collins et al.'s (1999) automatic parser (with a success rate of about 80%), followed by a manual tagging procedure that is supported by a special user-friendly software tool that enables the annotators to work with (i.e. modify) the automatically derived graphic representations of the trees;

(iii) the tectogrammatical (underlying) syntactic layer: tectogrammatical tree structures (TGTSs) are being assigned to a subset of the set tagged according to (ii); by now, the experimental phase has resulted in 20 samples of 50 sentences each; the TGTSs, based on dependency syntax, are much simpler than structural trees based on constituency (minimalist or other), displaying a much lower number of nodes and a more perspicuous patterning; their basic characteristics are as follows (a more detailed characterization of tectogrammatics and motivating discussion, which cannot be reproduced here, can be found in Sgall et al. 1986; Hajičová et al. 1998):

(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;

(f) elementary coreference links (both grammatical and textual) are indicated.

Thus, a TGTS node label consists of the lexical value of the word, of its '(morphological) grammatemes' (i.e. the values of morphological categories), its 'functors' (with a more subtle differentiation of syntactic relations by means of 'syntactic grammatemes' (e.g. 'in', 'at', 'on', 'under'), of the attribute of Contextual Boundness (see below), and of values concerning intersentential links (see below).

3 From Contextual Boundness to the Topic and the Focus of the Sentence

The dependency based TGTSs in PDT allow for a highly perspicuous notation of sentence structure, including an economical representation of TFA, understood as one of the main aspects of (underlying) sentence structure along with all other kinds of semantically relevant information expressed by grammatical means. TFA is accounted for by one of the following three values of a specific TFA attribute assigned to every lexical (autosemantic) occurrence: t for 'contextually bound' (prototypically in Topic), c for 'contrastive (part of) Topic', or f ('non-bound', typically in Focus). The opposition of contextual boundness is understood as the linguistically structured counterpart of the distinction between "given" and "new" information, rather than in a straightforward etymological way (see Sgall, Hajičová and Panevová 1986, Ch. 3). Our approach to TFA, which uses such operational criteria of empirical adequateness as the question test (with the item corresponding to a question word prototypically constituting the focus of the answer), represents an elaboration of older ideas, discussed especially in Czech linguistics since V. Mathesius and J. Firbas, in the sense of an explicit treatment meeting the methodological requirements of formal syntax.

The following rules determine the appurtenance of a lexical occurrence to the Topic (T) or to the Focus (F) of the sentence:

(a) the main verb (V) and any of its direct dependents belong to F iff they carry index f;

(b) every item i that does not depend directly on V and is subordinated to an element of F different from V, belongs to F (where "subordinated to" is defined as the irreflexive transitive closure of "depend on");

(c) iff V and all items k_j directly depending on it carry index t, then those items k_j to which some items l_m carrying f are subordinated are called 'proxy foci' and the items l_m together with all items subordinated to one of them belong to F, where $1 \leq j,m$;

(d) every item not belonging to F according to (a) - (c) belongs to T.

To illustrate how this approach makes it possible to analyze also complex sentences as for their TFA patterns, with neither T nor F corresponding to a single constitutent, let us present the following example, in which (1') is a highly simplified linearized TGTS of (1); every dependent item is enclosed in a pair of parentheses; for the sake of transparency, syntactic subscripts of the parentheses are left out here, 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:

(1) České radiokomunikace musí v tomto roce rychle splatit dluh televizním divákům.
This year, Czech Radiocommunications have quickly to pay their debt to the TV viewers.
(1') ((České.f) radiokomunikace.t) ((tomto.t)
Czech Radiocommunications this
roce.Temp.t) splatit.Necess.f (rychle.f)
in-year must-pay quickly
(dluh.f ((televizním.f) divákům.f))
debt TV viewers

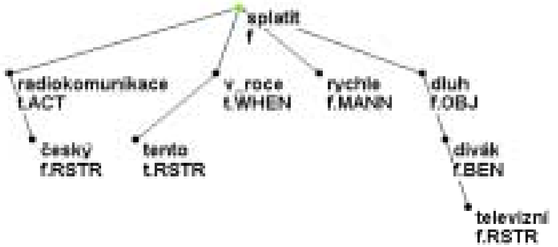

Figure 1.

