Topic-Focus Revisited (Through the Eyes of the Prague Dependency Treebank)

Eva Hajičová
Charles University in Prague
hajicova@ufal.mff.cuni.cz

Abstract

The distinctions covered by information structure of the sentence (its topic-focus articulation, TFA) are argued to be semantically relevant. In the Praguian theoretical framework of Functional Generative Description, their representation is integrated into the description of the underlying (tectogrammatical) level of language, which is suitable as the input to semantico-pragmatic interpretation. The phenomena connected with TFA on other levels (word order, particles, clefting, prosody etc.) serve as means expressing TFA. The primary opposition is the opposition of contextual boundness, from which the bipartition of the sentence into its Topic and Focus and other related notions can be derived. The present-day availability of corpora annotated in a systematic and linguistically-based manner allows for testing linguistic hypotheses, as the experience of the Prague Dependency Treebank indicates. It is the purpose of this contribution to sum up the hitherto reached insights in the domain of TFA that the annotated corpus has made available.

Keywords
Topic-Focus articulation, Prague Dependency Treebank

1 Introduction

Among the several linguistic issues Igor Mel‘čuk’s Meaning-Text Theory and the Praguian theory of Formal Generative Description of language have in common, is the view that the communicative organization of the sentence is an extremely difficult though frequently discussed subject but that it is still worth to be systematically studied (Mel´čuk, 2001, p.2). In my paper delivered at the MTT conference in Klagenfurt in 2007 (Hajičová, 2007) I presented a comparison of our standpoints to those embodied in other models, paying a special attention to Igor Mel’čuk’s Meaning Text Theory. I emphasized that a deeper empirical analysis of sentences (in their context) in various languages convincingly shows that the issues referred to as belonging to Topic-Focus Articulation (TFA in the sequel) - or communicative structure, information structure, theme-rheme or whatever terms are used) - are semantically relevant. Therefore within the Praguian theoretical framework of Functional Generative Decription I subscribe to, the representation of these phenomena is integrated into the description of the underlying, deep syntactic (tectogrammatical) level of language description. It is this level that is suitable as the input to semantico-pragmatic interpretation. The phenomena connected with TFA on other levels (word order, or also particles, clefting, prosody etc.) serve as means expressing TFA. I also argued that the primary opposition to be distinguished is the opposition of contextual boundness, from which the bipartition of the sentence into its Topic and Focus and other related notions can be derived. Such a description offers an adequate, effective and economic way of capturing the

corresponding semantically relevant distinctions. A well-suited way of testing the theoretical assumptions and hypotheses is the present-day availability of corpora annotated in a systematic and linguistically-based manner, as the experience of the Prague Dependency Treebank indicates. It is the purpose of this contribution to sum up the hitherto reached insights in the domain of TFA that the annotated corpus has made available.

2 Prague Dependency Treebank in a nutshell

The Prague Dependency Treebank (PDT, see e.g. Hajič, 1998; Hajič et al.. 2006; Mikulová et al. 2006) is an annotated collection of Czech texts, randomly chosen from the Czech National Corpus (CNK), with a mark-up on three layers: (a) morphemic, (b) surface shape ("analytical"), and (c) underlying (tectogrammatical). The current version (publicly available on http://ufal.mff.cuni.cz/pdt2.0), annotated on all three layers, contains 3168 documents (text segments mainly from journalistic style) comprising 49431 sentences and 833195 occurrences of word forms (including punctuation marks).

The annotation scheme of PDT is based on the framework of the Functional Generative Description (FGD; for a comprehensive description of this framework see (Sgall, Hajičová & Panevová, 1986). On the tectogrammatical level, every node of the tectogrammatical representation (TGTS, a dependency tree) is assigned a complex label consisting of the *lexical value* of the word, of its *'(morphological) grammatemes'* (i.e. the values of morphological categories), of its *'functors'* (with a more subtle differentiation of syntactic relations by means of *'syntactic grammatemes'*, e.g. 'in', 'at', 'on', 'under'), and the TFA attribute containing values for *contextual boundness*. In addition, some basic intersentential (discourse based) and coreferential (both grammatical and textual) links are also added. It should be noted that TGTSs may contain nodes not present in the morphemic form of the sentence in case of surface deletions.

