Text annotations in the Prague Dependency Treebank

Šárka Zikánová (Prague)

ABSTRACT

The contribution presents the prepared complex text annotations in the Prague Dependency Treebank (topic-focus articulation, coreference and bridging anaphora, discourse relations) and proposes solutions of the following theoretical and practical questions that arise from the interplay of the syntactic and discourse structure: advantages and disadvantages of text annotation on linear text compared to tectogrammatical trees (importance of the syntactic information for the interpretation of the discourse structure), the discrepancy between syntactic and discourse structure concerning the surface position of a discourse connective, and an unexpressed thought or assumption as a discourse argument.

KEY WORDS

discourse structure, discourse annotation, coreference, topic-focus articulation, Prague Dependency Treebank, Czech

1. INTRODUCTION

The advance of linguistic electronic corpora in the last years has allowed for a description of texts and languages in detail and from many points of view. The obvious result of the existence of such corpora is an easy accessibility of large and searchable data. A less apparent, but probably even more important consequence is the necessity to solve linguistic questions systematically, in all the range of occurring phenomena.

The present contribution introduces the complex scheme of annotation of text relations in the Prague Dependency Treebank as well as solutions of some problematic points of the annotation which can have theoretical consequences (the influence of the preliminary theoretical assumptions on the interpretation of the data, the discrepancy between syntactic and discourse structure, the role of unexpressed thoughts and assumptions in the discourse structure).

2. BASIC TERMS

The Prague Dependency Treebank contains a multi-level annotation of 49,431 Czech sentences from journalistic texts. The annotations capture morphology, technical syntactic relations, the syntactico-semantic structure (the so called tectogrammatics, which is the basic and the most important layer of the sentence structure), topic-focus articulation, coreference and discourse relations. The last three types of annotations crossing sentence borders represent a complex of text annotations.

The analyses of sentence structure in the Prague Dependency Treebank are based on the *Functional Generative Description* of language (Sgall et al., 1969). The Functional Generative Description understands the sentence structure as a complex of language layers connected by relations of forms and functions (Panevová, 1980): forms of the lower layers have certain functions on the higher layers (e.g. the form of nominative in morphology functions as a subject in syntax, the form of subject in syntax functions as an actor on the tectogrammatical layer).

Tectogrammatics is the layer of basic underlying syntactico-semantic description of sentence structure. A set of synonymous surface realizations of a sentence corresponds to one semantic entry on the tectogrammatical layer. A tectogrammatical sentence structure is represented as a dependency tree consisting of nodes connected by oriented edges. The edges in the tree are assigned a value of the type of dependency relation, called functor (e.g. Actor, Predicate, Patient). Furthermore, each node in a tree is assigned an extensive set of other semantic values, such as semantic part of speech, valency frame, grammatemes — number, gender, negation etc.).

The topic-focus articulation is represented in two ways in the tectogrammatical tree (Sgall, Hajičová, and Buráňová, 1980; Hajičová et al., 1998). Each node in the tree is marked up with a value of *contextual boundness* which indicates whether the given item is retrievable from the previous context or not. There are three values of contextual boundness: non-contrastive contextually bound node (t), contrastive contextually bound node (c), contextually non-bound node (f). The basic binary division according to the contextual boundness is supplemented by more detailed characteristics of *communicative dynamism*. Communicative dynamism denotes the extent to which each item contributes to the information flow. Items that are retrievable from the context have a low degree of communicative dynamism; on the contrary communicative dynamism of nodes expressing irretrievable information in a sentence is high. The nodes are ordered from the left to the right in the tree according to their communicative dynamism: the most important expressions in a sentence (the focus proper) are located at the rightmost position.

The annotation of the topic-focus articulation has been completely finished in the whole corpus and it was published in 2006 within the edition of the Prague Dependency Treebank (Hajič et al., 2006; Mikulová et al., 2005).

