What We Are Talking about and What We Are Saying about It

Eva Hajičová

Charles University
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague, Czech Republic
hajicova@ufal.mff.cuni.cz

Abstract. In view of the relationships between theoretical, computational and corpus linguistics, their mutual contributions are discussed and illustrated on the issue of the aspect of language related to the information structure of the sentence, distinguishing "what we are talking about" and "what we are saying about it".

1 Introduction

The name of the research domain of Computational Linguistics seems to be self-explanatory; however, there has always been a dispute what exactly 'computational' means (especially from the point of view of the relation between its theoretical and applied aspects and from the point of view of its supposedly narrowing scope due to the prevalent use of statistical methods). In addition, with the expansion of the use of computers for linguistic studies based on very large empirical language material, and, consequently, with the appearance of an allegedly new domain, corpus linguistics, a question has emerged what is the position of corpus linguistics with regard to computational linguistics.

After a summary of some of the issues related to the problem of 'how many linguistics there are' (Sect. 2), we briefly sketch in which respects the different 'linguistics' can mutually contribute to each other (Sect. 3). The main objective of our paper is to illustrate on an example of a linguistically based multi-layered annotation scenario (Sect. 4) and of a selected linguistic phenomenon, namely the information structure of the sentence (Sect. 5.1), how linguistic theory can contribute to a build-up of an integrated scenario of corpus annotation (Sect. 5.2) and, in the other direction, how a consistent application of such a scenario on a large corpus of continuous texts can provide a useful feedback for the theory (Sect. 5.3). In Section 6, some conclusions will be drawn from the personal experience with working with the given theory and scenario.

2 How Many Linguistics?

If the terms *computational linguistics* and *corpus linguistics* are understood rather broadly, as covering those domains of linguistics that are based on the use of computers and on the creation and use of corpora, respectively, then it can be seen that

A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 241–262, 2008.

the intersection of the two domains is very large. (Speaking of corpora, what we have in mind are corpora implemented in computers and patterned as data bases.) However, it is important to be aware also of a third domain that develops along with the two mentioned ones, and this is *theoretical linguistics*.

Certainly, there is no descriptive framework universally accepted; there are many different trends in linguistics, as there were a hundred years ago.[1] This diversity, which perhaps is even growing, offers certain advantages, among which there is the possibility of fruitful discussions. Different points of view help to throw light on problems discussed and to make choice between the available approaches or their parts. However, the diversity of views also constitutes a source of possible misunderstandings, especially if one of the various potential aims of research is seen as the only goal worth of serious studies, or as a goal of itself, standing higher than others.

If the different points of view and goals are soberly examined, then a highly effective collaboration of researchers working in the different domains can be achieved. It is important to look for reliable results, not deciding *a priori* whether they may be found in this or that trend, but rather basing the discussions on arguments. We do not understand it as appropriate to distinguish between "computational", "corpus" and "real" linguists, the more so if the latter were to be seen as those who avoid using computers for other aims than for the creation of large corpora and of search procedures, without using clear operational criteria for classifying the items to be described. The discussion on theoretical characterization of linguistic phenomena and the computerized checking of the adequacy of descriptive frameworks belong to fundamental goals in linguistics.

In the context of computational linguistics, dependency based grammar always has played an important role, competing with phrase structure (or transformation) based approaches. A framework of this kind offers a way to conceiving the core of language, based on prototypical phenomena, as patterned in a way that comes close to elementary logic, and thus to general human mental capabilities.[2]

3 Mutual Enrichment: Task of Corpus Annotation

From what has been said above, it is certainly significant to be aware of the requirements of a systematic, intrinsic collaboration (if not a symbiosis) of corpus oriented and computational linguistics with linguistic theory.

Several linguists still prefer to work without computers and computer corpora, or to avoid statistical methods, since these may appear as attempts to do without linguistic analyses, using just the outer "brute force". Nowadays, however, statistical methods do bring important results, thanks to factors such as their

[1] We are not concerned here with what is sometimes called hyphenated linguistics - socio-linguistics, ethno-linguistics, pragmalinguistics etc.

[2] On the other hand, contextually restricted rules are then needed for the handling of the large and complex periphery, containing secondary items of different levels, as well as all asymmetries between (underlying) sentence structure and morphemics (ambiguity, synonymy, irregularities, exceptions).

connection with different possibilities of automatic learning and of a computerized, more or less automatic search for appropriate classifications of linguistic phenomena on different layers.

Other researchers see an attractive goal, or even the center of all appropriate uses of computers in linguistics, in gathering large corpora with searching procedures.

Still others are aware of the fact that, along with the mentioned goals, there is also the need to use corpora for theoretical studies. According to their views, a linguist studying e.g. the system of tenses of the English verb should not only collect the occurrences of forms in *-ed*, *-ing*, etc., from a corpus and then select, comment and classify those of them that express tenses, but should also work with procedures that identify the forms of preterit, future, etc. and enable the researcher to start immediately to analyze their functions or their combinatorics, and so on. Similarly, it is of advantage to get at once all the occurrences of subjects, direct or indirect objects, etc. in a corpus. Such tasks require not only to assign part-of-speech (POS) annotations, but also to integrate syntactic annotations into the work with large corpora. As H. Uszkoreit ([1]) has put it: time has come for deep parsing, and thus, let us add, also for deep corpus annotation. A qualified choice between the existing theoretical approaches (or their parts and ingredients) is necessary to make it possible to use corpora effectively for the aims of theoretical linguistics, as well as of frameworks oriented towards pedagogical and other applications.

Such use of corpora in theoretical linguistic studies includes aims as the following:

(i) to offer new, substantially better conditions for most diverse kinds of research in linguistics itself as well as in neighboring domains ranging from theory of literature to information retrieval;

(ii) to check existing descriptive frameworks or their parts, having in mind improvements of their consistency, their enrichment or, in the negative case, the abandonment of falsified hypotheses;

(iii) on the basis of aligned corpora to compare descriptions of two or more languages, attempting at a formulation of procedures that would serve as sources for transfer components of translation systems or, as soon as the large multilingual lexical systems such as Wordnet become effectively usable, even as sources for the construction of an interlingua helping translate among whole groups of languages;

(iv) for all such and similar goals one of the most important ingredients is the search for suitable combinations of structural and statistically based procedures of most different kinds and levels, starting from an adequate linguistic background of a POS system with disambiguation; however, it is important to see the typologically determined differences between languages: if E. forms such as *give* (vs. *gives* or *gave*) are classified just as basic verb forms, without distinguishing their values of person and number, then the large set of tags used for a language with a rich morphology (as e.g. Czech, Russian, etc.) gives a much richer set of data (among which then

different ambiguities between morphemic cases and verb forms cause many difficulties); morphemic disambiguation is to be accompanied by procedures of syntactic annotation having their automatic and intellectual parts, the former being partly of a structural nature and partly stochastic.