In PDT, the attribute specifying TFA contains three values, one of which is assigned to every node of the tectogrammatical tree structure. The contextually bound nodes obtain either the values *t* or the value *c*; the value *t* stands for a contextually bound non-contrastive node, *c* for a contextually bound contrastive node; a contextually non-bound node gets the value *f* .

In the theoretical framework, a set of rules was formulated (see Sgall, 1979, p. 180; Sgall et al., 1986, pp. 216ff) based on the notion of contextual boundness; the rules determine the appurtenance of a lexical occurrence to the Topic (T) or to the Focus (F) of the sentence, and as such they reflect the aboutness relation (Focus of the sentence is ABOUT the Topic of the sentence).

The rules are specified as follows (*nb* stands for a contextually non-bound node, *cb* for a contextually bound node, which may be contrastive or non-contrastive):

(a) the main verb (V) and any of its direct dependents belong to F iff they carry index *nb*;
(b) every item 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 directly depending on V are *cb*, then it is necessary to specify the rightmost *k' node* of the *cb* nodes dependent on V and ask whether some of nodes *l* dependent on *k'* are *nb*;

if so, this *nb* node and all its dependents belong to F; if not so, then specify the immediately adjacent (i.e. preceding) sister node of *k'* and ask whether some of its dependents is *cb*; these steps are repeated until an *nb* node depending (immediately or not) on a cb node directly dependent on V is found. This node and all its dependent nodes are then specified as F.
(d) every item not belonging to F according to (a) - (c) belongs to T.

The application of the rules is illustrated by Fig. 1.

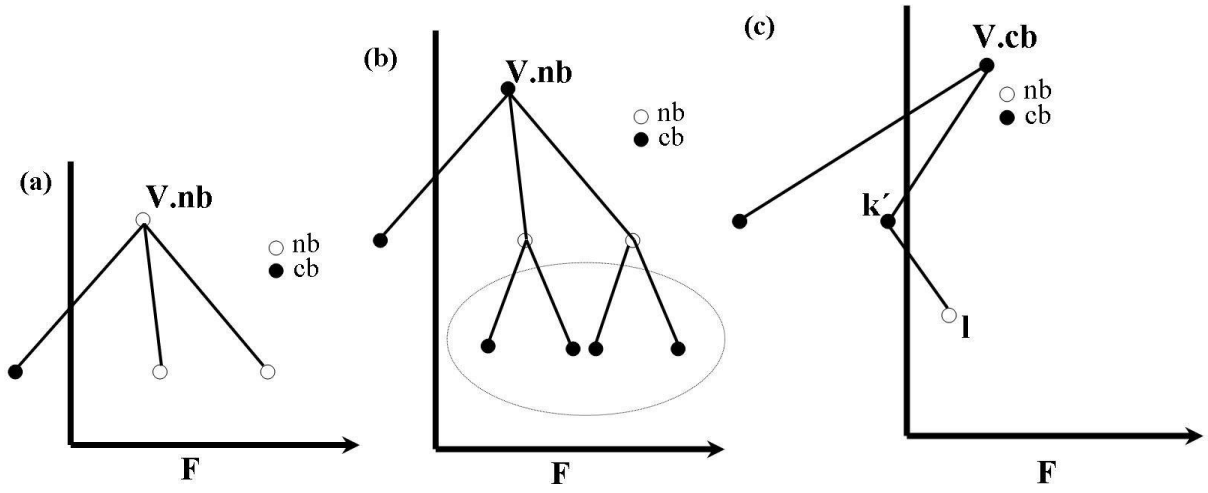

Fig. 1
Application of the algorithm dividing a TGTS into Topic and Focus

3 Annotated corpus as a testbed for a linguistic theory

3.1 Introduction

One of the advantages corpus annotation offers (if the annotation scheme is based on a sound linguistic theory and the annotation scenario is carefully, i.e. systematically and consistently designed) lies in the fact that the data acquired can be used for further linguistic research (see e.g. Hajičová & Sgall, 2006); this fact is well supported by the existence of annotated corpora of various languages: let us quote as examples the Penn Treebank for English (Marcus et al., 1993;1994), the PropBank and Penn Discourse Treebank developed also for English (Prasad et al., 2008; Miltasaki et al., 2008), the Tiger Treebank for German (Brants et al., 2002), SynTagRus for Russian (Boguslavsky et al., 2002) , or the Prague Dependency Treebank for Czech briefly characterized in Sect. 2 above. It is worth noting in this connection to see that also the Meaning-Text Theory is being discussed as a possible underlying linguistic theory for a scheme of treebank annotation (Mille et al., 2012).