The annotation of *coreference and bridging anaphora*, the second type of text annotations in the Prague Dependency Treebank (Nedoluzhko, 2011; Nedoluzhko and Mírovský, 2011), captures relations between single nodes referring to identical (coreferential) or semantically related entities (bridging anaphora, e.g. set — subset, part — whole). The annotation concentrates on nominal and pronominal expressions with specific and generic reference. The coreferential nodes can occur within a single tree as well as in different trees in a text, and as such they can constitute long chains passing through a text and providing its coherence. From the combination of the coreference and bridging anaphora data and the annotation of the topic-focus articulation a *salience* of certain entities in text can be set (Hajičová, 1993). The main part of the annotations was published in 2011 (Nedoluzhko et al., 2011). A supplement, namely the coreference of pronominal nodes of the 1st and 2nd persons, will be added to the next edition (2013). The entire present-day annotations contain 141,793 coreferential and bridging relations within 49,431 sentences.

Discourse structure, the last phenomenon of textual coherence, is analyzed in terms of discourse connectives (conjunctions, subjunctions, discourse adverbs) and their arguments (abstract objects as independent events, Asher, 1993). The annotation scenario of discourse relations in the Prague Dependency Treebank is inspired by the approach of the Penn Discourse Treebank (Prasad et al., 2008; Prasad et al., 2007). The annotators mark up the extent of both arguments, connect them with an oriented arrow signaling a discourse relation, assign a sense label to the arrow and link the relation with the appropriate discourse connective. The sense of component discourse arguments in a discourse relation (e.g. reason — result) is deducible from the orientation of the discourse arrow and the sense label. The extent of the arguments varies from a clause to a cluster of sentences.

It is typical for the complexity of a text that the structure of discourse relations is recursive and that a single text span can be an argument of different relations.

At the present stage of the annotations, all the occurrences of explicit discourse connectives have been annotated (6,571 discourse relations within 49,431 sentences). The future analysis will concentrate on further types of expressing discourse relations, especially on alternative lexicalizations of discourse connectives (e.g. *the reason is* instead of *because*) and on relations lacking an explicit discourse connective (*It rained. [therefore] They stayed at home.*)

The annotation of discourse relations has been available together with the completed annotation of coreference and bridging anaphora since 2012.

3. SOME QUESTIONS OF TEXT ANNOTATION

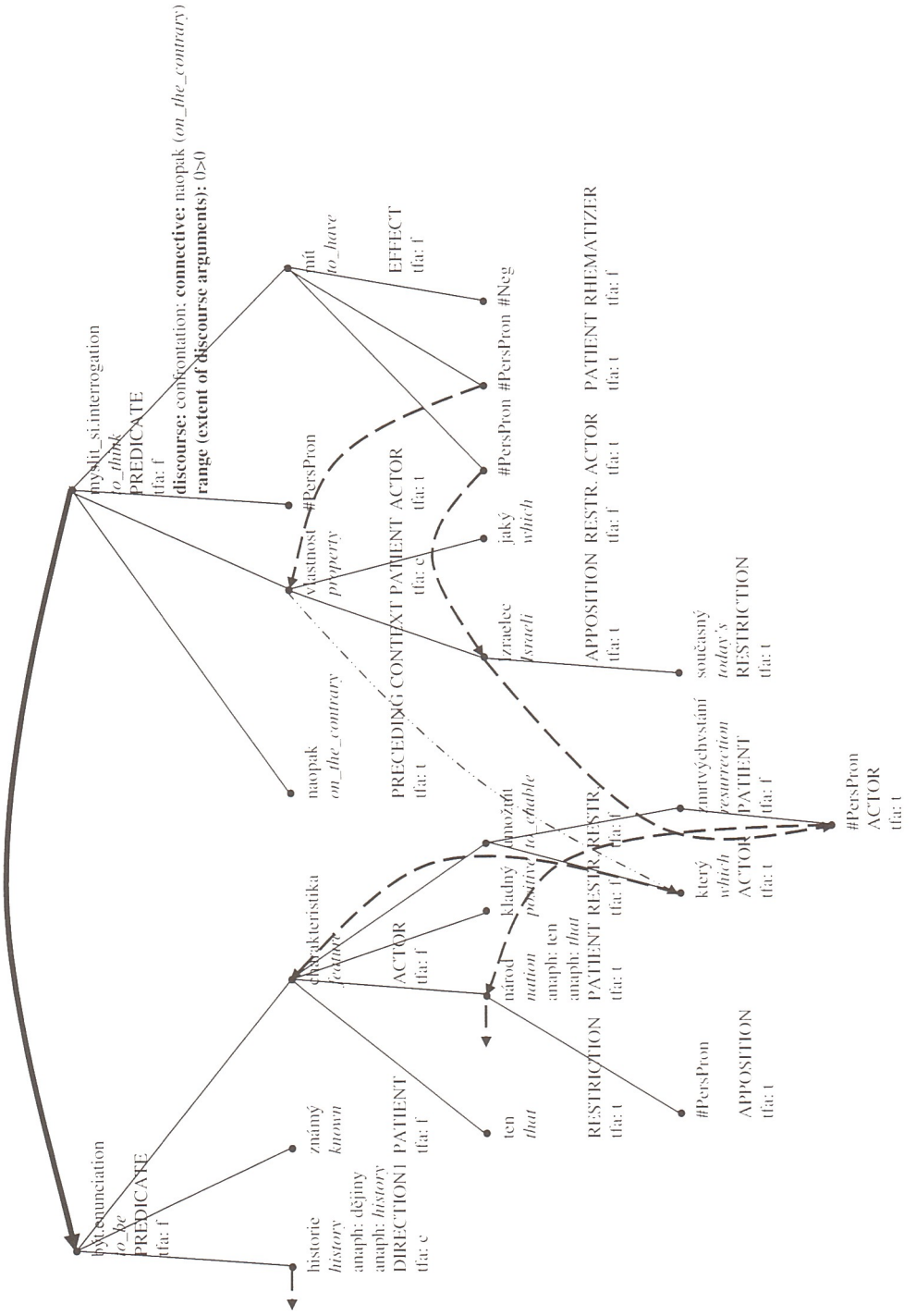
The necessity to formally treat text phenomena and to describe them fully including rare and even not known utterances raises some theoretical questions and requires finding systematic solutions of the problematic points. We want to introduce three examples of such complex questions which have arisen during the annotation of discourse relations:

