WHAT WE HAVE LEARNED FROM COMPLEX ANNOTATION OF TOPIC-FOCUS ARTICULATION IN A LARGE CZECH CORPUS

[Qu'est-ce que nous avons appris de l'annotation complexe en articulation en thème – rhème dans un grand corpus du tchèque]

Eva HAJIČOVÁ
Charles University in Prague

Abstract (En): After a short summary of the theory of Topic-Focus Articulation (TFA) the present contribution documents on several examples illustrating the annotation of the basic features of TFA on a large corpus (the Prague Dependency Treebank) that corpus annotation brings an additional value to the corpus if the following two conditions are being met: (i) the annotation scheme is based on a sound linguistic theory, and (ii) the annotation scenario is carefully (i.e. systematically and consistently) designed. Such an annotation is important not only for the surface shape of the sentence but even more for the underlying sentence structure: it may elucidate phenomena hidden on the surface but unavoidable for the representation of the meaning and functioning of the sentence.

Résumé (Fr): Après un bref résumé de la théorie de Topic-Focus Articulation (TFA), la présente étude démontre, à l'aide de plusieurs exemples illustrant l'annotation de principaux traits de TFA sur un large corpus (the Prague Dependency Treebank), que l'annotation du corpus apporte une valeur ajoutée au corpus, si deux conditions sont réunies : (i) le schéma de l'annotation est basé sur une théorie linguistique solide, (ii) le procédé d'annotation est établi avec soin (c'est-à-dire de façon systématique et cohérente). Une telle annotation est importante non seulement pour la structure de surface de la phrase mais encore davantage pour la structure phrastique sous-jacente, car elle est susceptible de mettre en évidence les phénomènes cachés au niveau de la structure de surface, mais incontournables lors de la représentation du sens et du fonctionnement de la phrase.

Keywords (En): Topic-Focus Articulation; corpus annotation; Prague Dependency Treebank.

1. Motivation

Corpus annotation may bring an additional value to the corpus if the following two conditions are being met: (i) the annotation scheme is based on a sound linguistic theory, and (ii) the annotation scenario is carefully (i.e. systematically and consistently) designed. The usefulness of annotated data for further linguistic research 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 by the teams at the University of Pennsylvania, the Tiger Treebank for German (BRANTS et al., 2002), or the Prague Dependency Treebank for Czech (HAJIČ, 1998; HAJIČ et al., 2006).

Corpus annotation is not a self-contained task. It offers a most useful support for natural language processing, it is an 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 it is possible to take into account in corpus annotation not only the surface shape of the sentence but even more importantly the underlying sentence structure: such an annotation may elucidate phenomena

hidden on the surface but unavoidable for the representation of the meaning and functioning of the sentence.

In the present contribution, we first give (in Sect. 2) a brief overview of the underlying theory of Topic Focus Articulation we subscribe to (abbreviated in the sequel as TFA) and we outline (in Sect. 3) how this theory is reflected in the Czech corpus annotation of the Prague Dependency Treebank (PDT). In Sect. 4 two linguistic hypotheses are presented to illustrate how theoretical hypotheses can be tested on language corpora. Some lessons learned in the course of our TFA annotation are given in Sect. 5 concerning (i) contrastive topic, (ii) focalizers and their scope, and (iii) some notes on the relationships of passivization, TFA and the use of indefinite article in English. Sect. 6 summarizes our experience.

2. Linguistic theory and corpus annotation

2.1. Underlying theory of TFA in a nutshell

Issues connected with the articulation of sentence with regard to their communicative rather than surface syntactic structure (functions such as subject, object, predicate) had been brought into the foreground of linguistic studies in Prague since Vilém Mathesius' first papers on the topic (e.g. MATHESIUS, 1929; 1939); as the Czech term he used (*aktuální členění větné*) was not directly translatable into English, Jan Firbas – on the advice of Josef Vachek (as acknowledged in FIRBAS, 1992, p. xii) and apparently inspired by Mathesius' use of the German term *Satzperspektive* in his fundamental paper from 1929 coined the term *functional sentence perspective* (FSP); German researchers in this field often speak about *Thema-Rhema Gliederung* and the prominent British linguist M.A.K. Halliday introduces the term *information subsystem* (HALLIDAY, 1967; 1967-68) or *information structure* (reflecting the given-new strategy) distinguishing it from *thematic structure*; the former term has been re-invented and is generally used nowadays by several modern linguists).

