The Tectogrammatics of English: on Some Problematic Issues from the Viewpoint of the Prague Dependency Treebank

Silvie Cinková – Eva Hajičová – Jarmila Panevová – Petr Sgall
Institute of Formal and Applied Linguistics
Charles University in Prague
{cinkova, hajicova, panevova, sgall}@ufal.mff.cuni.cz

1 Introductory Remarks

The present paper is aimed to illustrate how the description of underlying structures carried out in annotating Czech texts may be used as a basis for comparison with a more or less parallel description of English. Specific attention is given to several points in which there are differences between the two languages that concern not only their surface or outer form, but (possibly) also their underlying structures, first of all the so-called secondary predication (section 3.2). In section 4, we discuss the representations of these constructions in the PDT of Czech as compared with the corresponding annotation in the scenario of a treebank of English (PEDT), being developed in Prague as an English counterpart of PDT (Šindlerová et al., 2007, Bojar et al., 2007).

2 Tectogrammatics

In the Functional Generative Description (see Sgall et al., 1986, Hajičová et al., 1998), tectogrammatics is 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 (TRs) 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 TRs are more complex than mere dependency trees.

The TRs also reflect the topic-focus articulation (information structure) of sentences with a 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 in the TRs by the left-to-right order of the nodes, although in the surface the most dynamic item, i.e., focus proper, is indicated by a specific (falling) pitch.

In a theoretical description of language, the TRs are seen in a direct relationship to morphemic (surface) structures. This relationship is complicated by many cases of asymmetry – ambiguity, synonymy, irregularities, including the differences between communicative dynamism and surface word order (the latter belonging to the level of morphemics).

The core of a TR is a dependency tree the root of which is the main verb. Its direct dependents are arguments, 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, in other cases to an Experiencer (Bearer) of a state or process. If the valency frame of a verb contains only a single participant, then this participant is its Actor, even though (in marked cases) it corresponds to a cognitive item that primarily is expressed by Objective (see (1)).

(1) The book (Actor) appeared.

If the the valency frame of a verb contains just two participants, these are Actor and Objective, which primarily correspond to Agentive and Objective, although the Objective may also express a cognitive item that primarily corresponds to another argument (see (2)).

(2) The chairman (Actor) addressed the audience (Objective)

If the frame contains more than two items, then it is to be distinguished wheter the "third" of them is Addressee, Origin, or Effect (cf. the difference between e.g., (3) and (4).

(3) Jim (Actor) gave Mary (Addressee) a book (Objective).
(4) Jim (Actor) changed the firm (Objective) from a small shop (Origin) into a big company (Effect).

In a TR, 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 in the TRs only as indices of node labels: the syntactic functions of the nodes (arguments and adjuncts) are rendered here as functors, and the values of their morphological categories (tense, number, and so on) have the forms of grammatemes. Functors and grammatemes can be understood as indices of lexical items.

In annotating texts from the Czech National Corpus in the frame of the project of the Prague Dependency Treebank (PDT) (Hajič et al., 2006), 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, here the subscript *coord*, distinguishing between the coordinated items and an item depending on the coordination construction as a whole). Thus, the (primary) TGTS of the sentence (5), with many simplifications, is the tree presented in figure 1:

(5) Mary and Tom, who are our neighbours, have two children.

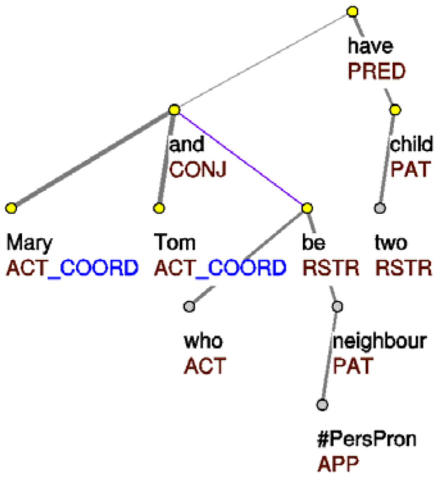

Figure 1

More details are presented in a linearized form of the corresponding TR in (5‘); note that (i) every dependent item (or a string of coordinated items) is embedded in its own pair of parentheses, and the functors are present here as subscripts of the parenthesis oriented towards the head, and (ii) the left-to-right order of the nodes, corresponding to the communicative dynamism, differs from the surface word order of the numeral *two*, which is contextually non-bound and is more dynamic than its head noun. Most of the grammatemes are left out.