(a) Some preliminary theoretical restrictions in annotation can result in the misrepresentation of data. Can we avoid this type of misrepresentation when annotating discourse relations on tectogrammatical trees rather than on a linear text?

The two following questions concern non-trivial phenomena in text:

(b) In some utterances, discourse connectives seem to occur externally from the connected discourse arguments. What are the consequences of this phenomenon for the description of a discourse structure in general?

(c) Other utterances of discourse connectives seem to relate assumptions or unexpressed thoughts of the text spans rather than certain text spans as discourse arguments. In case this really happens, how shall we treat it in the annotation?


být: enunciation
to_be
PREDICATE
tfa: f
myslit_si: interrogation
to_think
PREDICATE
tfa: f
discourse: confrontation; connective: naopak (on_the_contrary)
range (extent of discourse arguments): 0>0
historie
history
anaph: dějiny
anaph: history
DIRECTION1 PATIENT
tfa: c
tfa: f
známý
known
ten
that
RESTRICTION
tfa: t
charakteristika
feature
ACTOR
tfa: f
národ
nation
anaph: ten
anaph: that
PATIENT RESTR.
tfa: t
tfa: f
kladný
positive
#PersPron
APPOSITION
tfa: t
naopak
on_the_contrary
#PersPron
vlastnost
property
PRECEDING CONTEXT PATIENT ACTOR
tfa: t
tfa: c
tfa: t
mít
to_have
EFFECT
tfa: f
#PersPron #PersPron #Neg
jaký
which
zrušit
to_cancel
APPOSITION RESTR. ACTOR PATIENT RHEMATIZER
tfa: t
tfa: f
tfa: t
tfa: t
současný
today's
RESTRICTION
tfa: t
znovuvzkříšení
resurrection
PATIENT
tfa: f
který
which
ACTOR
tfa: t
#PersPron
ACTOR
tfa: t

3. 1 INFLUENCE OF PRELIMINARY THEORETICAL ASSUMPTIONS: ANNOTATION ON TREES AND ON A LINEAR TEXT

The question of preliminary theoretical assumptions is quite general for any scientific research: the methods and tools of the data analysis can undoubtedly influence the interpretation of the data. Therefore, it is crucial to be aware of the limits of the methods and to choose them appropriately so that they do not contradict the aims of an analysis and do not limit its results.