4 A Concrete Example of a Linguistically Based Annotation Scheme: Prague Dependency Treebank

Prague Dependency Treebank (PDT; for an overall characterization see e.g. [2]) is an annotated collection of Czech texts, randomly chosen from the Czech National Corpus, with a mark-up on three layers: (a) morphemic, (b) surface shape "analytical", and (c) underlying (tectogrammatical). The current version (the description of which is publicly available on http://ufal.mff.cuni.cz/pdt2.0, with the data themselves available at LDC under the catalog No. LDC2006T01), annotated on all three layers, contains 3,165 documents (text segments mainly from journalistic style) comprising of 49,431 sentences and 833,195 occurrences of tokens (word forms and punctuation marks) - Figure 1 illustrates sentence annotation on three layers.

On the tectogrammatical layer, which is our main concern in the present paper, every node of the tectogrammatical representation (TGTS, a dependency tree) is assigned a label consisting of: the *lexical value* of the word, its '(*morphological*) *grammatemes*' (i.e. the values of morphological categories), its *'functors'* (with a more subtle differentiation of syntactic relations by means of *subfunctors*, e.g. 'in', 'at', 'on', 'under'), and the topic-focus articulation (TFA) attribute containing values for *contextual boundness* (for a motivation for the introduction of this value see below Sect. 5.1). In addition, some basic coreferential links (including intersentential ones) 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.

Dependency trees on the tectogrammatical layer are projective (unimportant exceptions aside), i.e. for every pair of nodes in which a is a rightside (leftside) daughter of b, every node c that is less (more) dynamic than a and more (less) dynamic than b depends directly or indirectly on b (where *indirectly* refers to the transitive closure of *depend*). This strong condition together with similar conditions holding for the relationship between dependency, coordination and apposition, makes it possible to capture the tectogrammatical representations in a linearized way. Projective trees thus come relatively close to linear strings; they belong to the most simple kinds of patterning.

In the annotation of PDT, we work also with (surface) analytic representation, a useful auxiliary layer, on which the dependency trees include nodes representing the function words and the tree reflects the surface word order. This combination allows for non-projective structures in cases such as *A neighbour came in, who told us this* (with the relative clause dependent on the subject noun). We assume that such cases can be described as surface deviations from the underlying word order (i.e. in a tectogrammatical representation corresponding

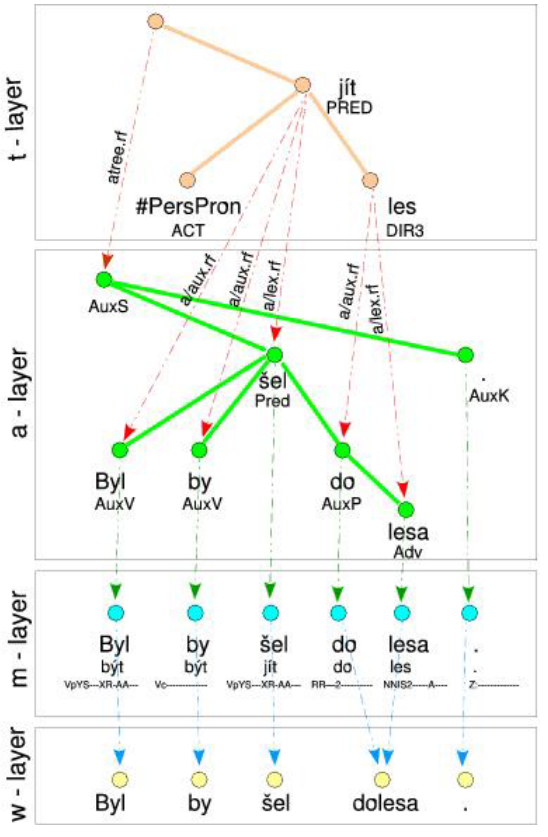

Fig. 1. Annotation layers of the Prague Dependency Treebank

to the example given above, the main verb is not placed between the subject and the dependent clause).

5 Illustration on a Concrete Linguistic Phenomenon

For our discussion of the mutual interlinking of theoretical linguistic description, corpus annotation and computational aspects, we have chosen the linguistic phenomenon of information structure as a universal feature of natural language pertaining to its function as a means of communication expressed in the surface shape of sentences in different ways, mostly dependent on the typological character of the language in question. A description of information structure (be it under the traditional terms of functional sentence perspective, theme-rheme articulation, topic and comment, or, as is the case in the theory we subscribe to, topic-focus articulation, TFA in the sequel) is nowadays regarded as a necessary part of language description in any linguistic theory, though the position within the framework and the detail in elaboration, the scope and depth of the

description differs from theory to theory. However, the different treatments of information structure share the underlying idea: a description of the structure reflecting the functioning of language in communication, which is different from the subject-verb-object structure (as described in any formalism)

5.1 The Phenomenon Under Scrutiny: Topic-Focus Articulation

The theoretical framework we subscribe to and on which the above mentioned annotation scenario of PDT is based is the Functional Generative Description (see [3], [4], [5], [6], [7]). This theoretical model works with an underlying syntactic level called tectogrammatics which is understood as the interface level connecting the system of language (cf. the notions of *langue*, linguistic competence, I-language) with the cognitive layer, which is not directly mirrored by natural languages. Language is understood as a system of oppositions, with the distinction between their prototypical (primary) and peripheral (secondary, marked) members. We assume that the tectogrammatical representations of sentences can be captured as dependency based structures the core of which is determined by the valency of the verb and of other parts of speech. Syntactic dependency is handled as a set of relations between head words and their modifications (arguments and adjuncts). However, there are also the relations of coordination (conjunction, disjunction and other) and of apposition, which we understand as relations of a "further dimension". Thus, the tectogrammatical representations are more complex than mere dependency trees.