(5‘) ((Mary Tom)$_{\text{Conj}}$ ($_{\text{Rstr}}$ be ($_{\text{Obj}}$ neighbour.Plur ($_{\text{App}}$ we))))$_{\text{Actor}}$ have ($_{\text{Obj}}$ child.Plur ($_{\text{Rstr}}$ two))

Rstr indicates here a restrictive adjunct, *App* one of Appurtenance (broader than possession), the other abbreviations being self-explaining.

Dependency trees are projective, 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 represent the TRs in a linearized way, as illustrated by (5‘) above. Projective trees thus come relatively close to linear strings; they belong to the most simple kinds of patterning.

3 Selected English Syntactic Constructions for Comparison

3.1 Introduction

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" (at least in some basic features) to several (even if typologically different) languages might serve as a kind of a "pivot" language in which the analysis of the source and the synthesis of the target languages of an automatic translation system may meet (see Vauquois' known "triangle" of analysis – pivot language – synthesis, Vauquois, 1975).

With this idea in mind, it is then interesting (again, both from the theoretical and the applied points of view) to design an annotation scheme by means of which parallel text corpora can be annotated in an identical or at least easily comparable way. It goes without saying, of course, that the question to which extent a certain annotation scenario designed originally for one language is transferrable to annotation of texts of another language is interesting in general, not just for parallel corpora.

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 (and sometimes the only possible) means expressing the so called second predication is a dependent clause (see Dušková et al., 1994, p. 542 ff.).

It is no wonder then that in our project, secondary predication has appeared as one of the most troublesome issues. In the present section, we devote our attention to one typical nominal form serving for the expression of secondary predication in English, namely infinitive (section 3.2), and look for its adequate representation on the tectogrammatical layer of PDT. The leading idea of our analysis is that we should 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.

The considerations included in the present section of our contribution resulted from our work on a project in which the PDT scenario (characterized above in section 2) was applied to English texts in order to find out if 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); see Šindlerová et al. (2007) and Bojar et al. (2007). This English counterpart of PDT (PEDT) comprises approx. 50,000 dependency trees, which have been obtained by an automatic conversion of the original Penn Treebank II constituency trees into the PDT-compliant a-layer trees (i.e., trees representing the surface shape of sentences). These a-layer trees have been automatically converted into t-layer trees.

3.2 Secondary Predication Expressed by Infinitive

Two classes of constructions are often distinguished: equi-NP deletion and raising. The distinction between the two classes of verbs was already mentioned by Chomsky (1965, pp. 22-23) who illustrated it on the examples (6) and (7):

(6) They expected the doctor to examine John.
(7) They persuaded the doctor to examine John.

Referring to Rosenbaum (1967), Stockwell et al. (1973), p. 521ff., discuss the distinction between *expect* and *require* (which is even clearer than Rosenbaum's distinction between *expect* and *persuade*) and point out that a test involving passivization may help to distinguish the two classes: while (8) and (9) with an equi-verb are synonymous (if their information structure is not considered), (10) and (11) with a raising verb are not:

(8) They expected the doctor to examine John.
(9) They expected John to be examined by the doctor.
(10) They required the doctor to examine John.
(11) They required John to be examined by the doctor.

The authors propose a deep structure indicated by (12) for *expect* (*hate* or *prefer*) and a deep structure that includes an animate object in addition to a sentential object for *require* and *persuade* (see (13)) while it is not important that this NP is then rewritten as S)

(12) They – AUX – VP [V(expect) NP (the doctor examine John)]
(13) They – AUX – VP [V(require) – NP (the doctor) – NP (the doctor examine John)]

Such a treatment of structures with equi verbs implies that there must be a position in the deep structure which is phonologically null (empty category PRO) and which is coreferential with one of the complementations of the equi verb; in our examples above, it is the object in (6). In theoretical linguistics, this issue is referred to as the relation of control (Chomsky, 1981; see also a detailed cross-linguistic study by Růžička, 1999; for Czech, see Panevová, 1986).

