Does Tectogrammatics Help the Annotation of Discourse?

Jiří Mírovský, Pavlína Jínová and Lucie Poláková
Charles University in Prague
Institute of Formal and Applied Linguistics

{mirovsky|jinova|polakova}@ufal.mff.cuni.cz

Abstract

In the following paper, we discuss and evaluate the benefits that deep syntactic trees (tectogrammatics) and all the rich annotation of the Prague Dependency Treebank bring to the process of annotating the discourse structure, i.e. discourse relations, connectives and their arguments. The decision to annotate discourse structure directly on the trees contrasts with the majority of similarly aimed projects, usually based on the annotation of linear texts. Our basic assumption is that some syntactic features of a sentence analysis correspond to certain discourse-level features. Hence, we use some properties of the dependency-based large-scale treebank of Czech to help establish an independent annotation layer of discourse. The question that we answer in the paper is how much did we gain by employing this approach.

Title and abstract in Czech

Pomáhá tektogramatika při anotaci diskurzních vztahů?

Abstrakt

V tomto příspěvku hodnotíme přínos, který představují syntacticko-sémantické stromy (tektogramatická rovina anotace) a celá bohatá anotace Pražského závislostního korpusu pro anotaci diskurzní struktury textu, tedy pro anotaci diskurzních vztahů, jejich konektorů a argumentů. Rozhodnutím anotovat diskurzní strukturu přímo na stromech se náš přístup liší od většiny podobně zaměřených projektů, které jsou obvykle založeny na anotaci lineárního textu. Naším základním předpokladem je, že některé syntaktické rysy větné analýzy odpovídají jistým rysům z roviny diskurzní struktury. Proto využíváme některé vlastnosti rozsáhlého závislostního korpusu češtiny k ustanovení nezávislé diskurzní anotační vrstvy. V tomto příspěvku odpovídáme na otázku, jaké výhody tento přístup přináší.

Keywords : Tectogrammatics, PDT, Discourse annotation

Keywords in Czech: Tektogramatika, PDT, Diskurzní anotace

Proceedings of COLING 2012: Posters, pages 853–862,
COLING 2012, Mumbai, December 2012.

1 Introduction

In recent years, there has been an increasing interest in studying linguistic phenomena going beyond the sentence boundary. Corpora of different languages conveying discourse-relevant annotation start to appear, e.g. RST Discourse Treebank (Carlson, Marcu and Okurowski, 2002), Penn Discourse Treebank (Prasad et al., 2008) – both for English, Hindi Discourse Relation Bank (Oza et al., 2009), Potsdam Commentary Corpus for German (Stede, 2004) etc. They usually have raw written documents as the annotation basis and the authors use and adjust for their purposes some of the well known discourse methodologies. In the discourse project for Czech, on the contrary to the majority, syntactic (tectogrammatical) trees have been used as the basis for the discourse annotation. Thus, the project makes use of the theoretical framework of the functional generative description (Sgall, Panevová and Hajičová, 1986), which gave rise to the dependency treebanking in Prague. The main goal of this paper is to report in detail on exploitations we were able to make of the syntactic annotation to establish an independent level of the discourse annotation. Annotation of the discourse structure here is understood as analyzing semantic relations between discourse units, in this phase of the project exclusively relations signalled by a specific discourse connective (henceforth DC). Some of the (not only) syntactic features were very helpful and enabled us to perform automatic extractions and conversions. The tectogrammatical layer of the Prague Dependency Treebank 2.0 (henceforth PDT, Hajič et al., 2006) provided most of the information we used, in less extent we used some features from the analytical layer and also the annotation of coreference.

1.1 Layers of Annotation in PDT

The data in our project come from the Prague Dependency Treebank 2.0, which is a manually annotated treebank of Czech journalistic texts, consisting of almost 50 thousand sentences. It is already provided with several layers of manual annotation: the morphological layer (where each token from the sentence gets a lemma and a morphological tag), the analytical layer (surface syntax in the form of a dependency tree, where each node corresponds to a token in the sentence), and the tectogrammatical layer (henceforth TR; underlying syntax and semantics, also in the form of a dependency tree). There is also a separate layer of manually annotated coreference and bridging anaphora (Nedoluzhko et al., 2011b), published as an extension to PDT.

1.2 Discourse Annotation in Two Steps

In our project so far, we have focused on discourse relations anchored by an explicit (surface-present) discourse connective. These relations and their connectives have been annotated throughout the whole treebank. However, all numbers reported in the paper refer to the training and development test parts of the whole data, i.e. 43,955 sentences (approx. 9/10 of the treebank).[1]

The annotation of discourse relations proceeded in two major steps. The first phase of the annotation was a thorough manual processing of the treebank focused on the inter-sentential relations (relations between sentences) signalled by explicit discourse connectives. Intra-sentential relations were only marked manually in cases where the TR did not provide enough or correct information for the subsequent automatic extraction of discourse relations. Other cases of intra-sentential relations, where the tectogrammatical annotation was adequate for the discourse interpretation, were left to the second phase.