The above-mentioned terminological differences often refer to some notional distinctions: the dichotomy reflected in the name of our theory, namely *Topic-Focus Articulation* (TFA) is not a mere "translation" or "rephrasing" of the terms used in FSP but indicates certain differences in the starting points:

(i) FIRBAS (1964) specifies the *theme* as the element (or elements) carrying the lowest degree of communicative dynamism within the sentence. This specification implies that every sentence *contains* an item with the lowest degree of communicative dynamism (CD), and thus would exclude the existence of sentences without a theme (so-called topicless sentences). It must be added that FIRBAS (1992) modifies his definition of theme by saying that in the absence of theme, the lowest degree of CD is carried by the first element of non-theme (in this reformulation he refers to Sgall's objection against Firbas' original definition of theme made at a FSP conference in Sofia in 1976).

(ii) Accepting Firbas' assumption that every item in the sentence carries a certain degree of CD, should not mean, however, that the notion of a bipartition (the focus of a sentence conveys an information *ABOUT* its topic) can be abandoned; an important argument for the necessity of a recognition of such a bipartition is the analysis of negation (see HAJIČOVÁ, 1972; 1973; 1984).

(iii) The so-called *factors* of linear arrangement, prosody, semantics and contexts as discussed by Firbas and his followers are not just four 'factors' of FSP but they fundamentally differ in their nature: the first two (word order and prosody) belong to the means of expression of information structure and the other two (semantics and context) to its functional layers.

(iv) Most importantly, TFA is understood as a structure belonging to the *underlying, deep structure* of sentences (tectogrammatics, literal meaning) because the TFA structure is semantically relevant.

In the theory of TFA, the underlying relation between topic and focus is based on the relation of 'aboutness': the topic is understood as a specification of "what we are talking about" and the focus as „what we are saying about topic". In other words, the speaker communicates something (the Focus of the sentence) about something (the Topic of the sentence); this relation can be schematically captured as:

F(T): the Focus holds about the Topic

and in case of negation:

~F(T): (in the prototypical case) the Focus does not hold about the Topic

The pragmatic background of this opposition is the cognitive dichotomy of *given* (*old*) versus *new*, but the two oppositions are not identical as can be illustrated by examples (1) and (2):

(1) John and Mary entered the dining-room. They first went to the window …
(2) Mary called Jim a Republican. Then he insulted HER.
(2') Mary called Jim a Republican. Then he INSULTED her.

In the second sentence in (1), 'the window' refers by no doubt to the window of the dining-room mentioned in the first sentence, thus to the cognitively 'given' or 'old' information, but this sentence is 'about' John and Mary and it says about them that what they did, was to go to the window (rather than to the table, to the other door, etc., or just to look around in astonishment …). In the second sentence in (2) as well as in (2'), both 'he' and 'her' are cognitively 'given', the pronouns refer to John and Mary, respectively, but only (2') is 'about' John and Mary and says that the relation between them was an insult (thus somehow implying that calling somebody a Republican is not, or need not be an insult), while the second sentence in (2) is about Jim's insult saying that this insult was directed against Mary (thus implying that calling somebody a Republican is understood as an insult). The difference in the TFA of the second sentences in (2) and (2') is

indicated, in spoken language, by the position of the intonation center, denoted here by capitals.