The different behaviour of verbs in the structures verb plus infinitive is discussed also in traditional grammars of English. Quirk et al. (2004) observe a certain gradience in the analysis of three superficially identical structures, namely N_1 V N_2 to-V N_3 (see their Table 16.64a, p. 1216) illustrated by sentences (14), (15) and (16) (their A, B, and C, respectively), each of which conforms to this pattern:

(14) We like all parents to visit the school.
(15) They expected James to win the race.
(16) We asked the students to attend a lecture.
(17) James was expected to win the race.

The authors claim that there is a strong reason to see a clear distinction between (14) and (16): in (14) they postulate a structure in which N_2 functions as the subject of the infinitival clause while in (16) the N_2 should be analyzed as the object of the main clause. However, according to the authors, (15) partakes in both these descriptions: from the semantic point of view, the same analysis as that of (14) would be appropriate; from the structural viewpoint, the analysis similar to that of (16) is preferable. This is supported by the fact that N_2 may become the subject of the passive sentence (17). With this analysis, N_2 behaves like an object in relation to the verb of the main clause and like a subject in relation to the infinitival clause. The authors use the term raised object to characterize this situation and they support their analysis by several criteria.

4 Solutions Proposed

4.1 Subject Raising

In the scenario of PEDT (the Prague English Dependency Treebank), the distinction between the structures with the so-called raising verbs and control verbs is preserved. The sentence (18) (see figure 2) is a typical example for the subject raising construction in English, see also a possibility of (18a) in English:

(18) John seems to understand everything.
(18a) It seems that John understands everything.

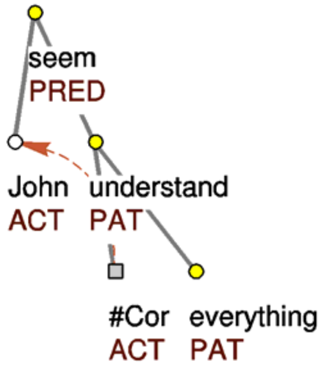

Figure 2

However, its Czech counterpart *zdát se* is connected with certain constraints: this verb must be determined by verbo-nominal (or only nominal) complement, see ex. (19). With verbo-nominal complement it has an analogical structure to the English example in figure 2, see figure 3. These constraints, however, eliminate this verb from the "pure" raising constructions; see also the unacceptability of (20) in Czech:

(19) Jan se zdá (být) smutný.
Lit. John Refl. he-seems (to-be) sad.
(20) * Jan se zdá rozumět.
Lit. John Refl. he-seems to-understand

In English, the modal and phase verbs are considered as belonging to the class of subject raising verbs. In the PDT scenario (as well as in the theoretical framework for it, FGD) most of these verbs are treated as auxiliaries, and their modal meanings are described by morphological grammatemes assigned to the autosemantic verb. As for modal verbs, this approach is adopted for PEDT as well (see Cinková et al., 2006, p. 88f.). This approach is planned for the treatment of phase verbs, too (*Jan začal pracovat* [John started to work]*, Jan začínal pracovat* [John was going to start to work]) could be described as multi verbal predicates).

The underlying structure proposed for subject raising constructions in Czech as well as in English is, however, identical to the control verb constructions, where ACT (i.e., the first argument of the control verb) controls Sb (subject) of the infinitive clause (see section 4.3).

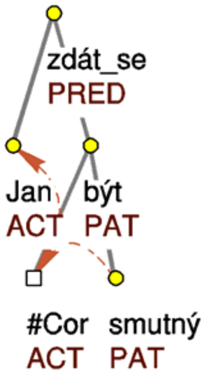

Figure 3

4.2 Object Raising

The English verbs used as clear examples of object raising verbs have no Czech counterparts with infinitive constructions; cf. (21) and figure 4 for English:

(21) John expects Mary to leave.