The usefulness of PDT annotation for the study of Czech syntax has been documented by many papers by Jarmila Panevová and her students (see e.g. Panevová, 2003; 2004; 2008; 2011; Panevová & Ševčíková 2011); in the present contribution, we adduce some examples from the domain of topic-focus articulation and coreference relations; the existence of a parallel Czech-

English Treebank based on the same principles as PDT makes it possible also to make some contrastive observations.

3.2 The bipartition of a sentence into Topic and Focus

The algorithm dissecting the sentence into its Topic (T) and Focus (F) mentioned in Sect. 2 above is based on the hypothesis that the division of the sentence into its T and F can be derived from the contextual boundness of the individual lexical items contained in the underlying representation of the sentence.

The results of the implementation are quite encouraging and they allow for some interesting observations: in 85,7% the verb belongs to Focus; in 8,58% the verb belongs to Topic but there always was a node or nodes depending directly on the verb that were contextually non-bound and thus belong to Focus; only in 4,41% of sentences the Focus was more deeply embedded (i.e. depends on some contextually-bound node). The algorithm failed in 1,2% cases when its application has led to an ambiguous partition and in 0,11% cases where no Focus was identified. Looking at these figures, we see another interesting result of the implementation of the algorithm and its application on the annotated corpus: in 95% of the cases the hypothesis (present also in the Functional Sentence Perspective theory as proposed by Jan Firbas, (see e.g. Firbas, 1959; 1992 on the transitional character of the verb) that in Czech the boundary between Topic and Focus is in the prototypical case signalized by the position of the verb was confirmed.

To validate the results of the automatic procedure in comparison with "human" annotation, a subset of the corpus (with the TFA assignment hidden) was selected and human annotators were asked to mark, on the basis of their native speakers' judgements what is the sentence 'about', that is, which part of the sentence is its Topic and which is its Focus. These 'human' assignments were then compared with the results of the automatic procedure (Zikánová et al., 2007; Zikánová &Týnovský, 2009). When evaluating the results, the main observation was that the correspondence supports the algorithm; the most frequent differences, if any, concerned the difference in the assignment of the verb to topic or to focus. This confirms again the transitional character of the verb in Czech.

The results then can be summarized as follows: in Czech, the boundary between Topic and Focus can be determined in principle on the basis of the consideration of the status of the main predicate and its direct dependents. The TFA annotation leads to satisfactory results in cases of rather complicated "real" sentences in the corpus. Certain modifications of the annotation procedure are necessary, but the material gathered and analyzed in this way may be further used for the study of several aspects of discourse patterning (Hajičová, 2012 and Sect. 3.6 below).

3.3 Systemic ordering

Another hypothesis that has already been tested on our annotated corpus concerns the order of elements in the Focus. It is assumed that in the focus part of the sentence the complementations of the verb (be they arguments or adjuncts) follow a certain canonical order in the underlying structure, the so-called systemic ordering. In Czech, also the surface word order in Focus corresponds to the systemic ordering in the prototypical case.

The following underlying systemic ordering is postulated For Czech (see Sgall et al., 1986), which, in the protoypical case, is in Czech reflected also by the surface word order in Focus: Actor – Time:*since-when* – Time:*when* – Time: *how-long* – Time:*till-when* – Cause – Respect – Aim – Manner – Place – Means – Dir:*from-where* – Dir:*through-where* – Addressee – Origin – Patient – Dir:*to-where* – Effect.

Systemic ordering as a phenomenon is supposed to be universal; however, languages may differ in some specific points. The validity of the hypothesis has been tested with a series of psycholinguistic experiments (with speakers of Czech, German and English); for English most of the adjuncts follow Addressee and Patient (Sgall et al., 1995). However, PDT offers a richer and more consistent material; preliminary results have already been achieved based on (a) the specification of Focus according to the algorithm mentioned above, (b) the assumed order according to the scale of systemic ordering (functors in TGTS), and (c) the surface word order (Zikánová, 2006). This information can be used to compare the order of the complementations in the actual sentence with the assumed order according to the scale of systemic ordering and to propose some more subtle formulation of the hypothesis or its modification, as documented by the studies of Rysová (2011a; 2011b).