When planning the discourse annotation in the Prague Dependency Treebank, we faced the question whether we should annotate discourse on plain (linear) texts, as in Penn Discourse Treebank, or whether we will make use of the tectogrammatic trees which were available in the Prague Dependency Treebank (Hajič et al., 2006).

The annotation on linear texts lets the annotators mark the extent of arguments independently, relying on their own interpretation of the sentence structure. Furthermore, annotators are not limited by clause or sentence borders. Thus, the internal structure of the arguments is not predefined: it is not related to the syntactic structure in any way — it can be disrupt, it can contain e.g. incomplete clauses or a combination of a clause and single words from another independent clause. This can be a great advantage of this approach which is theoretically independent.

On the other hand, the annotators of a linear text have to solve annotation of phenomena which are not in the spotlight of the analysis. These points should be marked homogeneously in order to provide inter-annotators' agreement, though. It is necessary then to set up rules for annotating these peripheral effects and to monitor the consistency of their annotation. This applies e.g. to punctuation, brackets, but also

Figure 1: Text annotations on tectogrammatic trees in the Prague Dependency Treebank

◀——	discourse relation	
◀— —	coreferential relation	
◀··– –	bridging anaphora	
tfa: t, c, f	topic-focus articulation:	non-contrastive contextually bound node (t), contrastive contextually bound node (c), contextually non-bound node (f)

Z historie jsou známy ty kladné charakteristiky vašeho národa, které umožnily jeho zmrtvýchvstání. O jakých vlastnostech současných Izraelců si naopak myslíte, že by je nemuseli mít?

Lit.:
From history are known those positive features of_your nation which enabled its resurrection.
About which properties of_today's Israelis REFL. on the contrary you_think, that they_would them not_need to_have?

Positive features of your nation are known from the history which enabled its resurrection.
On the contrary, what are the properties of today's Israeli people that you think they don't need to have?

to larger parts of a text which cannot be put aside in the linear text organization (author's speech interposed into direct speech, parentheses, digressions).

The inorganic parts of the text can be easily eliminated in the tectogrammatical structure which allows the annotators better to concentrate on the semantic content and intention of the text. A great advantage of tectogrammatical display is the rich labeling of the data. It enables to hide or show the phenomena that are (ir)relevant for the annotation; furthermore, searching for similar structures is quick and easy. This possibility of searching decreases the disagreement among annotators.

It is true that the annotators can be influenced by the tree structure when marking an extent of a discourse argument. They are more likely to respect clause and sentence boundaries than in a linear text. It is more apparent (and could be discouraging for them) when the structure of a discourse argument is different from the syntactic structure.

We decided to test both ways of annotation before starting complex annotations in the Prague Dependency Treebank. The result generally corresponded to our assumptions. Furthermore, it became apparent that discourse arguments follow the syntactic structure; a discourse argument containing a random text span from different clauses did not appear in the test annotation. However, a new and unexpected issue arose which turned out to be frequent: annotation and interpretation of ellipses. The discourse annotation marks up relations between discourse arguments (independent abstract objects, cf. Asher, 1993), in our case abstract objects containing a finite verb (clauses). If a discourse connective relates a clause and a construction with an elided finite verb, is the latter an adequate discourse argument? Should the relation be annotated at all (cf. 1)?

(1) Context: *Where is he?*) — *?[In hospital],* ***because*** *[he has had an accident].*

There are many types of ellipses and the treatment of the occurrences of one subtype should be identical. In this case, the tectogrammatical structure was very useful, as it contains an elaborated system of ellipses reconstruction (Mikulová et al., 2005). It would be a too hard task for annotators to do the reconstructions consistently on their own.

It was especially the frequency of ellipses and digressions and parentheses in our corpus which made us choose the annotation on the tectogrammatical trees as a basic type of annotation. In order to avoid the disadvantages of the annotation on tree structures, a special annotation tool was developed which enables the annotators to see the text in a linear display, which is directly connected with the tectogrammatical annotation, and to switch to a tree structure whenever it is needed.

3. 2 DISCREPANCY BETWEEN SYNTACTIC STRUCTURE AND DISCOURSE STRUCTURE

Although the structure of single discourse arguments is not random and generally corresponds with the structure of syntactic units, the way of connecting discourse

arguments seems to be less restricted than connections of syntactic units (cf. Lee et al., 2006). The core of discourse connectives — conjunctions and subjunctions — functions on two levels: they connect syntactic units as well as discourse arguments. There are text spans where the syntactic units and discourse arguments connected by one discourse connective do not correspond to each other (cf. 2).