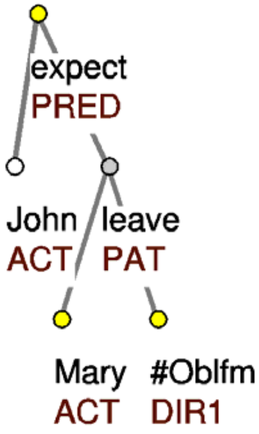

Figure 4

However, the subclass of verbs displaying this operation, called sometimes ECM (exceptional case marking), share this behavior with Czech constructions of *accusativus cum infinitivo* (AccI in sequel). It concerns the verbs of perception (see (22a) and figure 5 for English and (22b) and figure 6 for Czech):

(22a) John hears Mary cry/crying.
(22b) Jan slyší Marii plakat.

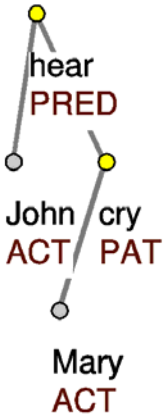

Figure 5

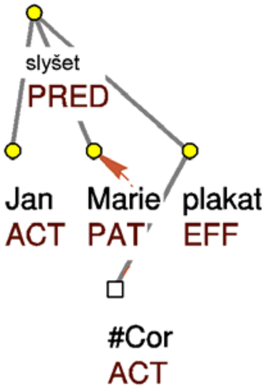

Figure 6

There are two possible ways to reflect the underlying structures of these sentences:

The approach (A) is influenced by the English tradition: The verbs of perception proper (such as *to see, to hear*) are understood in English as two-argument structures; if their second argument is expressed by secondary predication, the first argument of the secondary predication is raised up and it receives ("exceptionally") the Accusative form. The structure given in figure 5 would yield the surface structure (22a) as well as the surface structure (22c):

(22c) John hears that Mary cries.
(22d) Jan slyší, že Marie pláče.

However, the synonymy illustrated by (22a) and (22c) does not hold in all contexts, see (23a), (23b), (23c) and (23d), and also (24a) and (24b):

(23a) Jan slyšel, že Carmen zpívá Dagmar Pecková.
Lit. Jan heard that Carmen-Acc sings Dagmar Pecková
(23b) Jan slyšel, že Dagmar Pecková zpívá Carmen.
Lit. Jan heard that Dagmar Pecková sings Carmen
(23c) Jan slyšel Dagmar Peckovou zpívat Carmen.
Lit. Jan heard Dagmar Pecková to-sing Carmen
(23d) ?Jan slyšel Carmen zpívat Dagmar Peckovou.
Lit. Jan heard Carmen-Acc to-sing Dagmar Peckova-Acc
(24a) Jan slyšel tu skladbu hrát kapelu Olympic.
Lit. Jan heard the piece-Acc to-play the band Olympic-Acc
(24b) Jan slyšel, že/jak tu skladbu hraje kapela Olympic.
Lit. Jan heard that/how the piece-Acc plays the band Olympic-Nom

In the pairs (23a), (23b) vs. (23c), (23d) the difference between the meanings of the polysemic verb *slyšet* [to hear] is reflected: while in (23a) and (23b) Jan is either the direct hearer of the singing or he may be only told about the singing, in (23c) and (23d), if it is possible at all, he must be a direct listener. Moreover, the possible pre-posing of the object of the dependent clause (see (23a) and (24a) for Czech) has no counterpart in English.

In the approach (B) reflecting the situation in Czech the verbs of pereception are understood as three-argument structures with the underlying structure given in figure 6 corresponding to the sentence (22d), which differs from the underlying structure of ex. (22c) given in figure 5.

Under the approach (A), the formulation of the conditions under which the secondary predication could be nominalized by an infinitive clause seems to be very complicated while with the approach (B) the raised object is understood as a part of a cognitive operation, the result of which is manifested on the level of underlying structure.

4.3 Control (Equi) Verbs

As for the control verbs, the underlying structure proposed for Czech seems to be suitable for the PEDT scenario as well, see (25), (26) and figure 7, 8. A special node with lemma *Cor* is used for the controllee and an arrow leads from this node to its controller. The list of the verbs sharing the attribute of control will be nearly identical for both languages.