The tectogrammatical representations reflect also the topic-focus articulation (information structure) of sentence, including the scale of communicative dynamism (underlying word order) and the dichotomy of contextually bound (CB) and non-bound (NB) items, which belong primarily to the topic and the focus, respectively. The scale is rendered by the left-to-right order of the nodes; in the surface structure of the sentence, the most dynamic item, i.e. focus proper, is indicated by a specific (falling) pitch and not necessarily by the word order.

The core of a tectogrammatical representation is a dependency tree the root of which is the main verb. Its direct dependents are arguments (primarily obligatory), i.e. Actor, Objective (Patient), Addressee, Origin and Effect, and adjuncts (of location and direction, time, cause, manner, and so on). Actor primarily corresponds to a cognitive (intentional) Agentive (or Experiencer, i.e. Bearer of a state or process). If the valency frame of a verb contains only a single participant, than this participant is its Actor, even though (in marked cases) it corresponds to a cognitive item that primarily is expressed by some other participant.

In a tectogrammatical representation, there are no nodes corresponding to the function words (or to grammatical morphs). Correlates of these items (especially of prepositions and function verbs) are present there only as indices of node labels: the syntactic functions of the nodes (arguments and adjuncts) are rendered as functors and subfunctors, and the values of their morphological categories (tense, number, and so on) have the forms of grammatemes.

In annotating texts from the Czech National Corpus in the frame of the project of the Prague Dependency Treebank, we work with several specific deviations

from theoretically conceived TRs described above. The most important of these deviations is that the tectogrammatical tree structures (TGTSs) we work with in PDT differ from TRs in that they have the form of trees even in cases of coordination; this is made possible by the coordinating conjunctions being handled as specific nodes (with a specific index).

In terms of the communicative function of language, an adequate explanation of information structure of the sentence may be based on the relation of *aboutness*: the speaker communicates something (the Focus of the sentence) about something (the Topic of the sentence), schematically: F(T): the Focus holds about the Topic F(T): negation: (in the prototypical case) the Focus does not hold about the Topic

A supportive argument for such a treatment is offered by the discussions on the *kinds of entailments* as opened by [8](esp. p. 173ff.), who distinguishes a formal logical relation of entailment and a formal logical relation of *presupposition*. He illustrates this distinction on the analysis of the sentence (1a): according to Strawson, (1a) as well as its negation (1b) implies (2). If John's children were not asleep, the sentence (1a) would be *false*; however, if John did not have children, then (1a) as well as its negation (1b) would not be false but *meaningless*. Strawson concludes that (2) is a presupposition of (1a) and as such it is not touched by the negation contained in (1b).

(1) a. All John's children are asleep.
b. All John's children are not asleep.

(2) John has children.

In a similar vein, [9] discusses the classical example (3a) and, most importantly, notices the difference between (4a) and (5a) by saying (p.96) "we might ... have felt a shade more squeamish if we had written (4a) instead of (5a)".

(3) a. The King of France is bald.
b. The King of France is not bald.

(4) a. The King of France visited the exhibition yesterday.
b. The King of France did not visit the exhibition yesterday.

(5) a. The exhibition was visited yesterday by the King of France
b. The exhibition was not visited yesterday by the King of France.

In his analysis of identifying reference in statements, Strawson (p. 98) suggests that a speech episode "He was saying that the King of France visited the exhibition yesterday." might be described as "he was saying what the king of France is like", in which the clause beginning with "what" specifies "the topic of the statement, what it can be said ... to be *about*"; while *what is said about its topic* is eliminated from the description in favour of the interrogative expression". He adds (imprecisely, influenced apparently by his native tongue) that "the placing of an expression at the beginning of a sentence, in the position of grammatical subject, serves, as it were, to announce the statement's topic" (p. 99).

Applying Strawson's considerations to an analysis of (3a) and its negation (3b), we may say that (3a) is *about* the King of France and therefore the King's existence (referential availability) is *presupposed* and entailed also by its negative counterpart (3b); otherwise (3a) would have no truth value, it would be meaningless. The same holds true about (4a). However, no such existential (referential) presupposition is present in (5a). The truth/falsity of (5a) and (5b) does not depend on the referential availability of the entity "King of France". These sentences are not about the King of France but about the exhibition; the existence (referential availability) of the King of France is not presupposed.

To describe the difference between the cases such as in (4a) and in (5a), we have introduced([10]; see also the commentary by [11]) a third kind of entailment in addition to *meaning proper* and presupposition, namely the so-called allegation. While (i) meaning proper can be characterized as an assertion A entailed by an assertion carried by a sentence S, the negation of A being entailed by the negation of S, and (ii) *presupposition* as an assertion A entailed by an assertion carried by a sentence S, and also by the negation of S, (iii) an *allegation* is an assertion A entailed by an assertion carried by a sentence S, with which the negative counterpart of S entails neither A nor its negation.

This distinction can be further illustrated by examples (6a) and (8a). Both (6a) and (6b) implies that we were defeated (i.e. (7) is a presupposition of both of them), they are statements about our defeat.

(6) a. Our defeat was caused by John.
b. Our defeat was not caused by John.

(7) We were defeated.

The situation is different with (8a) and (8b): it is possible to imagine that (8b) can be used in both contexts (9) and (10), which indicates that (7) is an allegation rather than a presupposition of (8a). In terms of the 'aboutness' relation, (8a) and (8b) are statements about John rather than about the defeat.

(8) a. John caused our defeat.
b. John did not cause our defeat.

(9) We were defeated because the whole team performed badly.

(10) Though it is true that John has a reputation of a rather bad player, Paul was in a very good shape and we won.

Returning to our presentation of the relation between the communicative function of language and the information structure of the sentence given at the beginning of the preceding section, we can explain the difference between (6a) and (8a) in terms of the scope of negation and the 'aboutness' relation as reflected by TFA as follows:

(i) in the prototypical case: the scope of negation constituted by the Focus: Focus (F) does not hold of Topic: F(T).
(ii) in a secondary case, the assertion holds about a negative Topic: F(T)

Compare the possible interpretations of (11) implied by the questions (12a) and (12b):

(11) Bert did not come because he was out of money.