2.2 Semantic relevance of TFA

Related to the attempt of the TFA proponents at an explicit description of the dichotomy of topic and focus and thus of its integration into a formal description of language as early as at the beginning of the 1960's, was the consideration which level of language description TFA belongs to. Since then it is commonly accepted in the linguistic literature on semantic aspects of sentence structure that TFA is semantically relevant, even if truth conditions are taken into account. We will not recapitulate the discussions of these issues here, but we want to recall them by re-quoting some of the examples hinting at the semantic relevance of TFA as they appeared in some writings: from Chomsky's observations not at the time fully reflected in his model, through Lakoff's sentences that had led him to formulate an alternative model of transformational grammar, so-called generative semantics, with different semantic representations for his (a) and (b), to Rooth's very influential (esp. in the circles of semanticists, but, fortunately also among linguists of different streams); to document that these observations were duly documented and analyzed by the theory of TFA, we also quote three of the many examples adduced by Sgall and his colleagues.

(3) (a) Everybody in this room knows at least two LANGUAGES.
(b) At least two languages are known by everybody in this ROOM. (CHOMSKY, 1957;1965)
(4) (a) Many men read few BOOKS.
(b) Few books are read by many MEN. (LAKOFF, 1971)
(5) (a) I only introduced BILL to Sue.
(b) I only introduced Bill to SUE. (ROOTH, 1985)
(6) (a) Londoners are mostly at BRIGHTON.
(b) At Brighton, there are mostly LONDONERS. (SGALL, 1967)
(7) (a) I work on my dissertation on SUNDAYS.
(b) On Sundays, I work on my DISSERTATION.
(8) (a) English is spoken in the SHETLANDS.
(b) In the Shetlands, ENGLISH is spoken. (SGALL et al., 1986)

As can be seen from the above mentioned examples, the means of expression of TFA are multifarious and should be distinguished from the uniform semantic function (this is the main reason why we do not agree with the subsumption of the function of the dichotomy and the means of its expression under 'four' factors as if they were of the same notional category). Let us mention here, again very briefly, four such means:

(i) Surface order of words, a most visible means esp. in languages that do not have a grammatically fixed word order. In the first writings, esp. from the tranformationalist circle, the authors assumed that it is the order of quantifiers that is responsible for the semantic differences (see (3) and (4) above, for examples taken from English), or perhaps the difference between active and passive constructions (but compare the possible Czech translation of (3) and (4), in which no passivization is necessary to make the change in the order). However, as illustrated by (7) and (8), the presence of a quantifying expression is not crucial.

(ii) In spoken language, the most important means of expressing the difference in TFA is the sentence prosody including the placement of the intonation center; in our more recent work with spoken language corpora, the F0 characteristics of the curve was attested as to be a marker of a "contrastive topic" (VESELÁ et al., 2003). HALLIDAY (1967) adduces a brilliant example of the importance of the placement of the intonation center, pointing out the necessity to pronounce the warnings at the bottom of an elevator in London underground stations (9)(a) with the (normal) placement of the intonation center at the end and comparing it with the inadequacy of (9)(b) with its funny interpretation "you should carry a dog". The greater was our surprise to see that in the newly opened modern underground stations in London at the time of the centenery celebrations was (9)(c), which evoked the same funny intepretations as Halliday's (9)(b). The intention of the authors of this change was – perhaps in addition to make the instruction shorter and thus more urgent – to read the instruction with the non-normal position of the intonation center at the beginning of the sentence, as in (9)(d).

(9) (a) Dogs must be CARRIED.
(b) DOGS must be carried. (HALLIDAY, 1967)
(c) Carry DOGS. (a warning in London underground, around 2000)
(d) CARRY dogs.

(iii) Another possible means for expressing TFA are specific syntactic constructions such as the *it*-clefts (in contrast to *wh*-clefts) in English, cf. (10)(a); in Czech, the sentence – unless we do want to make it more emphatic, with a subjective order and the placement of the intonation center on the subject in the front position – can be translated as (10)(b), with the same TFA.

(10) (a) It was JOHN who talked to few girls about many problems.
(10) (b) S málo děvčaty mluvil o mnoha problémech HONZA.

(iv) A specific device is morphemic means indicating which element of the sentence is its topic or focus, such as the particles *ga* and *wa* in Japanese and similar morphemic means used in some other languages such as Yukaghir, Tagalog etc.