The second phase of the annotation consisted predominantly of an automatic procedure that extracted mostly tectogrammatical features and used them directly for the annotation of the intra-sentential discourse relations. A detailed description of the second phase can be found in Jínová, Mírovský and Poláková (2012b).

The main theoretical principle of the annotation was naturally the same for both the phases. It has been inspired partially by the lexical approach of the Penn Discourse Treebank project (Prasad et al., 2008), and partially by the already mentioned tectogrammatical approach and the functional generative description (Sgall, Panevová and Hajičová, 1986, Mikulová et al., 2005). A discourse connective in this view takes two discourse arguments (verbal clauses) as its arguments. The semantic relation between the arguments is represented by a discourse arrow (link), the direction of which also uniformly defines the nature of the argument (e.g. *reason* - *result*).[2] However, the annotation itself proceeded in each of the phases differently. During the manual annotation (phase 1), the annotators first searched for possible discourse connectives in the texts and then assigned relations, arguments, connectives and discourse types to the tree structures. In the automated annotation (phase 2), the relations and their discourse types were identified and annotated first (mostly automatically), then we searched for their connectives (also mostly automatically).

[1] Thus the last tenth of the treebank, evaluation test data, remains (as far as possible) unobserved.
[2] For further information on the annotation guidelines, see the annotation manual (Poláková et al., 2012) or http://ufal.mff.cuni.cz/discourse/

Type of the relation	number
Intra-sentential relations	**12,673**
- automatic vertical	3,090 (2,599+491)
- automatic horizontal	7,392
- manual vertical	510
- manual horizontal	1,681
Inter-sentential (all manual)	**5,514**
Total	**18,187**

TABLE 1 – Overview of discourse relations annotated in PDT

Table 1 shows the summary of all relations annotated during both phases. The intra-sentential relations are divided into two categories – vertical and horizontal. Vertical relations correspond to dependency relations, horizontal relations correspond to coordinations. Also the number of inter-sentential relations (relations between sentences) and the total number of all relations are presented.[3]

2 Intra-sentential Relations

In this Section, we focus on the annotation of the intra-sentential discourse relations (mostly phase 2) and discuss and evaluate features that helped automatize the annotation. All topics are discussed only briefly here, a detailed analyses is given in Jínová, Mírovský and Poláková (2012b).

Concerning the intra-sentential relations, i.e. the syntax-based ones, we were able to automatically convert 10,482 (3,090 vertical and 7,392 horizontal) tectogrammatical relations to discourse relations. However, for 491 of them, the discourse type had to be set manually, as explained below (second number in the parenthesis in the second row of Table 1). Mostly during the first phase of the annotation, 2,191 (510 vertical and 1,681 horizontal) intra-sentential discourse relations were annotated completely manually.

2.1 Discourse Types

An ideal case for the automatic treatment was a tectogrammatical relation with an exact semantic counterpart on the level of discourse analysis, e.g. *reason-result* (signaled by functors REAS, CSQ, CAUS), *concession* (CNCS),

[3] Let us emphasize again: all numbers refer to the training and development test parts of the data (9/10 of the treebank, 43,955 sentences).

conjunction (CONJ) (all automatic horizontal relations and 2,599 completely automatic vertical relations). Because of rich variety of connectives, some manual work preceded in case of temporal relations (491 relations).

2.2 Detection of Discourse Connectives

In most cases, the discourse connectives of intra-sentential discourse relations could be automatically detected on the basis of the information on the tectogrammatical and analytical layers. With the exception of 31 atypical cases (which were fixed manually), discourse connectives could be detected automatically for all 10,482 intra-sentential discourse relations.

2.2.1 Grammatical Coreference and Expression *což*

Pronoun-like expression *což*[4] (roughly *which* in English) represents an intra-sentential connective with the conjunction meaning and is, at the same time, inflected and plays a role of a participant of the clause structure. To make it possible to associate this connective with the discourse relation automatically, the grammatical coreference[5] had to be used. The deictic part of the expression *což* can refer both to a verbal phrase (*the war unites us* in Example 1), and to a nominal phrase (*a love to war* in Example 2). However, it functions as a DC only when it refers to a verbal phrase (Example 1).

(1) *Válka nás sjednocuje, což pro nás není přirozené.*
The war unites us, which is not natural for us.

(2) *Cítil jsem z nich lásku k válce, což je něco proti přírodě.*
I felt from them a love to war, which is something against nature.