(12) a. What about Bert? I am saying about Bert that he did not come because he was out of money
Topic: Bert
Focus: (he) did not come because he was out of money

b. Why didn't Bert come? I am saying about the fact that Bert did not come that this was caused by the fact that he was out of money:
Topic: Bert did not come
Focus: (because) he was out of money

c. Bert came, but for some other reason I am saying about the fact that Bert came (i.e.about his presence) that it was not because he was out of money but because ...
Topic: Bert came
Focus: not because he was out of money

In the interpretation indicated by (12b), the scope of negation is restricted to the Topic part of the sentence; the assertion triggered (on this reading) by the *because*-clause in Focus is not touched by negation (the reason of Bert's not-coming (absence) is ...).

However, there is another possible reading of (11), namely (12c), e.g. if the sentence is followed by: ... *but because he was on his leave of absence.*

Under this interpretation, Bert's being out of money is neither entailed nor negated. The scope of negation again concerns Focus, schematically: F(T). What is in the scope of negation is neither asserted, nor presupposed; the *because*-clause triggers as allegation.

These considerations – in addition to examples of evident semantic differences between sentences such as (15) through (21) quoted below - have led us to the conclusion that TFA undoubtedly is a *semantically relevant* aspect of the sentence as such should be represented at a level of an integrated language description capturing the meaning of the sentence. This level can be understood as the *'highest'* level of the *language description* viewed from the point of view of the hierarchy from *function to form*. The inclusion of TFA into this level can serve well as a starting point for connecting this layer with an interpretation in terms of *intensional semantics* in the one direction and with a description of the *morphemic and phonemic means* expressing TFA in the other direction (see below Sect. 5.1).

The semantico-pragmatic interpretation of sentences (for which the tectogrammatical representations represent a suitable input) may then include an application of Tripartite Structures (Operator - Restrictor - Nuclear Scope), as outlined by B. H. Partee in [7]. Let us briefly recall some of the characteristic sentences discussed there (with their relevant tectogrammatical representations, TRs) and specify (in a maximally simplified notation) which parts of their individual readings belong to the Operator (O), Restrictor (R) and Nuclear Scope (N) of the

corresponding tripartite structures. We assume that in the interpretation of a declarative sentence, O corresponds to negation or to its positive counterpart (the assertive modality)[3] or to some other operators such as focusing particles, R corresponds to Topic (T), and N to Focus (F).

(13) a. John sits by the TELEVISION.
b. O ASSERT, R John, N sits by the TELEVISION.
c. O ASSERT, R John sits, N by the TELEVISION.

Sentence (13a) may be analyzed in two ways: ether (i) it conveys an information about John (i.e. John being its Topic and the rest its Focus), or (ii) it conveys an information about John's sitting (i.e. with both John and the verb in the Topic). If the sentence includes a focusing particle such as *only, also, even* etc., the particle occupies its prototypical position in the TR, so that the focus of the particle is identical with the F of the sentence on either reading. If the focusing particle is included in T, its own focus (which differs from the sentence F in such marked cases) does not cross the boundary between the T and the F of the sentences, see (14) in the context indicated in the brackets (and discussed in more detail below as sentence (22)).

(14) (Everyone already knew that Mary only eats vegetables.) If even Paul knew that Mary only eats vegetables, then he should have suggested a different restaurant.

In linguistic literature, many examples have been adduced which indicate that the difference of the meaning between the members of the given pairs of sentences is given by their topic-focus structure, though not always the difference in this structure is being referred to (see ex. (15a, 15b) and (14)). Let us give here just a couple of examples (the original sources of the examples are given in brackets; the capitals denote the assumed position of the intonation centre, which is crucial for the interpretation of the given sentences).

(15) a. Everybody in this room knows at least two LANGUAGES.
b. At least two languages are known by everybody in this ROOM. ([12], [13])

(16) a. Many men read few BOOKS.
b. Few books are read by many MEN. ([14])

(17) a. Londoners are mostly at BRIGHTON.
b. At Brighton, there are mostly LONDONERS. ([15])

(18) a. I work on my dissertation on SUNDAYS.
b. On Sundays, I work on my DISSERTATION.

[3] In the interpretation, we use the ASSERT operator introduced by Jacobs (1984).

(19) a. English is spoken in the SHETLANDS.
b. In the Shetlands, ENGLISH is spoken. ([6])

(20) a. I only introduced BILL to Sue.
b. I only introduced Bill to SUE. ([16])

(21) a. Dogs must be CARRIED.
b. DOGS must be carried. ([17])
c. Carry DOGS. (a warning in London underground, around 2000)
d. CARRY dogs.

We have discussed these and several other sentences in our previous writings on TFA (see the References below) and therefore we present them here without a more detailed analysis, just for a support for our claim that the differences in the surface shape of these sentences have a common denominator, i.e. that they are due to the differences of the means of expression of an underlying phenomenon of TFA. These means of expression may concern (a) the surface *order of words*, (b) the sentence prosody, (c) the syntactic constructions, and (d) morphemic means. It goes without saying that this is an open list, especially if languages belonging to other than the Indoeuropean type are taken into account.

The most frequently and extensively discussed means of expression of the information structure is the order of words; as a matter of fact, in some approaches, the differences in the information structure are even identified with the differences in the order of words in the surface shape of the sentence. It has been sometimes claimed that with respect to the order of elements, the presence of a quantifying expression is crucial; as the examples quoted above demonstrate, there are no quantifiers present in (18) and (19) and yet the difference in meaning cannot be excluded. (18b) is about my work on dissertation, and may be true also in a context when I am preoccupied also by other things on Sundays, while this is not the case in (18a) which is about Sundays and indicates that my (only) preoccupation on Sundays in working on my dissertation. Such an "exhaustive listing" (for this notion, see [18], esp. p. 307) is also implied by (19a), and the sentence cannot stand alone e.g. in a textbook on geography since it would not convey a true information (it brings a false information about English), while (19b) is true about the Shetlands rather than about English.

(b)The order of words in the surface shape of the sentence might be the same and yet the sentences acquire different information structure and differ in their meanings which is reflected by *sentence prosody* including the placement of the intonation center. This holds e.g. about sentences in (20) and (21) above. Sentences (20a) and (20b) differ in their truth conditions: leaving aside the possible ambiguities of the placement of the verb within topic or focus, (20a) can be uttered in a situation when the speaker did not introduce other people to Sue except for Bill, this is not the case of (20b): the speaker may have introduced other people to Sue but the only person he introduced Bill to, was Sue.