3. The reflection of the TFA theory in corpus annotation

The Prague Dependency Treebank (PDT, see HAJIČ et al., 2006; MIKULOVÁ et al., 2006) is an annotated (electronic) collection of Czech texts with a mark-up on three layers: (i) morphemic, (ii) surface shape, and (iii) underlying (tectogrammatical) incl. underlying dependency relations such as Actor, Patient, Addressee, Temporal, Local, Manner etc. and values concerning TFA. The current version (annotated on all three layers of annotation) contains 3168 documents with 49442 sentences and 833357 occurrences of forms. In addition to these three layers, the current annotation also covers some basic relations of textual coreference and fundamental discourse relations.

Each node of the dependency tree representing the sentence on the tectogrammatical (underlying) level is assigned one of the three possible values of a special TFA attribute, namely *t* for a contextually bound non-contrastive node, *c* for a contextually bound contrastive node and *f* for a contextually non-bound node. These values serve then as a basis for the bipartion of the sentence into Topic and Focus; an algorithm of such a bipartition was formulated and tested on the whole PDT collection (see below).

4. Annotated corpus used for testing of linguistic hypotheses

As an illustration of what possibilities a consistent and systematic annotation of a text corpus offers for linguistic theory, we present in our contribution two examples.

Hypothesis A1:

The global division of the sentence into its TOPIC (what the sentence is about) and its FOCUS (what is said about the topic) can be made on the basis of contextual boundness.

Some first formulations of the steps of a possible algorithm for a (global) division of a sentence into its Topic and Focus based on this hypothesis are given in SGALL (1979; see also SGALL et al., 1986: 216f). The original algorithm was later implemented and then tested on the whole of PDT and the results were reported in HAJIČOVÁ, HAVELKA and VESELÁ (2005).

The basic steps of the algorithm are as follows:

(a) if the main verb carries *f*, it belongs to Focus (F); else, it belongs to Topic;

(b) all the nodes directly dependent on the main verb and carrying *t* belong to Topic, together with all nodes depending on them;

(c) all the nodes directly dependent on the main verb and carrying *f* belong to Focus, together with all nodes depending on them;

(d) if the main verb carries *t* and all nodes directly depending on the main verb carry also *t*, then follow the rightmost edge leading from the main verb to the first node(s) on this path carrying the value *f*; this/these node(s) and all the nodes depending on it/them belong to Focus.

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 was contextually non-bound and thus belongs 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 FSP theory, see Firbas on the transitional character of the verb) that in Czech the boundary between Topic and

Focus is in the prototypical case signalled 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', 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 and 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 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, in press).

Hypothesis A2:

In the focus part of the sentence the complementations of the verb (be they arguments or adjuncts, in the sense of underlying, tectogrammatical dependency relations) follow a certain canonical order (not necessarily the same for all languages).

Before the *A2* hypothesis was formulated, a series of psycholinguistic experiments (with speakers of Czech, German and English) was carried out to establish a tentative ordering. However, the PDT offers a richer and more consistent material for its testing as the underlying dependency relations within the sentence are annotated and the appurtenance of the elements into Focus can be determined by the implemented TFA algorithm (see *A1* above). 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).

5. Lessons learned

In addition to the two examples given in the previous Section, the manual annotation itself and the annotated corpus material have provided some other interesting observations and suggestions.

5.1 Contrastive topic

The original formulation of the TFA theory worked with the notion of contextual boundness, which served as the basis for the recognition of the Topic-Focus dichotomy. However, thanks to 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 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 (see above about F0) and by the use of a long form of pronoun in the topic position in Czech, see (11), with the intonation center marked by capitals.

(11) 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 (11), *jeho* is the long form of Acc.sing. of the pronoun 'on' (he), the short form of this pronoun being *ho* as in (12).

(12) 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 and in accordance with it to introduce a third value of the TFA attribute in the annotation scheme of PDT, namely the value *c* (HAJIČOVÁ et al., 2007).