4 Degrees of Salience in a Discourse

During the development of a discourse, in the prototypical case, a new discourse referent emerges as corresponding to a lexical occurrence that carries the index f; its further occurrences in the discourse carry t and are primarily guided by the scale of their degrees of salience. This scale, which was discussed by Hajičová and Vrbová (1982), has to be reflected in a description of the semantico-pragmatic layer of the discourse. In this sense our approach can be viewed as pointing to a useful enrichment of the existing theories of discourse representation (cf. also Kruijffová 1998, Krahmer 1998; Krahmer and Theune 1999).

In the annotation system of PDT, not only values of attributes concerning sentence structure are assigned, but also values of attributes for coreferential links in the discourse, which capture certain features typical for the linking of sentences to each other and to the context of situation and allow for a tentative characterization of the discourse pattern in what concerns the development of salience degrees during the discourse.

The following attributes of this kind are applied within a selected part of PDT, called 'model collection' (for the time being, essentially only pronouns such as 'on' (he), including its zero form, or 'ten' (this) are handled in this way):

COREF: the lexical value of the antecedent,
CORNUM: the serial number of the antecedent,
CORSNT: if the antecedent in the same sentence: NIL, if not: PREVi for the i-th preceding sentence.

An additional attribute, ANTEC, with its value equal to the functor of the antecedent, is used with the so-called grammatical coreference (relative clauses, pronouns such as 'se' (-self), the relation of control).

On the basis of these attributes (and of further judgments, concerning especially associative links between word occurrences), it is possible to study the referential identity of different word tokens in the flow of the discourse, and thus also the development of salience degrees.

The following basic rules determining the degrees of salience (in a preliminary formulation) have been designed, with x(r) indicating that the referent r has the salience degree x, and $1 \leq m,n$:

(i) if r is expressed by a weak pronoun (or zero) in a sentence, it retains its salience degree after this sentence is uttered: n(r) --> n(r);

(ii) if r is expressed by a noun (group) carrying f, then n(r) --> 0(r);

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

(iv) 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[1];

(v) if r neither is included in S, nor refers to an associated object, then n(r) --> n+2(r).

These rules, which have been checked with several pieces of English and Czech texts, capture such points as e.g. the fact that in the third utterance of *Jim met Martin. He immediately started to speak of the old school in Sussex. Jim invited him for lunch* the weak pronoun in object can only refer to Martin, whose image has become the most salient referent by being mentioned in the second utterance; on the other hand, the use of such a pronoun also in the subject (in *He invited him for lunch*) would make the reference unclear.

Since the only fixed point is that of maximal salience, our rules technically determine the degree of salience reduction (indicating 0 as the maximal salience). Whenever an entity has a salience distinctly higher than all competing entities which can be referred to by the given expression, this expression may be used as giving the addressee a sufficiently clear indication of the reference specification.[2]

5 Illustrations

The development of salience degrees during a discourse, as far as determined by these rules, may be illustrated on the basis of five sentence tokens (utterances) from PDT, starting from (1), which constitute a segment of a newspaper text (we indicate the numerical values of salience reduction for every noun token that is a referring expression). We present here - similarly as with (1') in Section 3 above - highly simplified representations of these sentences, with parentheses for every dependent member and the symbols t, c, and f for contextual boundness; numbers of the degrees of salience (more precisely, of salince reduction) for every referring expression are inserted in the sentences themselves. This example should enable the reader to check (at least in certain aspects) the general function of the procedure we use, as well as the degree of its empirical adequacy in the points it covers, and also our consistence in assigning the indices. We are aware of the preliminary character of our analysis, which may and should be enriched in several respects (not to cover only noun groups, to account for possible episodic text segments, for oral speech with the sentence prosody, for cases of deictically, rather than anaphorically conditioned salience, etc.).