M.A.K.Halliday quotes in his pioneering analyses of the relations between grammar and intonation ([17]) the example given above as (21a) and (21b). (21a) is a

warning at the bottom of the escalators in London underground and Halliday jokingly remarks that if pronounced as in (21bb), it would lead to a false assumption that (on the escalator) everybody has to carry a dog. His warning apparently has not reached the ears/eyes of the builders of the new underground stations around 2000, since these stations have been equipped by a shortened warning the natural pronunciation of which would be as indicated in (21c). This, however, would lead to the same funny interpretation as (21b) rather than to the intended interpretation (21a), unless the inscription is pronounced with the placement of the intonation centre (unusual for English) at the beginning (as in (21d)).

The respect to the prosodic expression is most perspicuously reflected in the above mentioned doctoral dissertation on *'association with focus'* by [16]. Rooth postulates the so-called 'association with focus" connected with E. particles (called focussing particles or focalizers) such as '*only*', '*even*', '*also*', etc and its assumed realization by a pitch accent (typically with a falling intonation contour).The question arises whether these particles always stipulate association with a focussed element in their scope. As demonstrated by [7], an association with the Focus of the sentence in not necessarily the case, see (14) above reproduced here as (22).

(22) a. Everyone already knew that Mary only eats vegetables.
b. If even PAUL knew that Mary only eats vegetables, then he should have suggested a different restaurant.

There are two 'focalizers' in B, namely only which introduces material repeated from the previous sentence (sometimes called a second occurrence focus), and even associated with *Paul*, which carries the intonation center. [19] observed that the acoustic realization of "second- occurrence focus" is different from the 'regular' focus; [20] refer to her analysis and claim that the second occurrence focus is not only marked differently from the 'regular' focus but also differs acoustically from the non-focused expressions. This confirms the suggestions given in [7]: the authors differentiate focus of the focussing particle (its scope) from the Focus of the sentence (i.e. the part of the sentence which is about its Topic) and illustrate this distinction by (23a) and its interpretation in terms of the tripartite structure in (23b), within the context indicated in the brackets.

(23) a. What did even PAUL realize? Even Paul realized that Jim only admired MARY.
b. O ASSERT, R (O even, R realized, N Paul), N (O only, R Jim admired, N Mary)

When deciding on the status of the given elements of the sentence in its TFA, not only the position of the intonation center should be taken into account but the whole intonation contour of the sentence (its F0 characteristics) should be considered. Such an evaluation of the F0 characteristics has led us to introduce the notion of "contrastive topic" (see e.g. [21], [22]).

As Firbas in [23] noticed, it is not always the case that the most dynamic element of Focus is to be prosodically marked. (24) is his example of an 'automatic

placement' of the intonation center at the end of the sentence even if the subject is (a part) of the focus .

(24) A boy came into the room.

Since the grammatically fixed word order of English does not always allow to linearly order the elements of a sentence as to reflect the information structure of the sentence (and passivization as in (15) and (16) above or some other syntactic restructuring cannot be applied), the use of *italics* in written English can be used to denote the position of IC. This has been observed by Alena Skaličková in the 1970's and her observation reoccurred in a paper by [24], analyzing the use of italics to mark focus in English translations of Spanish and Portuguese original texts.

(c) Among the specific *syntactic constructions* as the means of expression of TFA in English, the *it*-clefts (in contrast to the pseudo-clefts (as *wh*-clefts) are often referred to, which make it possible to 'prepose' the focussed element and thus to give it some kind of prominence. The rest of the sentence is then understood as being in a kind of 'shadow', backgrounded. The 'preposing' of the focused element is prototypically accompanied by the shift of the intonation center to the clefted element, see (25a).

(25) a. It was JOHN who talked to few girls about many problems.

b. With few girls talked about many problems
S málo děvčaty mluvil o mnoha problémech
John-Nominative.
HONZA.

Cleft constructions may serve also as an additional support for the view that not only the division of the sentence into its Topic and Focus, but also the degrees of communicative dynamism (underlying word order, see below in Sect. 5.2) as such play their role in the semantic interpretation of the sentence.

(26) a. It was JOHN who talked about many problems to few girls.

b. About many problems talked with few girls
O mnoha problémech mluvil s málo děvčaty
John-Nominative.
HONZA.

The (preferred) interpretation of (25a) indicates that there was a group of few girls with which John talked about many problems, not necessarily the same set of many problems; the (preferred) interpretation of (26a) suggests that there was a (single) set of many problems about which John talked with few girls (not necessarily with a single group of girls).

(d) Notorious examples of *morphemic* means expressing the TFA are the Japanese particles *ga* and *wa* discussed in linguistic literature since Kuno's ([18]; [25]) pioneering analysis of the function of these particles in the information structure of Japanese (most recently, the thematic function of 'wa' was analyzed e.g. by [26]).

There are many other examples of languages where morphemics serves as (one of the means of expression) of information structure quoted in linguistic literature up to now. Let us only mention two of them discussed by [27](p. 177) referring also to [28]. Information structure is expressed obligatorily and by using morphological means in Yukaghir, a Paleo-Asiatic language ([29]). There are three series of forms for each transitive verb there (distinguished from one another by the presence or absence of personal inflection, by morphological exponents, and by the presence or absence of certain prefixes) which are used whether the rheme-component coincides with the subject of the verb, its object, or the verb itself, respectively. In addition, a suffix is attached to the subject or object under conditions that pertain to the distribution of the rheme. In Tagalog, an Indonesian language, the theme of the sentence is distinguished by means of certain particles (articles) and word order; the syntactic roles of the given participants are indicated by an appropriate from of the verb ([30]).

5.2 From the Theory to an Annotation Scheme

For the theoretical description of TFA, the crucial issue is which basic oppositions are to be captured. In the approach of the Functional Generative Description to TFA, which we subscribe to, the basic opposition is seen in the opposition of contextual boundness. This opposition is represented in the underlying structure: for every autosemantic lexical item in the sentence (i.e. for every node of its tectogrammatical representation) it is specified whether it is (a) contextually bound (*cb*), i.e. an item presented by the speaker as referring to an entity assumed to be easily accessible by the hearer(s), more or less predictable, readily available to the hearers in their memory, or (b) contextually non-bound (*nb*), i.e. an item presented as not directly available in the given context, as cognitively 'new'. While the characteristics 'given' and 'new' refer only to the cognitive background of the distinction of contextual boundness, the distinction itself is an opposition understood as a grammatically patterned feature, rather than in the literal sense of the term. This point is illustrated by (27): both Tom and his friends are 'given' by the preceding context (indicated here by the preceding sentence in the brackets), but their linguistic counterparts are structured in the given sentence as non-bound (which is reflected in the surface shape of the sentence by the position of the intonation center).