5.2 Focalizers and their scope

FIRBAS (1957) observed a rhematizing function of the adverb *even*; in his later paper (FIRBAS, 1959) he speaks about a class of intensifying elements. The relation between some specific class of sentence elements to TFA was also mentioned by SGALL (1967) in connection with examples with quantifiers (such as *mostly)*. DANEŠ (1985) distinguishes in this connection direct restrictors (*jen 'only'*), indirect (*vyjma 'except for'*) and contextualizers *(také 'also', a přece 'and still'*). In all these writings, an observation is made that there is a class of sentence elements that is closely related to the indication of the focus of the sentence.

Connected to this, is the relation between the semantic scope of negation and topic/focus articulation; let us mention here already VACHEK (1947), the analysis of the negative particle *niet* in Dutch (KRAAK, 1966), the analysis of the German

nicht (ZEMB, 1968) and our systematic attention paid to the relation between TFA and the semantic scope of negation with the relevant consequences for the relation of presupposition in HAJIČOVÁ (1973; 1975) and our comparison of the scope of negation with the function of focalizers in HAJIČOVÁ (1995).

FIRBAS (1959: 53) characterizes his intensifying elements as follows: "intensifying elements […] are, as it were, superimposed on the sentence structure, considerably *changing* its FSP by rhematizing (frequently even turning into rheme proper) the element to which they are made to refer". Our position is different: the focalizer (prototypically, by its word-order position and also with regard to the placement of the intonation center) just *indicates* which element(s) of the sentence are its focus (but see below); the TFA of the sentence is a part of the underlying structure of the sentence (its meaning), the position of the focalizer and the prosody of the sentence are the outer forms (expression) of this function and as such do not change the function.

Interesting examples are also those sentences that contain two focalizers, which need not even be in separate clauses, cf. (13) in English and its Czech counterpart in (13'), and also (14) with negation and a focalizer.

(13) (Preceding context: Who has sent just a postcard even to John?) MARY has sent just a postcard even to him.
(13') (Kdo poslal i Honzovi jenom pohlednici?) I jemu poslala jen pohlednici MARIE.

(14) Jen dobré srdce bezmocným nepomůže.
Only good heart the-helpless-Dat will-not-help.

Sentences (13) and (14) document that the class called 'focalizers' need not be only indicators of focus. In HAJIČOVÁ, PARTEE and SGALL (1998, Sect. 6.3), the sentences quoted here as (15) and (16) are given used in the contexts (again, the placement of the intonation center is indicated by capitals): *Who criticized even MOTHER TERESA as a tool of the capitalists?* and *Is there a film only JIM liked?*, respectively.

(15) JOHN criticized even Mother Teresa as a tool of the capitalists.
(16) Only Jim liked AMADEUS.

These observations, first, have served as a further argument for the introduction of the notion of contrastive topic (see the use of the long form of the Czech pronoun *jemu* in (13')) and to the suggestions to differentiate between a global focus (*JOHN, AMADEUS* in (15) and (16), respectively) and a focus of a focalizer (*Mother Teresa, Jim* of focalizers *even* and *only* in the same sentences). At the same time, the analysis of the large PDT corpus has indicated that the class of focalizers is bigger than originally (and usually) assumed and that it contains such Czech particles that can be translated into English as *also, alone, as well, at least, even, especially, either, exactly, in addition, in particular, just, merely, only, let alone, likewise, so much as, solely, still/much less, purely, too* etc.

To summarize our observations presented in Sect. 5.2, there is a special class of particles that have a specific position in the TFA of the sentence; these particles have some common features with negation. The so-called focalizer can occur also in the topic of the sentence and there can be more than a single focalizer in a sentence. It is therefore necessary to distinguish between the focus of the whole sentence and the focus of a focalizer. The scope of a focalizer has important consequences for the semantic interpretation of the sentence.

5.3 Passivization, TFA and indefinite article in English

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.

Looking at these cases in more detail (HAJIČOVÁ et al., 2011), 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 E. 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 (17) and (18) and the sentence elements in question are printed in italics.

(17) (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.

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

(18) (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.

(18') 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. (19).

(19) *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.)

(19') 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. (20).

(20) (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 .

(20') 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.