3.4 Contrastive topic

The original formulation of the TFA theory works with the binary distinction between contextually bound and non-bound nodes. However, a more consistent work with the empirical material during the corpus annotation, an observation was made that in some sentences a part of the Topic can be distinguished that actually expresses a contrast, though different from the contrast expressed – by default – in the Focus. (Focus is understood by most of researchers as a choice of alternatives thus actually involving a contrast to the non-selected alternatives.) This contrastive (part of the) Topic can be distinguished from the other part(s) of the Topic by two features: by some specific intonation contour and by the use of a long form of pronoun in the topic position in Czech, see (1), with the intonation center marked by capitals.

(1) Milena nás seznámila se svým BRATREM. *Jeho* jsme pozvali do PRAHY a do *Brna* jsme jeli s NÍ.
Milena – us – acquainted – with- - her – BROTHER. *Him* – (we)Aux - invited – to PRAGUE – and - to - *Brno* - (we)went – with – HER.

In (1), *jeho* is the long form of Acc.sing. of the pronoun ‘on’ (he), the short form of this pronoun being *ho* as in (2).

(2) Pozvali jsme ho do PRAHY.
(we)invited - Aux. – him - to – PRAGUE

This observation (see Koktová 1999) has led us to introduce the notion of a contrastive topic into the TFA theory (see Hajičová, Partee & Sgall, 1998) and in accordance with it to introduce a third value of the TFA attribute in the annotation scheme of PDT, namely the value *c*. The PDT material with such a more subtle differentiation of contextually bound nodes has made it possible to study the phenomenon of contrast in a more detail (see Hajičová & Sgall 2001); when applied

to a small corpus of Czech spoken discourse it was also possible to trace the difference between contrastive topic and focus with respect to sentence prosody (Veselá, Peterek & Hajičová, 2003).

3.5 Passivization in English as one of the means of TFA

A quite self-evident basic hypothesis says that in English passivization is one of the possibilities how to "topicalize" Patient (Object). A natural, though rather simplified implication is that such a topicalized Patient can be used with an indefinite article only in specific cases.

For the purpose to check under which conditions such an implication holds, we have used another Praguian corpus, namely the parallel corpus of English and Czech called Prague Czech-English Dependency Treebank. This corpus consists of 49208 sentences with the total number of 54304 predicates (roughly: clauses). In the corpus, there are 194 cases which seemingly contradict the above mentioned assumption, i.e. in which a subject of a passive sentence is accompanied by an indefinite article (see Hajičová, Mírovský & Brankatschk, 2011).

Looking at these cases in more detail, most frequent constructions are those with General Actor, i.e. an Actor that is not expressed in the surface shape of the sentences. The surface subject has the function of the Patient. The placement of an indefinite expression at the front position (even though it is the focus of the sentence) is due to the grammatically fixed English word-order. In the Czech counterparts, the Patient is placed at the final position, in the normal focus position. These cases are exemplified here by sentences in (3) and (4) and the sentence elements in question are printed in italics.

(3) (Preceding context: Soviet companies would face fewer obstacles for exports and could even invest their hard currency abroad. Foreigners would receive greater incentives to invest in the U.S.S.R.)
Alongside the current non-convertible ruble, *a second currency* would be introduced that could be freely exchanged for dollars and other Western currencies .

(3') Cz. Zároveň se současným nekonvertibilním rublem bude zavedena *druhá měna*, která by mohla být volně směnitelná za dolary a další západní měny.

(4) (Preceding context: He notes that industry executives have until now worried that they would face a severe shortage of programs once consumers begin replacing their TV sets with HDTVs. Japanese electronic giants, such as ..., have focused almost entirely on HDTV hardware, and virtually ignored software or programs shot in high-definition.)
And *only a handful of small U.S. companies* are engaged in high-definition software development.

(4') Cz. A vývojem softwaru pro vysoké rozlišení se zabývá *jen hrstka malých amerických společností.*

A second group of cases can be characterized by the use of the indefinite article in the meaning "one of the", cf. (5).

(5) *A seat on the Chicago Board of Trade* was sold for $ 390,000, unchanged from the previous sale Oct. 13. (The following context: Seats currently are quoted at $ 361,000 bid, $395,000 asked. The record price for a full membership on the exchange is $550,000, set Aug. 31 , 1987.)

(5') Cz.: *Členství v Chicagské obchodní radě* bylo prodáno za 390 000 dolarů, což je o 5 000 dolarů méně než při posledním prodeji minulý čtvrtek.