(27) (Tom entered together with his friends.) My mother recognized only HIM, but no one from his COMPANY.

In the prototypical case, the head verb of the sentence and its immediate dependents (arguments and adjuncts) constitute the Topic of the sentence if they are contextually bound, whereas the Focus consists of the contextually non-bound items in such structural positions (and of the items syntactically subordinated to them). Also the semantically relevant scopes of focus sensitive operators such as *only, even*, etc. can be characterized in this way.

The bipartition of the sentence into the Topic and Focus (reflecting the aboutness relation as discussed above in Sect. 5.1) can then be specified by the

following set of the rules determining the appurtenance of a lexical occurrence to the Topic (T) or to the Focus (F) of the sentence (see [31]; [6], pp. 216ff)

(a) the main verb (V) and any of its direct dependents belong to F iff they carry index *nb*;
(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 *cb*, 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.

There are two reasons why to distinguish the opposition of contextual boundness as a primary (primitive) one and to derive the Topic-Focus bipartition from it. First, and most importantly, the Topic/Focus distinction exhibits – from a certain viewpoint - some recursive properties, exemplified first of all in sentences which contain embedded (dependent) clauses. The dependent clause D functions as a sentence part of the clause containing the word on which D depends, so that the whole structure has a recursive character; one of the questions discussed is whether the T-F articulation should be understood as recursive, too. Several situations arise: (i) one of the clauses may be understood as the F of the whole sentence, though each of the clauses displays a T-F articulation of its own; (ii) in a general case the boundary between T and F may lie within one of the clauses.

The second argument is related to the fact that Topic/Focus bipartition cannot be drawn on the basis of an articulation of the sentence into constituents but requires a more subtle treatment. In early discussions on the integration of the topic-focus articulation into a formal description of grammar, the proponents intended to specify this aspect of the structure of the sentence in terms of the type of formal description they subscribed to. Within the framework of generative transformational grammar, [32] (p. 205) defined focus as "a phrase containing the intonation center", i.e. in terms of constituency (phrase-structure) based description (see also Jackendoff 1972, p. 237). Such a description served as a basis also for several studies on the relationship between syntax and semantics (e.g. [33]; [34]; [35]): the boundaries between topic and focus or some more subtle divisions were always supposed to coincide with the boundaries of phrases. Sgall and his followers (see already [15]) work within a framework of dependency grammar and define the boundary between the two parts on the basis of syntactic dependency, of the opposition of contextual boundness and of the left-to-right order of nodes. The boundary between Topic and Focus can then be characterized as intersecting an edge between a governor and its dependent (the latter may be a single node or a subtree), with the provision that whatever is to the right of the given dependent in the tectogrammatical dependency tree, belongs to the Focus, the rest to the Topic (see Sgall's definition above).

However, the definition of Focus (and of presupposition, in Chomskyan terms) as a phrase is untenable since it is not always possible to assign the focus value

to a part of the sentence that constitutes a phrase. This claim is supported by examples as those adduced by [36]: in the given context, the Focus of the sentence is *for a week to Sicily*, which would hardly be specified as a constituent under the standard understanding of this notion. These examples, however, bring no difficulties for a dependency-based description.

(28) John went for a week to Sicily. (He didn't't go only for a weekend to his parents.)

It was convincingly argued by [37]; [38]; [39] that it is advisable to postulate a common structure for accounting both for the syntactic structure of the sentence as well as for its information structure. For that purpose, he proposes a modification of categorial grammar, the so-called combinatory categorial grammar. A syntactic description of a sentence ambiguous as for its information structure should be flexible enough to make it possible to draw the division line between Topic and Focus also in other places that those delimiting phrases; in [38] (p.5), the author claims that e.g. for the sentence *Chapman says he will give a policeman a flower* his "theory works by treating strings like *Chapman says he will give, give a policeman*, and *a policemen a flower* as grammatical constituents" and thus defining "a constituent" in a way that is different from the "conventional linguistic wisdom". In other words, Steedman proposes to work with non-standard constituents, as can be illustrated by (29) with the assumed intonation center at the last element of the sentence: the division of (29) into Topic and Focus is ambiguous because the verb may belong either to the topic or to the focus part of the sentence.

(29) Fred ate the BEANS.

The representation of such an ambiguity in a dependency framework like that of the Praguian Functional Generative Description causes no difficulty. In case the root of the tree (the verb) is cb, then it depends on the *cb/nb* feature of its dependents whether *Fred ate* or just *ate* are the elements of the Topic (answering the question *What did Fred eat?*, or *Who did eat what?*, respectively. If the verb is *nb*, then again two divisions are possible: either the whole sentence is the Focus (*What happened?*), or the verb and the object are the elements of the Focus (*What did Fred do?*). In the underlying tree structure, the *cb* nodes depend on the verb from the left, the *nb* nodes from the right. A division line between Topic and Focus is then drawn as characterized above.

In (29), we assumed the (normal) placement of the intonation center on the object *beans*. However, as also discussed by Steedman, the sentence may have different intonation patterns, and this may reduce its ambiguity: if the intonation center is on *Fred*, then *Fred* is the sentence Focus and the rest is the Topic (*Who ate the beans? Fred.*). If the intonation center is on the verb, then only the verb is the Focus the rest being the Topic (*What did Fred do with the beans? (He) ate (them).*) This again can be easily captured in the dependency representation of the meaning of the sentence by the assignment of the primary opposition of *cb/nb* nodes.

The above considerations of the theoretical status of TFA within a formal descriptive framework have led us to introduce, in the annotation scheme of the underlying layer of PDT, a specific TFA attribute as a part of the annotation of each node of the tectogrammatical tree structure, with the choice of one of the following three values: t for a non-contrastive contextually bound node, c for a contrastive contextually bound node, and f for a contextually non-bound node.

5.3 From the Annotation Scheme to the Theory

Any modern linguistic theory has to be formulated in a way that it can be tested by some testable means. One of the ways how to test a theory is to use it as a basis for a consistent annotation of large language resources, i.e. of text corpora. Annotation may concern not only the surface and morphemic shapes of sentences, but also (and first of all) the underlying sentence structure, which elucidates phenomena hidden on the surface although unavoidable for the representation of the meaning and functioning of the sentence, for modeling its comprehension and for studying its semantico-pragmatic interpretation. One of the aims the PDT was designed for was to use it as a testbed for the theoretical assumptions encapsulated in the Functional Generative Description as briefly sketched in Sect. 5.1 above.