6. Summary

Every linguistic theory needs testing and evaluation; annotated text and spoken corpora are suitable (and hitherto unsurpassed) tools for that purpose. The results are invaluable: in our contribution we have tried to document that a consistent and systematic testing brings new findings, and these findings then may lead to additions or modifications of the theory.

Acknowledgement

The work on the final version of this contribution was supported by the grant of the Czech Grant Agency GAČR P406/12/0658.

REFERENCES

BRANTS Sabine, Stefanie DIPPER, Silvia HANSEN, Wolfgang LEZIUS, and George SMITH (2002), The TIGER treebank, in: Erhard HINRICHS and Kiril SIMOV (eds.), *Proceedings of the First Workshop on Treebanks and Linguistic Theories* (TLT 2002), Sozopol, Bulgaria.

CHOMSKY Noam (1957), *Syntactic Structures*, The Hague, Mouton.

CHOMSKY Noam (1965), *Aspects of the Theory of Syntax*, Cambridge, Mass., The M.I.T. Press.

DANEŠ František (1985), *Věta a text*, Praha, Academia.

FIRBAS Jan (1959), Thoughts on the Communicative Function of the Verb in English, German and Czech, *Brno Studies in English*, Brno.

FIRBAS Jan (1964), On defining the theme in functional sentence perspective, *Travaux Linguistique de Prague* 1, Prague, 267-280.

FIRBAS Jan (1992), *Functional sentence perspective in written and spoken communication*, Cambridge/London, Cambridge - London University Press.

FIRBAS Jan (1957), K otázce nezákladových podmětů v současné angličtině. Příspěvek k teorii aktuálního členění větného, *ČMF* 39, p. 22-42; p. 165-173. (An abbreviated and modified English version of this contribution was published as Non-thematic subjects in Contemporary English, TLP 2, 1966, p. 239-256.)

HAJIČ J., PANEVOVÁ J., HAJIČOVÁ E., SGALL P., PAJAS P., ŠTĚPÁNEK J., HAVELKA J., MIKULOVÁ M., ŽABOKRTSKÝ Z. and M. ŠEVČÍKOVÁ-RAZÍMOVÁ (2006), Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, Philadelphia, PA, USA. LDC Catalog No. LDC2006T01 URL<http://ufal.mff.cuni.cz/pdt2.0/>

HAJIČ Jan (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, Prague, Karolinum, p. 106-132.

HAJIČOVÁ Eva, MÍROVSKÝ Jiří and BRANKATSCHK Katya (2011), A contrastive look at information structure: A corpus probe. *6th Congrès de la Société Linguistique Slave*, Aix-en-Provence, 1-3 September, Univ. de Provence, pp. 47-51.

HAJIČOVÁ Eva, PARTEE Barbara and SGALL Petr (1998), *Topic-Focus Articulation, Tripartite Structures and Semantic Content*, Dordrecht, Kluwer Academic Publishers.

HAJIČOVÁ Eva, SGALL Petr and VESELÁ Kateřina (2007), Contextual Boundness and Contrast in the Prague Dependency Treebank, In: *Language Context and Cognition: Interfaces and Interface Conditions*, ed. Andreas Spaeth, Berlin – New York, Walter de Gruyter, p. 231- 243.

HAJIČOVÁ Eva (1984), Presupposition and allegation revisited. *Journal of Pragmatics* 8, p. 155-167; amplified as "On presupposition and allegation" in: *Contributions to functional syntax, semantics and language comprehension*, ed. by SGALL P., Amsterdam, Benjamins – Prague, Academia, p. 99-122.

HAJIČOVÁ Eva (1972), Some Remarks on Presuppositions. *The Prague Bulletin of Mathematical Linguistics* 17, p. 11-23.

HAJIČOVÁ Eva (1973), Negation and topic vs. comment. *Philologica Pragensia* 17, p. 18-25.

HAJIČOVÁ Eva (1975), *Negace a presupozice ve významové stavbě věty*. Praha, Academia.

HAJIČOVÁ Eva (1995), Postavení rematizátorů v aktuálním členění věty. *Slovo a slovesnost* 56, p. 241-251.