(25) John refused to cooperate.

(26) The editor recommended the author to correct the errors immediately.

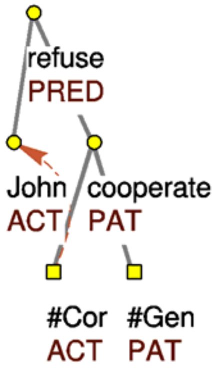

Figure 7

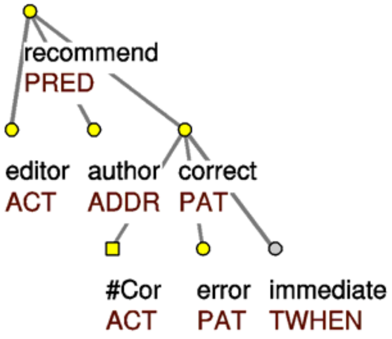

Figure 8

We have concluded that though the notions of raising and control are assumed not to be theory dependent and therefore applicable in both scenarios (for PDT as well as for PEDT), the differences between these two classes are not substantial (and they seem to be overestimated in the theoretical works).

4.4 Nominal Predicates

Analogical control constructions appear with some adjectives in the position of the nominal predicates in sentences with copula, see (27), (28) and figure 9 for English:

(27) John is eager to please.
(28) John is eager to be pleased.

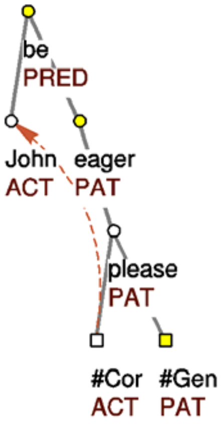

Figure 9

The corresponding underlying structures for Czech sentences (29a), (30a) are similar to those for English (29b), (30b):

(29a) Jan je schopen to udělat.
(29b) John is able to do it.
(30a) Jan je ochoten být očkován.
(30b) John is willing to be vaccinated.

However, the list of English adjectives complemented by an infinitive clause is wider than in Czech. In (31), (32) and figure 10 a control between ACT and the Sb of infinitive clause could be seen:

(31) She was quick to shut the door.
(32) Bob was reluctant to respond.

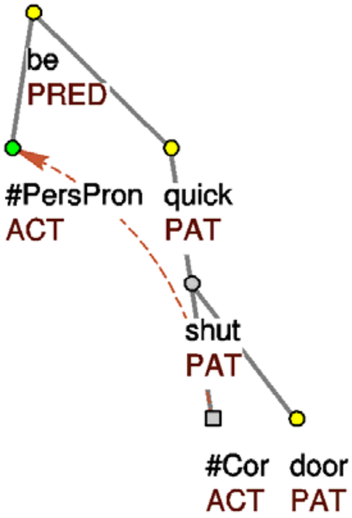

Figure 10

4.5 Tough Movement

The object-to-subject raising (sometimes called tough movement) takes place with some evaluative adjectives in complex predicates, see (33a) and its transformed version after the raising operation (33b, figure 11):

(33a) It is difficult to please John.
(33b) John is difficult to please.

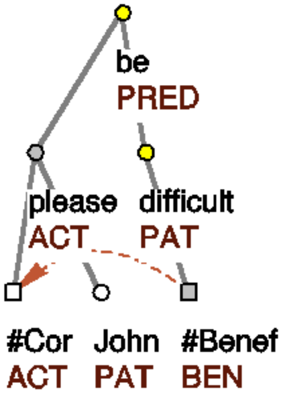

Figure 11

This type of raising has no counterpart in Czech.

4.6 Causative Constructions

Causativity of constructions such as (34) and (35) is expressed by the lexical meanings of the “semiauxiliaries” *to make, to get, to have* and by the secondary predication denoting the caused event filling the position of the PAT(ient) of the semiauxiliary causative verb.

(34) John made Mary stay.
(35) John had Mary clean the window.