There are a total of 355 occurrences of the expression *což* in our data, 220 occurrences have a grammatical coreference link to a finite-verb node, 11 occurrences have this link to a coordination of finite-verb nodes. Therefore, thanks to the grammatical coreference, it was possible to automatically distinguish these 231 (220+11) occurrences from the rest and identify the expression *což* as a discourse connective in these contexts.

2.3 Scope of Arguments

In all intra-sentential relations, the scope of arguments is defined as the effective

[4] It has arisen from relative pronoun *co* (*what*) and particle *-ž* which is no longer used as a separate word in Czech.
[5] Grammatical coreference has been annotated in the PDT for expressions for which it is possible to identify the coreferred part of the text on the basis of grammatical rules (this applies e.g. for relative pronouns, reflexive pronouns or for participants of control verbs (see Mikulová et. al, 2005)).

subtree[6] of the root node of the argument (the root node of the argument can either be a finite verb or a node coordinating finite verbs[7]), excluding all nodes of the other argument of the relation. In all 10,482 automatically annotated intra-sentential relations, the tectogrammatical tree structure correctly defined the scope of the arguments, independently of the fact whether the argument was formed on the surface by a continuous sequence of words or not.

For the 2,191 manually annotated relations, in all but 146 cases the scope of arguments was also equal to the effective subtree of the root node, in the 146 cases the annotator had to define a different scope of the argument.

3 Inter-sentential Relations

In this section, we focus on the annotation of the inter-sentential discourse relations (phase 1). Unlike for the intra-sentential relations, the inter-sentential discourse relations (relations between sentences) had to be annotated completely manually.[8] However, in the following subsections, we discuss and evaluate features of the tectogrammatical layer that contributed notably to the annotation.

3.1 Expressions with the PREC Label

Although the annotation on the tectogrammatical layer does not in principle surpass sentence boundaries (i.e. each sentence is represented by an individual tree), one special mark has been adopted for expressions that signal (mostly) an inter-sentential relation (it is often the case with connectives such as *proto* (*therefore*), *ovšem* (*however*), *tedy* (*hence*)), see Mladová (2008). An expression marked with the functor PREC (a reference to PREceding Context) on the tectogrammatical layer thus indicates a possible presence of a discourse relation, but, at the same time, it does not interpret the semantic type of the relation, neither says anything about the scope and the position of the other discourse argument (see Example 3).

(3) *Rádi bychom ale začali u středních odborných učilišť.*
V jejich případě ***ovšem*** *záleží také na domluvě s ministerstvem hospodářství.*

But we would like to start with the vocational schools.
In their case, ***however****, also the arrangement with the Ministry of Economy matters.*

Expressions with label PREC proved to be a very important clue during the annotation process – they served as a clear signal of a possible discourse relation

[6] Effective subtree of a node is a set of nodes that linguistically depend (transitively) on the given node, taking all effects of coordinations etc. into account.
[7] possibly transitively, i.e. through other coordinating nodes
[8] See Jínová, Mírovský and Poláková (2012a) for the evaluation and analysis of the inter-annotator agreement.

in the context and were used after each part of the annotation to check the completeness of the annotation. The total number of occurrences of expressions with label PREC in our data is 5,441. The vast majority of them – 4,313 – were added as a connective to a discourse arrow (3,910 to inter-sentential relations, 403 to intra-sentential relations). The remaining occurrences of these expressions were marked by an annotator´s comment in the data and will be analyzed according to their function in some next phase of the work.

3.2 Role of Textual Coreference

In the PDT, textual coreference has been annotated for all syntactic nouns (substantives and pronouns behaving as nouns) and some adjectives throughout the whole corpus. Coreferred expressions are not necessarily only other nouns, they can also be verbs or other parts of text, if it is an appropriate interpretation of the context (for details see Nedoluzhko, 2011a). From the theoretical point of view, textual coreference is not a part of the tectogrammatical layer of the PDT but it contributes largely to the representation of meaning.

3.2.1 Connectives with a Deictic Part

One aspect of textual coreference proved to be partly helpful in determining discourse connectives. Many connectives in Czech (and also in other languages) have arisen from a connection of a preposition and a deictic element.[9] The deictic part of these prepositional phrases refers to some previous context and the coreference link helps decide if the phrase in a given context functions as a DC or not. For the DC function of such prepositional phrases, verbal antecedent of the deictic part is characteristic (for a detailed analysis, see Poláková, Jínová, Mírovský, 2012). In Example 4, the deictic element *tomu* (dative form of *that*) of the phrase *naproti tomu* (*in contrast with that*, lit. *opposite that*) has in the PDT annotation a referential link to the verb *dosáhnout* (*to achieve*) in the sentence 1.