HALLIDAY M. A. K. (1967), *Intonation and Grammar in British English*. The Hague, Mouton.

HALLIDAY M.A.K. (1967-8), Notes on transitivity and theme in English. *Journal of Linguistics* 3 (1967), p. 37-81, p. 199-244; 4 (1968), p. 179-215.

KOKTOVÁ Eva (1999), *Word-Order Based Grammar*. Berlin, Mouton De Gruyter.

KRAAK Albert (1966), *Negative Zinnen. Een Methodologishe Analyse*, Hilversum.

LAKOFF George (1971), On generative semantics, in: Steinberg and Jakobovits (1971), p. 232-296.

MARCUS Mitch, SANTORINI B. and MARCINKIEWICZ M-A. (1993), Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics*, 19:2, p. 313-330.

MARCUS Mitch, KIM G., MARCINKIEWICZ M. A., MACINTYRE R., BIES A., FERGUSON M., KATZ K., and SCHASBERGER B. (1994): The Penn Treebank: annotating predicate argument structure, in: *Proceedings of the human language technology workshop*. Morgan Kaufmann Publishers Inc.

MATHESIUS Vilém (1929), Zur Satzperspektive im modernen Englisch, *Archiv für das Studium der neueren Sprachen und Literaturen* 155, p. 202-210.

MATHESIUS Vilém (1939), O tak zvaném aktuálním členění větném, *Slovo a slovesnost* 5, p. 171-174; translated as: On information-bearing structure of the sentence, in: S. KUNO (ed.): *Harvard Studies in Syntax and Semantics*, 1975, p. 467-480

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.

ROOTS Mats (1985), *Association with Focus*. PhD Thesis, Univ. of Massachusetts, Amherst.

RYSOVÁ Kateřina (2011a), The unmarked word ordedof free verbal modifications in Czech (with the main reference to the influence of verbal valency in the utterance, in: ARISTA Javier Martin (eds). *44th Meeting of SLE 2011, Book of abstracts*, Logrono, p. 277-278.

RYSOVÁ Kateřina (2011b), The unmarked word order of inner participants, with the focus on the systemin ordering of Actor and Patient, in: GERDES Kim, HAJIČOVÁ E. and WANNER L. (eds), *Int. Conference on Dependency Linguistics* (Depling 2011), Barcelona, p. 183-192.

SGALL Petr (1967), Functional Sentence Perspective in a generative description of language. *Prague Studies in Mathematical Linguistics* 2, Prague, Academia, p. 203-225.

SGALL Petr (1979), Towards a definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics* 31, p. 3-25; 32, 1980, p. 24-32; printed in *Prague Studies in Mathematical Linguistics* 78, 1981, p. 173-198.

SGALL Petr, HAJIČOVÁ Eva and PANEVOVÁ Jarmila (1986), *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Prague, Academia and Dordrecht, Reidel.

STEINBERG Danny D. and JAKOBOVITS Leon A. (eds., 1971), *Semantics*, Cambridge, Mass., The M.I.T. Press

VACHEK Josef (1947), *Obecný zápor v angličtině a češtině. Příspěvky k dějinám řeči a literatury anglické* (Prague Studies in English), Vol. 6, Prague, Filosofická fakulta Univerzity Karlovy, p. 7-72.

VESELÁ Kateřina, PETEREK, Nino and HAJIČOVÁ Eva (2003), Topic-Focus articulation in PDT: Prosodic characteristics of contrastive topic. *The Prague*

Bulletin of Mathematical Linguistics 79-80, Praha, Univerzita Karlova, p. 5-22.

ZEMB Jean-Marie (1968), *Les structures logiques de la proposition allemande*. Paris, O.C.D.L.

ZIKÁNOVÁ Šárka and TÝNOVSKÝ Miroslav (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, Peter Lang, Frankfurt am Main, Germany, p. 343-353.

ZIKÁNOVÁ Šárka, TÝNOVSKÝ Miroslav and HAVELKA Jiří (2007), Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations, in: *The Prague Bulletin of Mathematical Linguistics*, No. 87, p. 61-70.