As mentioned in Sect. 5.2, one of the hypotheses of the TFA account in FGD concerned the possibility of the derivation of the bipartition of a sentence into its Topic and Focus on the basis of the feature of contextual boundness of the individual lexical items contained in the sentence. To illustrate the hypothesis on a PDT example, let us take the Czech sentence (30) and its (very simplified) annotation on the tectogrammatical layer (in the preferred reading) as given in Figure 2.

(30) Nenadálou finanční krizi podnikatelka řešila
The sudden financial crisis(Acc.) the entrepreneur(Nom.) solved
jiným způsobem.
by other means.

The application of the rules quoted in Sect. 5.2 gives the following result: Topic: *Nenadálou finanční krizi podnikatelka* [the sudden financial crisis the enterpreneur] Focus: *řešila jiným způsobem* [solved by other means]

The implementation of an algorithm based on the quoted rules has led to a differentiation of five basic types of Focus and it significantly supported the hypothesis that in Czech the boundary between T and F is signalized by the position of the verb in the prototypical case (the boundary between T and F: immediately before the verb in 95% of the cases) and it has also been confirmed that the TFA annotation leads to satisfactory results even with rather complicated "real" sentences in the corpus.

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)

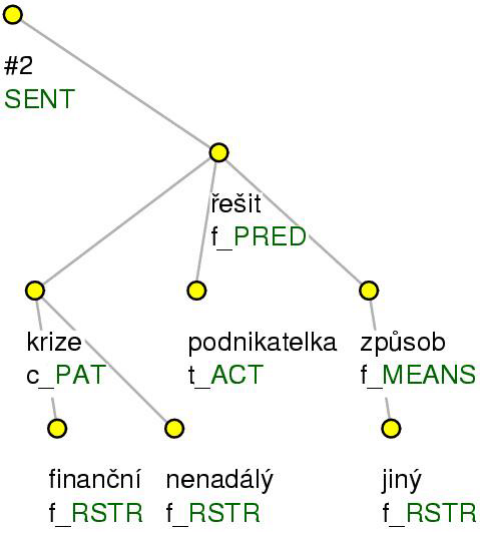

Fig. 2. The preferred TGTS of sentence (30)

follow a certain canonical order in the TRs, the so-called systemic ordering (not necessarily the same for all languages). In Czech, also the surface word order in Focus corresponds to the systemic ordering in the prototypical case.

For Czech, the following systemic ordering is postulated (see [6]): 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 ([40]). However, PDT offers a richer and more consistent material; preliminary results have already been achieved based on (a) the specification of F according to the rules mentioned above, (b) the assumed order according to the scale of systemic ordering (functors in TGTS), and (c) the surface word order ([41]). These results have led to a fruitful reconsideration and possible modification of the theoretical assumptions.

A general assumption common to any postulation of a deep (underlying) layer of syntactic description is the belief that languages are closer to each other on that level than in their surface shapes. This idea is very attractive both from the theoretical aspects as well as from the point of view of possible applications in the domain of natural language processing: for example, a level of language description considered to be "common" (in its structure, not of course in their repertoire of features) to several (even if typologically different) languages might serve as a kind of "pivot" language in which the analysis of the source and the synthesis of the target languages of an automatic translation system may meet.

With this idea in mind, it is interesting (again, both from the theoretical and the applied points of view) to design and test an annotation scheme by means of which parallel text corpora can be annotated in an identical or at least easily comparable way.

These considerations have motivated one of our current project in which the PDT scenario (described above in Sect. 3) is being applied to English texts in order to find out whether such a task is feasible and if the results may be used for a build-up of a machine translation system (or other multilingual systems).

To this end, a parallel Czech and English corpus (Prague Czech-English Dependency Treebank, see [42]) is built, with the intention to apply of the original annotation scheme designed for the annotation of Czech sentences on the tectogrammatical layer to English parallel texts.

It is well known from classical linguistic studies (let us mention here – from the context of English-Czech contrastive studies – the writings of Czech anglicists Vilém Mathesius, Josef Vachek and Libuše Dušková) that one of the main differences between English and Czech concerns the degree of condensation of the sentence structure following from the differences in the repertoire of means of expression in these languages: while in English this system is richer (including also the forms of gerund) and more developed (the English nominal forms may express not only verbal voice but temporal relations as well), in Czech, the more frequent means expressing the so called second predication (and sometimes the only possible one, see (32) below) is a dependent clause (see [43], p. 542 ff.).

It is no wonder then that in our project, secondary predication has appeared as one of the most troublesome issues. Therefore, we devote our attention to two typical nominal forms serving for the expression of secondary predication in English and look for their adequate representation on the tectogrammatical layer of PDT, namely (1a) infinitive (see (31)) and (2) gerunds (see (32)). The leading idea of our analysis is that we aim at a representation that would make it possible to capture synonymous constructions in a unified way (i.e. to assign to them the same TGTS, both in the same language and across languages) and to appropriately distinguish different meanings by the assignment of different TGTSs.

(31) Jan slyší Marii(Acc.) plakat(Inf.).
John hears Mary cry.

(32) Jan očekává, že Marie odejde.
John expects Mary to leave.
or:
John expects that Mary leaves.

(33) Viděl jsem, že jeho úspěch roste.
(I) saw that his success grows.
I saw his success growing.

This is still a work in progress ([44]) but the preliminary investigations in this direction and a consistent effort to confront the application of the PDT

annotation on both Czech and English as typologically different languages scenario have brought several interesting stimuli for the theoretical considerations.

6 Conclusion

Our experience has convinced us that a corpus annotation on an underlying level is a feasible task, not only if the predicate – argument structure is to be captured but also with respect to the information structure of the sentence reflecting the communicative function of language, which indicates **what we are talking about and what we are saying about it**. To this aim strong **interconnections** between theoretical research and corpus annotation efforts as well as a due regard to computational aspects of the enterprise are necessary and mutually enriching. Such cooperation is also fruitful for **applications** such as automatic and machine assisted translation on different layers of complexity, communication with intelligent systems, information retrieval, grammar checking and so on.