We do not reflect several peripheral points, such as the differences between surface word order and the scale of CD (underlying WO), mainly caused by the fact that a dependent often precedes its head word on the surface (in morphemics), although if the dependent has f (as e.g. *rychle* (*quickly*) has in (1)), then it follows its head under CD (with the exceptions of focus sensitive particles, cf. Hajičová, Partee and Sgall 1998); our translations are literal.

(1) České radiokomunikace.1 musí v tomto roce.1 rychle splatit dluh.0 televizním divákům.0

In this year, Czech Radiocommunications have quickly to pay their debt to the TV viewers.

(1') ((České.f) radiokomunikace.t) ((tomto.t)
Czech Radiocommunications this
roce.Temp.t) splatit.Necess.f (rychle.f)
in-year must-pay quickly
(dluh.f ((televizním.f) divákům.f))
debt TV viewers

(2) Jejich.1 vysílače.1 dosud pokrývají signálem.0 programu.0 ČT.1 2.0 méně než-polovinu.0 území.0 republiky.0.

Their transmitters hitherto cover by-signal of-the-program ČT2 less than a-half of-the-territory of-the-Republic.

(2') ((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))

(3) Na moravsko-slovenském pomezí.1 je řada míst.0, kde nezachytí ani první program.0 České televize.1.

On the-Moravian-Slovakian borderline there-is a-number of-places where (they) do-not-get even the-first program of-Czech Television.

[1] Only immediate associative links are taken into account for the time being, such as those between (*Czech*) *crown* and *money*, or between *TV* or (*its*) *signal* and (*its*) *viewer*.

[2] These tentative rules, which have been presented at several occasions (starting with Hajičová and Vrbová 1982) for the aims of a further discussion, still wait for a systematic testing and evaluation, as well as for enrichments and more precise formulations. These issues may find new opportunities now, when e.g. a comparison with the centering theory gets possible and when a large set of annotated examples from continuous texts in PDT is available. An automatic derivation of such features can only be looked for after the lexical units included get a very complex and subtle semantic classification.

(3') ((na-moravsko-slovenském.t) pomezí.t) je.f (řada.f (míst.f ((kde.t) (oni.t) (ne.f) zachytí.f ((ani.f) (první.f) program.t ((České.t) televize.t)))))

(4) Do rozdělení.1 federace.1 totiž signál.1 zajišťovaly vysílače.0 v SR.0.

Until the-division of-the-federation as-a-matter-of-fact the-signal.Accusative provided transmitters.Nominative in S(lovac)R(epublic).

(4') (do-rozdělení.t (federace.t)) (totiž.t) (signál.t) zajišťovaly.t (vysílače.f (v-SR.f)).

(5) Česká televize žádá urychlenou výstavbu nových vysílačů.

Czech Television requires quick construction of-new transmitters.

(5') ((Česká.t) televize.t) žádá.f ((urychlenou.f) výstavbu.f ((nových.f) vysílačů.t))

The development of salience reduction of the referents most frequently mentioned in (1) - (5) is characterized in Tab. 1, which includes numbers of salience reduction degrees and of those rules from Section 3 that are the main sources of the degrees. Two further remarks may be added, concerning details of our analysis that have not been discussed above and may not be directly found in the previous publications we refer to: (a) a noun group consisting of a head with t or c and of one or more adjuncts with f constitutes a referring expression as a whole, in the prototypical case, and gets degree 0, if it occurs in F; this concerns e.g. the group *vysílače v SR* ('transmitters in the Slovac Republic') in sentence (4), or *ČT 2* (*CTV 2*) in (2); here *2* is treated as an adjunct of *CT*; (b) the difference between the degrees 0 and 1 is not sufficient for a safe choice of reference, so that, e.g., the reference of the pronoun *jejich* (*their*) after (1) by itself is indistinct, and only inferencing helps to establish that *České radiokomunikace* (*Czech Radiocommunications*) are referred to (viewers normally do not have transmitters at their diposal).