Exceptionally, but still, there occurred cases which can be interpreted as a contrast in the topic part, cf. (6).

(6) (Preceding context: DOT System. The `` Designated Order Turnaround " System was launched by the New York Stock Exchange in March 1976 , to offer automatic, high-speed order processing.) *A faster version* , the SuperDot , was launched in 1984 .

(6') Cz. Rychlejší verze SuperDot byla spuštěna v roce 1984.

It is a matter of course that a more systematic investigation of the mentioned issue is necessary; it will be also of interest to look at these structures in a spoken corpus of English to see whether a 'fronted' Patient into the subject position accompanied by an indefinite article in English is marked by some specific features of the intonation contour that would indicate its appurtenance to Focus or to a contrastive part of the Topic.

3.6 (Some) heuristics guiding the development of the activation degrees

In our previous studies of some aspects of discourse patterns, we formulated the following hypothesis: A finite mechanism exists that enables the addressee to identify the referents on the basis of a partial ordering of the elements in the stock of knowledge shared by the speaker and the addressees (according to the speaker's assumption), based on the degrees of activation (salience) of referents.

The research question we asked then is whether it is possible to combine the "dynamic" (communication based) view of language and discourse with the description of (underlying) sentence syntax the TFA aspect of which was described above in Sect. 2. Or, in other words, how to combine the "dynamic" (communication based) view of language and discourse (and textual coreference) with the description of (underlying) sentence syntax.

To this aim, we introduced (Hajičová & Vrbová, 1982; for a more detailed treatment see esp. Hajičová, 1993;1997; 2003a; 2003b; 2012) the notion of the stock of knowledge assumed by the speaker to be shared by him and the hearer; this stock of shared knowledge, of course, is not a undifferentiated collection, but a hierarchized structure based on the different degrees of salience (activation) of its elements. This scale has to be reflected in a description of the semantico-pragmatic layer of the discourse.

The following three basic heuristics (a) through (c) based on the position of the items in question in the topic or in the focus of the sentence, on the means of expression (noun, pronoun) and on the previous state of the activation can be formulated to determine the degrees of salience of the elements of the stock of shared knowledge:

(a) In the flow of communication, a discourse referent enters the discourse, in the prototypical case, first as contextually non-bound, thus getting a high degree salience. A further occurrence of the referent is contextually bound, the item still has a relatively high degree of salience, but lower than an element referred to in the focus (as contextually non-bound) in the given sentence, see (7).

(7) The night before her mother left, Irena introduced her to her companion, Gustaf, a Swede. The three of them had dinner in a restaurant, and the mother, who spoke not a word of French, managed valiantly with English. Gustaf was delighted: with his mistress, Irena, he spoke only French, and he was tired of that language, which he considered pretentious and not very practical. (Kundera, p. 22)

In the first sentence of this paragraph, 'Gustaf, a Swede' is introduced for the first time; he is rementioned simply as 'Gustaf' in the topic part of the third sentence (the sentence is 'about' him) and in the following sentences, as a relatively salient element, with no competitor in reference, he is referred to just by the pronoun 'he'. We should say in this connection that the stock of knowledge of the speaker/hearer contains also some permanently salient referents such as: *here, now* etc., which can stand in the topic part of the sentence without having been mentioned in the previous co-text.

(b) If an item is not referred to in the given sentence, the degree of salience is lowered; the fading is slower with a referent that had in the previous co-text occurred as contextually bound; this heuristics is based on the assumption that a contextually bound item has been 'standing in the foreground' for some time (as a rule, it was introduced in the focus, then used as contextually bound, maybe even several times) and thus its salience is reinforced; it disappears from the set of the highly activated elements of the stock of shared knowledge in a slower pace that an item which has been introduced in the focus but then dropped out, not rementioned. If the referent has faded too far away it has to be re-introduced in the focus of the sentence, see (8):

(8) In 1921 Arnold Schoenberg declares that because of him German music will continue to dominate the world for the next hundred years. Twelve years later he is forced to … After the war … he is still convinced … He faults Igor Stravinskij for paying too much attention to his contemporaries …
<two pages later> As I said, he was living in the very lofted spheres of mind, … The only great adversary worthy of him, the sublime rival whom he battled with verve and severity, was Igor Stravinskij.
(Kundera pp. 144, 146)

(c) If the difference in the degree of salience of two or more items is very small, then the identification of reference can be done only on the basis of inferencing. In the segment of (9), both Milada and Irena are highly salient items and can be referred to by the pronoun *she*; the assumed concrete reference assignment is indicated by M for Milena and I for Irena.