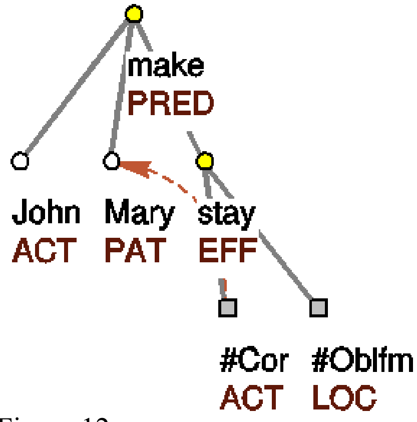

Figure 12

The constructions with the Czech verb *nechat* [to let] and the analogical underlying structure (with raised subject-to-object position) correspond to this type of causativity.

5 Conclusions

In our contribution, we have briefly discussed certain issues of secondary predication in which English differs from Czech with the result that most of them probably can be handled without differences in underlying structures of the two languages.

There are, of course, other cases in which the TRs of the two languages certainly differ. We want only to note here that not all such differences concern syntactic relations (functors). Thus in the case of such grammatical categories as definiteness or as tense and verbal aspect the differences can be captured by distinctions in the repertoires and values of grammatemes (representing morphological values).

Acknowledgements

This work was funded by GA-CR 405/06/0589, MSM 0021620838, and in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

References

O. Bojar, S. Cinková, J. Ptáček (2007). Towards English-to-Czech MT via Tectogrammatical Layer. In NEALT Proceedings Series: Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories (TLT 2007). 1. Bergen, Norway: North European Association for Language Technology, pp. 7-18.

S. Cinková, J. Hajič, M. Mikulová, L. Mladová, A. Nedolužko, P. Pajas, J. Semecký, J. Šindlerová, J. Toman, Z. Urešová, Z. Žabokrtský (2006). Annotation of English on the Tectogrammatical Level. Technical report UFAL TR 2006-35. Prague.

M. Čmejrek, J. Cuřín, J. Havelka, J. Hajič, V. Kuboň (2005) Prague Czech-English Dependency Treebank Version 1.0. *EAMT 2005 Conference Proceedings,* p. 73-78.

N. Chomsky (1981.) Lectures on Government and Binding. Dordrecht: Foris.

L. Dušková et al. (1994). Mluvnice současné angličtiny na pozadí češtiny [Grammar of Present-Day English on the Background of Czech], Academia, Prague.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková-Razímová (2006). Prague Dependency Treebank 2.0. CD-ROM. Linguistic Data Consortium, Philadelphia, PA, USA.

LDC Catalog No. LDC2006T01 URL<http://ufal.mff.cuni.cz/pdt2.0/>, quoted 2008-12-02.

E. Hajičová, B.H.Partee, P. Sgall (1998), Topic-Focus Articulation, Tripartite Structures and Semantic Content. Dordrecht: Kluwer.

M. Mikulová, 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.

J. Panevová (1986). The Czech Infinitive in the Function of Objective and the Rules of Coreference. In: J. L. Mey, ed. Language and Discourse: Test and Protest. Amsterdam: Benjamins, 123-142.

R.Quirk, S. Greenbaum, G. Leech, J. Svartvik (2004). A Comprehensive Grammar of the English Language. Longman. First published 1985.

P. S. Rosenbaum (1967). The Grammar of English Predicate Complement Constructions. The MIT Press, Cambridge, Mass.

R. Růžička, 1999, Control in Grammar and Pragmatics. Amsterdam/Philadeplhia: Benjamins.

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

R. P.Stockwell, P. Schachter, B. Hall Partee (1973). The Major Syntactic Structures of English. Holt, Winehart and Winston, New York.

J. Šindlerová, L. Mladová, J. Toman, S. Cinková (2007). An application of the PDT scheme to a parallel treebank. Proceedings of the conference Treebanks and Linguistic Theory 2007, Bergen, pp. 163-174.

B. Vauquois (1975). Some problems of optimization in multilingual automatic translation. First National Conference on the Application of Mathematical Models and Computers in Linguistics. Varna, May 1975.