(2) *He said he would come* ***but*** *he added he would have to leave early.*

From the syntactic point of view, *but* connects main clauses *he said...* ***but*** *he added...* Semantically, the relation of opposition appears rather between the dependent clauses: *he would come* ***but*** *he would have to leave early*. This can be simply proved by a substitution synonymous with (2): *He said he would come* ***but*** *he would have to leave early.*

This observation has several consequences for linguistic description of discourse structure as well as for computational information retrieval. A discourse connective (conjunction, subjunction) typically occurs as a part of one of the related discourse arguments. However, it can be placed externally, too. We assume that some restrictions on the extra-placement of discourse connectives can be still set. The discourse connective can be expressed higher in the syntactic tree than it would correspond to the semantic content (i.e. in main clauses containing the discourse argument rather than within the discourse argument itself). This is typical of main clauses with a vague or almost empty semantic content: in the case of (2) a repeated act of speaking, in other cases simple verbs of existence or assertion (*It is fact that... Sometimes it happens that...*).

On the other hand, the non-symmetric lower position of a discourse connective is not excluded in Czech, either (cf. 3).

(3) *Byl nemocný. Protože* ***ale*** *nechtěl hledat lékaře, neléčil se.*
Lit.: *He_was ill. Since* ***but*** *he_didn't_want to_look_for a_doctor, he_wasn't_treated REFL.*
He was ill. ***But*** *since he didn't want to look for a medical doctor, he didn't undergo any treatment.*

The discourse connective *ale (but)* semantically relates the arguments *he was ill* ***but*** *he wasn't treated*, syntactically expressed as main clauses. The discourse connective itself is placed in the embedded clause rather than in the matrix clause. This case is slightly different from the previous one because the connective is not placed externally from the discourse argument; but there is still a discrepancy between the semantic structure and syntactic position of the discourse connective here, since the connective effects higher in the syntactic structure than it is placed.

Generally, we assume that syntactic structure limits the position of discourse connectives so that a discourse connective is placed within a syntactic complex containing the relevant discourse argument. It is a question of further research whether and under what conditions a discourse connective can be placed absolutely externally from such a syntactic complex.

3. 3 UNEXPRESSED THOUGHT AS A DISCOURSE ARGUMENT

In examples (2) and (3), the discourse connectives relate discourse arguments whose positions are not trivial to find in the text. The case may be even more complicated: there are discourse connectives whose discourse argument(s) or their parts are not present in the text at all, cf. (4):

(4) *She is at home* ***because*** *the light is on.*

The discourse connective *because* prototypically connects arguments expressing a reason and its result. In this case, the fact that *the light is on* is not a reason for the fact that *somebody is at home*. There is the causal relation rather between unexpressed thoughts *I am saying/ I can say (that she is at home)* ***because*** *I know / I can see (that the light is on).*

This analysis results in a question what the existence of this phenomenon means for an automatic information retrieval as well as how it should be treated in the annotation. In the automatic information retrieval, this could lead to a problem: whenever a program finds a discourse connective in a text, it will try to find appropriate discourse arguments and interpret them semantically (e.g. *because* needs two arguments, the one in which *because* is present expresses reason, the other one expresses result). This is not correct in our example, the core of the arguments does not occur in the text. There are two criteria which can help to solve this task in an automatic procedure. The first of them is frequency: if the absence of discourse arguments needed is generally rare, it can be omitted in automatic processes; the omission will not increase the frequency of errors significantly. The second criterion is structural typology: it is possible that the surface absence of a discourse argument is typical of certain discourse connectives or certain structures only. They can be treated separately, then.

If we want to follow the criteria of frequency and structural typology, we need to have this phenomenon annotated. It would be a too complex and unnecessary task to reconstruct the unexpressed thoughts in trees (in addition to the existing reconstructions of clear ellipses), the annotators' agreement would be certainly questionable. Therefore, we decided to annotate these occurrences just by adding a remark to the discourse connective ("Argument(s) of this discourse connective is/are not expressed", followed by a short explanation); the remark enables us to find all these instances again and to work with them in the next phase of annotation.