(9) Milada had been a colleague of Martin's working at the same institute. Irena had recognized her [M] when she [M] first appeared at the door of the room, but only now, each of them with a wine glass in hand, is she [I] able to talk to her [M]. She [I] looks at her [M]: Milada still has the

same shape face …
(Kundera, p. 9)

The mentioned three basic heuristics served as a basis for our formulation of several rules for the assignment of the degrees of salience, which we have applied to numerous text segments to check how the determination of these degrees may help reference assignment.

The following basic rules determining the degrees of salience (in a preliminary formulation and taking into account only nominal referents) have been designed, with $dg_i(r)$ indicating the salience degree of the referent r after the *i*-th sentence S_i of a document is uttered:

(i) if r is expressed by a weak pronoun (or zero, i.e. deleted in the surface shape) in a sentence, it retains its salience degree after this sentence is uttered: $dg_i(r) := dg_{i-1}(r)$;
(ii) if r is expressed by a noun (group) carrying nb, then $dg_i(r) = 0$;
(iii) if r is expressed by a noun (group) carrying cb, then $dg_i(r) = 1$;
(iv) $dg_i(q) := dg_i(r) +2$ obtains for every referent q that is not itself referred to in S_i, but is immediately associated with an item present here;
(v) if r neither is included in S_i, nor refers to an associated object, and has been mentioned in the focus of the preceding sentence, then $dg_i(r) := dg_{i-1}(r) +2$.
(vi) if r neither is included in S_i, nor refers to an associated object, and has been mentioned in the topic of the preceding sentence, then $dg_i(r) := dg_{i-1}(r) +1$.

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. It should be emphasized that what matters is the relation higher/lower degree, rather than the absolute numerical value; also the difference of 1 is too small to be relevant (see point (c) and the ex. (9) above).

As we have pointed out in our previous papers, such an analysis of discourse makes it possible to throw some light on several issues of discourse structure:

(a) Certain patterning can be readily observed: e.g. a more or less regular change of groupings of items on the "top of the stock" if the discourse fluently passes from one group of items talked about to another group, or a cluster of items staying on the top with other items just entering the stage and leaving it very quickly.

(b) The proposed representation of the flow of discourse offers one way of segmenting the discourse more or less distinctly into smaller units according to which items are the most activated ones in these stretches.

(c) The proposed representation of the flow of discourse can serve as a basis for the identification of 'topics' of the discourse. It is often disputable to determine 'the' topic of a given discourse; however, the discourse topic(s) occur (or at least the items associated with these topic(s), whatever the notion of association may be understood to stand for) most probably among the

items staying longer (or more frequently) among the most activated items, i.e. on the top of the stock

(d) An interesting issue for further investigation is that of the identification of possible thresholds that may be used together with other prerequisites to study the possibility/impossibility of pronominal reference, for the use of a full definite NP in the Topic, for the necessity to use stronger means for a reintroduction of some already mentioned item, and for the necessity of such a reintroduction to occur in the Focus. The presence of 'competitors', of course, is highly relevant for such investigations. The results of such inquiries may be helpful both for the identification of pronominal reference and for the generation of referring expressions in texts (summaries, question-answering etc.).

(e) Last but not least, the possibility to follow the development of the degrees of activation of individual elements in a text may also help to judge the cohesion of the given text or discourse: a frequent 'popping-up' and then rapid disappearing of different elements on the scene may indicate some unwanted frequent interruptions of the flow of discourse, and thus a lack of cohesion.

For Czech, we are now working on a project based on the data available in the Prague Dependency Treebank.Thanks to such a richly annotated corpus, we basically have at our disposal all information we need for an application of our rules for activation assignment: the underlying sentence representation with restored (superficial) deletions as well as with part-of-speech information, the Topic-Focus assignment (via the TFA attribute with values contextually-bound and contextually non-bound) and coreferential chains for nominal and pronominal realization of referential expressions. The task we face now is to implement the activation algorithm on (selected but full) documents, to visualize the 'activation' diagram, and, most importantly, to evaluate the results to see whether we achieve what we have envisaged.