after	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
CRC	1 (iii)	1 (iii)	3 (iv)	5 (v)	7 (v)
CTV	3 (iv)	1 (iii)	1 (iii)	2 (iv)	1 (iii)
CTV1	2 (iv)	2 (iv)	0 (ii)	2 (iv)	3 (iv)
CTV2	2 (iv)	0 (ii)	2 (iv)	2 (iv)	3 (iv)
viewer	0 (ii)	2 (iv)	2 (i)	3 (iv)	3 (iv)
sig.	3 (iv)	0 (ii)	2 (iv)	1 (iii)	3 (iv)
CR	3 (iv)	1 (iii)	3 (iv)	3 (iv)	3 (iv)
CSF	-	-	3 (iv)	1 (iii)	3 (v)
terr.	3 (iv)	0 (ii)	2 (iv)	2 (iv)	4 (v)
tr.	-	1 (iii)	2 (iv)	0 (ii)	0 (ii)

Table 1.

Abbreviations:
CRC - Czech Radio(tele)communications
CTV - Czech TV
CR - Czech Republic
CSF - (CS) Federation
CTV1(2) - 1st (2nd) program of CTV
tr. - transmitter
terr. - territory of CR
sig. - signal of CTV

Even with this short piece of discourse, its segmentation is reflected, if its first subsegment, discussed up to now (sentences (1) - (5)), is compared with its continuation, i.e. sentences (6) - (9), given below. While the first segment deals primarily with CTV and its signal (cf. the relatively high salience of *CTV, CTV1, CTV2, RC, signal* and *viewer* in most parts of the segment), sentences (6) – (9) are devoted to financial issues, as can be seen from the following facts: (a) *money* gets degree 0 after (6), in which it functions as its focus proper (the most dynamic item), (b) *Czech crown* gets degree 1 after (7), in which it is an embedded part of the focus, and (c) the group *financial coverage* gets degree 1 in sentence (8).

The continuation is presented here without the TGTSs:

(6) Naše společnost může úkol splnit, ale chybějí nám peníze.

Our company can the-task.Accusative fulfil, but is-lacking us.Dative the-money.Nominative.

(7) Letos by výstavba technického zařízení v sedmi lokalitách stála 120 miliónů korun, ale můžeme uvolnit jen 80 miliónů.
This-year, would the-construction of-technical equipment in seven localities cost 120 million crowns, but we-can spend only 80 million.

(8) Proto o finančním zabezpečení jednáme s Českou televizí, uvádí ekonomický ředitel Českých radiotelekomunikací Miroslav Cuřín.
Therefore about (its) financial coverage we-discuss with Czech Television, states the-economic director of-Czech Radiotelcommunications M. C.

(9) Dalších 62 miliónů korun si vyžádá výstavba vysílačů a převaděčů signálu v pohraničí.
Further 62 million crowns.Accusative Refl. will-require the-construction.Nominative of-transmitters and transferrers of-the-signal in the-border-area.

6 Conclusions

We are aware that, along with the rules characterized above, there are other factors that have to be investigated, which are important for different kinds of discourses. This concerns various aspects of the discourse situation, of domain knowledge, of 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, which 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. Discourse analysis of this kind may also be useful for application domains such as text segmentation (in accordance with topics of individual segments), or data mining (specifying texts in which a given topic is actually treated, rather than being just occasionally mentioned).

References

Böhmová A. and E. Hajičová (1999). The Prague Dependency Tree Bank I: How much of the underlying syntactic structure can be tagged automatically? The Prague Bulletin of Mathematical Linguistics 71, 5-12.

Collins M., Hajič J., Brill E., Ramshaw L. and C. Tillmann (1999). A statistical parser for Czech. In: Proceedings of 37th Annual Meeting of ACL, Cambridge, Mass.: M.I.T. Press, 505-512.

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. (in press). Disambiguation of rich inflection (Computational morphology of Czech). Prague:Karolinum.

Hajič J. 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.

Hajič J. and Hladká B. (1998). Czech language processing - POS tagging. In: Proceedings of the First International Conference on Language Resources & Evaluation, Granada.

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.

Kruijff-Korbayová I. (1998): The dynamic potential of topic and focus: A Praguian approach to Discourse Representation Theory. Prague: Charles University, Faculty of Mathematics and Physics, Ph.D. dissertation.

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.