4. CONCLUSION

The prepared complex text annotations in the Prague Dependency Treebank will serve as a rich source of data for the linguistic research of a text structure as well as for automatic information retrieval. The preparation of the data requires defining the analyzed phenomena in the most precise way and to evaluate the methods of the annotation carefully, with regard to the aim of the annotation. Thus, we decided to

annotate the text relations not on linear texts, but on tectogrammatical trees which makes it possible for the annotators to use the information from the tectogrammatical layer. First non-trivial assumptions about the function of discourse connectives in Czech texts have been set: the surface position of a discourse connective is less restricted than a position of a syntactic conjunction/subjunction; furthermore, a discourse argument need not be expressed in a text. These hypotheses will be elaborated further and tested on a larger extent of data.

ACKNOWLEDGEMENTS

The research reported in this contribution is supported by the grant project P406/12/0658 ("Coreference, discourse relations and information structure in a contrastive perspective") of the Grant Agency of the Czech Republic.

REFERENCES

Asher, N. (1993) *Reference to Abstract Objects in Discourse*, Dordrecht: Kluwer Academic Publishers.

Hajič, J. et al. (2006) *Prague Dependency Treebank 2.0*, Philadelphia: Linguistic Data Consortium.

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

Hajičová, E. (1993) *Issues of Sentence Structure and Discourse Patterns*, Prague: Charles University.

Hoffmannová, J. (1984) Typen der Konnektoren und deren Anteil an der Organisierung des Textes, in Kořenský, J. and J. Hoffmannová (eds), *Text and the Pragmatic Aspects of Language*, Linguistica X, Prague: ČSAV, 23–39.

Hošnová, E. (2005) *Studie z novočeské syntaxe (konjunkce, pronominalizace)* (Studies from the syntax of Modern Czech [conjunctions, pronominalization]), Prague: Karolinum.

Hrbáček, J. (1994) *Nárys textové syntaxe* (A sketch of the syntax of text), Prague: Trizonia.

Lee, A. et al. (2006) Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax? in Hajič, J. and J. Nivre (eds), *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories (TLT 2006)*, Prague: ÚFAL MFF UK, 79–90.

Mann, W. C. and S. A. Thompson (1987) Rhetorical structure theory: Description and construction of text structures, in Kempen, G. (ed.), *Natural Language Generation: New Results in Artificial Inteligence, Psychology, and Linguistics*, Nijhoff, 279–300.

Mikulová, M. et al. (2005) *Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank: Annotation Manual*, Prague: Universitas Carolina Pragensis.

Nedoluzhko, A. and J. Mírovský (2011) *The Extended Textual Coreference and Bridging Relations*, Technical report, Prague: ÚFAL MFF UK.

Nedoluzhko, A. (2011) *Rozšířená textová koreference a asociační anafora. Koncepce anotace českých dat v Pražském závislostním korpusu* (Extended textual coreference and bridging relations in the Prague Dependency Treebank), Prague: ÚFAL MFF UK.

Nedoluzhko, A. et al. (2011) *Extended Textual Coreference and Bridging Relations in PDT 2.0*. Data/software, Prague: ÚFAL MFF UK. Avalaible at: https://ufal-point.mff.cuni.cz/xmlui/handle/11858/00-097C-0000-0005-BCCF-3 [2012-04-22].

Panevová, J. (1980) *Formy a funkce ve stavbě české věty* (Forms and functions in the structure of the Czech sentence), Prague: Academia.

Pasch, R. et al. (2003) *Handbuch der deutschen Konnektoren*, Berlin: de Gruyter.

Prasad, R. et al. (2007) *The Penn Discourse Treebank 2.0 Annotation Manual*. Avalaible at: http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf [2012-04-22].

Prasad, R. et al. (2008) *Penn Discourse Treebank Version 2.0*. Philadelphia: Linguistic Data Consortium.

Schiffrin, D. (1994) *Approaches to Discourse*. Blackwell Publishing.

Sgall, P., E. Hajičová and E. Buráňová (1980) *Aktuální členění věty v češtině* (Topic-focus articulation in Czech), Prague: Academia.

Sgall, P. et al. (1969) *a Functional Approach to Syntax in Generative Description of Language*, New York: American Elsevier.