4 Summary

Corpus annotation offers a most useful support for natural language processing, it is a irreplaceable resource of linguistic information for the build-up of grammars, and, most importantly, it provides an invaluable test for linguistic theories standing behind the annotation schemes. One of the important features is that in corpus annotation it is possible to take into account not only the surface shape of the sentence but even more importantly the underlying sentence structure: such an annotation may ellucidate phenomena hidden on the surface but unavoidable for the representation of the meaning and functioning of the sentence. The aim of our contribution was to indicate some of the possibilities an annotated corpus offers in the domain of information structure of the sentence and discourse analysis.

Note
Illustrations in Sect. 3.5 are taken from Milan Kundera's book Ignorance (translated from the French original by Linda Asher, Harper/Collins Publishers, New York, 2002.

Acknowledgements

This paper was written under the support of the grant of the Czech Republic Grant Agency P406/12/0658 and has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

Bibliography

Boguslavsky, I., I. Chardin, S. Grigorieva, N. Grigoriev, L. Iomdin, L. Kreidlin & N. Frid. 2002. Development of a dependency treebank for Russian and its possible applications in NLP. In: *Proceedings of LREC*, 852-856.

Brants, S., S. Hansen, W. Lezius & G. Smith. 2002. The TIGER treebank. In Hinrichs E. and Simov K. (eds.), *Proceedings of the First Workshop on Treebanks and Linguistic Theories* (TLT 2002). Sozopol, Bulgaria.

Firbas, J. 1959. Thoughts on the Communicative Function of the Verb in English, German and Czech. *Brno Studies in English*, Brno.

Firbas, J. 1992. *Functional sentence perspective in written and spoken communication.* Cambridge/London: Cambridge - London University Press.

Hajič, J. 1998. Building a syntactically annotated corpus:The PragueDependency Treebank. In: Hajičová E. (ed) *Issues of Valency and Meaning*. Studies in Honour of Jarmila Panevová, Prague: Karolinum, 106-132.

Hajič, J., J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský & M. Ševčíková-Razímová. 2006. Prague Dependency Treebank 2.0. CD-ROM. Linguistic Data Consortium, Philadelphia, PA, USA. LDC Catalog No. LDC2006T01 URLhttp://ufal.mff.cuni.cz/pdt2.0/

Hajičová, E. 1993. *Issues of sentence structure and discourse patterns*. Prague:Charles University.

Hajičová, E. 1997. Topic, focus and anaphora. Paper presented at the 16th International Congress of Linguists CIL 16, Paris (*Abstracts of the Congress*, p. 109).

Hajičová, E. 2003a. Contextual boundness and discourse patterns. Paper presented at the 17th International Congress of Linguists CIL 16, Prague (*Abstracts of the Congress*, p. 388).

Hajičová, E. 2003b. Aspects of Discourse Structure. In: Menzel, W. & C. Vertan (eds.) *Natural Language Processing between Linguistic Inquiry and System Engineering* Iasi, 47-56.

Hajičová, E. 2007. The Position of TFA (Information Structure) in a Dependency Based Description of Language. Invited paper for the *3rd MTT conference*, Klagenfurt. **(**http://meaningtext.net/mtt2007/proceedings/

Hajičová, E. 2012. Contextual boundness and discourse patterns. Presented at the *Symposium on Knowledge and Discourse,* Barcelona, April 2012, to be printed as a chapter in the collective volume of invited contributions.

Hajičová, E., J. Mírovský & K. Brankatschk. 2011. A contrastive look at information structure: A corpus probe. *6th Congres de la Societe Linguistique Slave*, Aix-en-Provence, 1-3.September, Univ. de Provence, 47-51.

Hajičová, E. , B. Partee & P. Sgall. 1998 .*Topic-Focus Articulation, Tripartite Structures and Semantic Content*, Dordrecht: Kluwer Academic Publishers.

Hajičová, E. & P. Sgall. 2004. Degrees of Contrast and the Topic-Focus Articulation. In: Steube, A. (ed). *Information Structure – Theoretical and Empirical Aspects.* Berlin - New York: Walter de Gruyter, 1-13.

Hajičová, E. & P. Sgall. 2006. Corpus annotation as a test of a linguistic theory. In *Proceedings of LREC 2006,* 879-884.

Hajičová, E. & J. Vrbová. 1982. On the role of the hierarchy of activation in the process of natural language understanding. In: Horecký J., ed. (1982), Coling 82 – *Proceedings of the Ninth International Congress of Computational Linguistics,* Prague – Amsterdam.