Acknowledgments

The author gratefully acknowledges the most useful comments and suggestions given by Jan Hajič, Jarmila Panevová, Petr Sgall, Barbora Vidová Hladká and Šárka Zikánová after having read the pre-final version of the manuscript. The present paper has been written under the support of the grant MSM0021620838 of the Ministry of Education of the Czech Republic, and the EU project Euromatrix (IST-FP6-034291-STP).

References

1. Uszkoreit, H.: New Chances for Deep Linguistic Processing. In: Huang, C.R. (ed.) Frontiers in Computational Linguistics, Shangwu Press Beijing (2004)
2. Hajič, J.: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Hajičová, E. (ed.) Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová, pp. 106–132. Karolinum, Charles University Press, Prague, Czech Republic (1998)
3. Sgall, P.: Zur Frage der Ebenen im Sprachsystem. In: Travaux linguistiques de Prague 1, pp. 95–106 (1964)
4. Sgall, P.: Generative bschreibung und die ebenen des sprachsystems. Zeichen und System der Sprache III, 225–239 (1966)
5. Sgall, P.: Generativní popis jazyka a česká deklinace. Academia, Prague (1967)
6. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Reidel Publishing Company, Academia and Dordrecht, Prague (1986)
7. Hajičová, E., Partee, B.H., Sgall, P.: Topic-focus articulation, tripartite structures, and semantic content. Kluwer, Amsterdam (1998)
8. Strawson, P.P.: Introduction to Logical Theory. Methuen, London (1952)

9. Strawson, P.P.: Identifying Reference and Truth Values. Theoria, 96–118 (1964)
10. Hajičová, E.: On presupposition and allegation. In: Partee, B.H., Sgall, P. (eds.) Discourse and Meaning, pp. 99–122. John Benjamins Publ. House, Amsterdam (1996)
11. Partee, B.H.: Allegation and Local Accommodation, pp. 65–86 (1996)
12. Chomsky, N.: Aspects of the Theory of Syntax. The MIT Press, Cambridge (1965)
13. Chomsky, N.: Syntactic Structures. Mouton, The Hague (1957)
14. Lakoff, G.: On Generative Semantics, pp. 232–296 (1971)
15. Sgall, P.: Functional Sentence Perspective in a Generative Description of Language. The Prague Bulletin of Mathematical Linguistics, 203–225 (1967)
16. Rooth, M.: Association with Focus. PhD thesis, Univ. of Massachusetts, Amherst (1985)
17. Halliday, M.A.K.: Intonation and Grammar in British English. Mouton, The Hague (1967)
18. Kuno, S.: Functional sentence perspective. Linguistic Inquiry, 296–320 (1972)
19. Bartels, C.: Acoustic correlates of second occurrence focus: Towards and experimental investigation. In: Kamp, H., Partee, B.H. (eds.), pp. 11–30 (1997)
20. Beaver, D., Clark, Z.B., Flemming, E.: T.Jaeger, F., Wolters, M.: When semantics meets phonetics: Acoustical studies of second-occurrence focus. Language, 245–276 (2007)
21. Hajiéová, E., Sgall, P.: Degrees of Contrast and the Topic-Focus Articulation. In: Steube, A. (ed.) Information Structure - Theoretical and Empirical Aspects, pp. 1–13. Walter de Gruyter, Berlin (2004)
22. Veselá, K., Peterek, N., Hajiéová, E.: Topic-Focus Articulation in PDT: Prosodic Characteristics of Contrastive Topic. The Prague Bulletin of Mathematical Linguistics, 5–22 (2003)
23. Firbas, J.: Functional Sentence Perspective in Written and Spoken Communication. Cambridge University Press, Cambridge (1992)
24. Gaby, S.: The use of italics as stylistic devices marking information focus in English translation. In: Proceedings of the Corpus Linguistics Conference, Birmingham, p. 55 (2007)
25. Kuno, S.: The structure of the Japanese language. Cambridge, Mass (1973)
26. Fazuo, F.: A consideration of the thematiser 'wa' (in Japanese), pp. 147–160 (2003)
27. Novák, P.: Remarks on devices of functional sentence perspective. Papers on Functional Sentence Perspective, pp. 175–178. Academia, Prague (1974)
28. Dahl, O.: Topic and comment: a study in Russian and general transformational grammar. Slavica Gothoburgensia 4, Göteborg (1969)
29. Krejnovič, E.A.: Jukagirskij jazyk. Leningrad, Moscow (1958)
30. Bowen, D.: Beginning Tagalog. Berkeley and Los Angeles (1965)
31. Sgall, P.: Towards a Definition of Focus and Topic. The Prague Studies of Mathematical Linguistics, 173–198 (1981)
32. Chomsky, N.: Deep Structure, Surface Structure and Semantic Interpretation, pp. 193–216 (1971)
33. Schmerling, S.: Aspects of English Sentence Stress. University of Texas Press, Austin, Texas (1971)
34. Selkirk, E.: Phonology and Syntax: The Relation between Sound and Structure. MIT Press, Cambridge (1984)
35. Selkirk, E.: Sentence Prosody: Intonation, Stress and Phrasing. In: Goldsmith, A. (ed.) Handbook of Phonological Theory, pp. 550–569. Blackwell, London (1995)
36. Hajièová, E., Sgall, P.: Topic and Focus in Transformational Grammar. Papers in Linguistics, 3–58 (1975)

37. Steedman, M.: Structure and Intonation. Language, 260–296 (1991)
38. Steedman, M.: Surface Structure and Interpretation. The MIT Press, Cambridge (1996)
39. Steedman, M.: Information structure and the syntax-phonology interface. Linguistic Inquiry, 649–689 (2000)
40. Sgall, P., et al.: Experimental Research on Systemic Ordering. Theoretical Linguistics, 97–239 (1995)
41. Zikánová, v.: What Do the Data in PDT Say about Systemic Ordering in Czech? The Prague Bulletin of Mathematical Linguistics, 39–46 (2006)
42. Cuřín, J., et al.: The Prague Czech-English Dependency Treebank 1.0 CD-ROM (2004) CAT: LDC2004T25, Linguistic Data Consortium (2004)
43. Dušková, L.: Mluvnice současné angličtiny na pozadí češtiny. Academia, Prague (1988)
44. Cinková, S., et al.: The tectogrammatics of English: on some problematic issues from the viewpoint of Prague Dependency Treebank (in preparation)