Koktová, E. 1999. *Word-Order Based Grammar*. Berlin: Mouton De Gruyter.

Marcus, M., G. Kim, M.A.Marcinkiewicz M, R. MacIntyre, A. Bies, M. Ferguson, K. Katz & B. Schasberger., 1994. The Penn Treebank: annotating predicate argument structure. In *Proceedings of the human language technology workshop*. Morgan Kaufmann Publishers Inc, 1994.

Marcus, M., B. Santorini & M-A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics*, 19(2):313-330.

Mel'čuk, I. A.. 2001. *Communicative Organization in Natural Language: The Semantic Communicative Structure of Sentences*. Amsterdam/Philadelphia: John Benjamins.

Mikulová, M., A. Bémová, J. Hajič, E. Hajičová, J. Havelka , V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá & Z. Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. *Annotation manual*. Tech. Report 30 ÚFAL MFF UK. Prague.

Mille, S., L. Wanner & A. Burga. 2012. Treebank annotation in the light of the Meaning-Text Theory. To be published in *Linguistic Issues in Language Technology*.

Miltsakaki, E., L. Robaldo, A. Lee & A. Joshi. 2008. Sense Annotation in the Penn Discourse Treebank. In: Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing,* Berlin/Heidelberg: Springer, 275–286.

Panevová, J. 2003. O jednom typu kauzativní konstrukce v češtině [On one type of causative constructions in Czech]. In: Banyś, W. L. Bednarczuk & K. Polański (eds.), *Etudes Linguistique Romano-Slaves* offertes *à Stanisław Karolak*. Cracovie: Oficyna Wydawnicza „Edukacja“, 379-385.

Panevová, J. 2004. Všeobecné aktanty očima Pražského závislostního korpusu [General actants in Prague Dependency Treebank]. In: Karlík, P. (ed.), *Korpus jako zdroj dat o češtině.* Brno: Masarykova univerzita, 41-46.

Panevová, J. 2008. České konstrukce tzv. slovanského akuzativu s infinitivem [Czech constructions of the so-called accusative with infinitive]. *Slovo a slovesnost* 69: 163 - 175.

Panevová , J. 2011. On the Syntax and Semantics of Czech Infinitival Constructions: A Case Study. In: *Slovo i jazyk*. Sbornik statej k vosmidesjatiletiju akademika Ju. D. Apresjana. Moskva: Jazyki slavjanskich kul'tur, 541 – 551.

Panevová, J. & M. Ševčíková. 2011. Počítání substantiv v češtině (Poznámky ke kategorii čísla) [Counting of nouns in Czech (Notes on the category of number)]. *Slovo a slovesnost* 72,: 163-176.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi A. & B. Webber. 2008. The Penn Discourse Treebank 2.0. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

Rysová, K. 2011a. The unmarked word ordedof free verbal modifications in Czech (with the main reference to the influence of verbal valency in the utterance. In. *44th Meeting of SLE 2011, Book of abstracts*, Logrono, 277-278.

Rysová, K. 2011b. The unmarked word order of inner participants, with the focus on the systemin ordering of Actor and Patient. In: Gerdes, K., E. Hajičová E. & L. Wanner (eds.) *Int. Conference on Dependency Linguistics* (Depling 2011), Barcelona, 183-192.

Sgall, P. 1979. Towards a definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics* 31:3-25; 32, 1980:24-32; printed in *Prague Studies in Mathematical Linguistics* 78, 1981, 173-198.

Sgall, P., E. Hajičová E. & J. Panevová. 1986., *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Prague:Academia and Dordrecht: Reidel.

Veselá, K., N. Peterek N. & E. Hajičová. 2003. Topic-Focus articulation in PDT: Prosodic characteristics of contrastive topic. *The Prague Bulletin of Mathematical Linguistics* 79-80:5-22.

Zikánová, Š. 2006. What do the data in PDT say about systemic ordering in Czech? *The Prague Bulletin of Mathematical Linguistics* 86:39-46.

Zikánová, Š., M. Týnovský M. & J. Havelka. 2007. Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. *The Prague Bulletin of Mathematical Linguistics* 87:61-70.

Zikánová, Š. & M. Týnovský. 2009. Identification of Topic and Focus in Czech: Comparative Evaluation on Prague Dependency Treebank. In: *Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure*. Formal Description of Slavic Languages 7, Frankfurt am Main: Peter Lang,, 343-353.