(4) 1. *Velmi dobrých výsledků **dosáhly** divize Montáže, Klimatizace a Dodavatelská divize.*
2. ***Naproti tomu** divize Odlučování měla za první tři měsíce ztrátu 1,8 milionu korun a divize Ventilátory tři miliony korun.*

1. *Very good results **were achieved** by the divisions of Assembly, Air Conditioning and Delivery.*
2. *In contrast with that [lit. **opposite that**], the division of Separation lost 1.8 million in the first three months and the division of Fans three million.*

We encountered 103 occurrences of a preposition plus a deictic element during

[9] These connectives were called alternative lexicalizations in the PDTB approach to the annotation of discourse (see Prasad et al., 2010).

the discourse annotation that can function as a DC in Czech. Only 11 instances of them had a referential link to a syntactic noun and therefore (besides other criteria such as the impossibility to replace the phrase in the given context by a regular connective) were not considered to be DCs.

4 Ellipsis Resolution

Missing or omitted nodes in structures with an ellipsis have been reconstructed on the tectogrammatical layer of the PDT. It proved to be helpful both in the annotation of intra-sentential and inter-sentential discourse relations, namely in case of reconstructed verbal nodes. Thus, we were able to mark 1,630 relations that have in one or both arguments an elided verb. Without the ellipsis resolved, the relations could be easily overlooked in the text or it would not be possible to annotate them in the trees at all. Example 5 shows a relation with an elided verb.

(5) *Zloději nechodí po horách, ale po domácnostech.*
Thieves do not visit mountains but households.

Conclusions and Perspectives

We have presented a discourse annotation project and discussed and evaluated how it benefited from the previous annotation of the underlying syntactic structure of sentences in PDT. Its main contribution was to the partially automatic annotation of the intra-sentential discourse relations; it helped find the arguments of the discourse relations, identify the connectives and assign the discourse senses. Resolved cases of ellipses in the trees made it possible to annotate relations with no surface-present finite verb and also made it easier to determine the argument extent, both for intra- and inter-sentential relations. As for the inter-sentential discourse relations alone, the marking of a majority of discourse connectives with the semantic label PREC (reference to PREceding Context) was a helpful feature. Grammatical and textual coreference helped distinguish some of the less typical connectives.

Acknowledgments

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875) and from the Ministry of Education, Youth and Sports in the Czech Republic, program KONTAKT (ME10018) and the LINDAT-Clarin project (LM2010013).

References

Carlson, L., Marcu, D., and Okurowski, M.E. (2002). *RST Discourse Treebank*, LDC2002T07 [Corpus]. *Linguistic Data Consortium*, Philadelphia, PA, USA.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z. and Ševčíková-Razímová, M. (2006). Prague Dependency Treebank 2.0. *Software prototype, Linguistic Data Consortium*, Philadelphia, PA, USA, ISBN 1-58563-370-4, www.ldc.upenn.edu, Jul 2006.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová-Řezníčková, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K. and Žabokrtský, Z. (2005). Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. *Praha: UFAL MFF.* Available at: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/index.html.

Jínová, P., Mírovský J. and Poláková, L. (2012a). Analyzing the Most Common Errors in the Discourse Annotation of the Prague Dependency Treebank. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories (TLT 11),* Lisbon, Portugal, November 2012.

Jínová, P., Mírovský J. and Poláková, L. (2012b). Semi-Automatic Annotation of Intra-sentential Discourse Relations in PDT. In *Procedings of the 24th International Conference on Computational Linguistics (COLING 2012), ADACA Discourse Workshop,* Mumbai, India, December 2012.

Mladová, L. (2008). Diskurzní vztahy v češtině a jejich zachycení v anotovaném korpusu. *Technical report no. 2008/TR-2008-40, Faculty of Mathematics and Physics, Charles University in Prague,* Czech Republic, ISSN 1214–5521, 136 pp., Dec 2008.

Nedoluzhko, A. (2011a). Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu). *Institute of Formal and Applied Linguistics, Charles University in Prague*, Czech Republic, ISBN 978-80-904571-2-6, 268 pp., Dec 2011.

Nedoluzhko, A., Mírovský, J., Hajičová, E., Pergler, J. and Ocelák, R. (2011b). Extended Textual Coreference and Bridging Relations in PDT 2.0. *Data/software, ÚFAL MFF UK, Prague, Czech Republic*, Dec. 2011. Available at: https://ufal-point.mff.cuni.cz/xmlui/handle/11858/00-097C-0000-0005-BCCF-3, Dec 2011.

Oza, U., Prasad, R., Kolachina S., Sharma, D.M. and Joshi, A.K. (2009). The Hindi Discourse Relation Bank. In *Proc. Linguistic Annotation Workshop (LAW 2009)*, Suntec, Singapore, August 2009, pp.158-161.

Poláková (Mladová), L., Jínová, P. and Mírovský, J. (2012). Interplay of Coreference and Discourse Relations: Discourse Connectives with a Referential Component. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Eropean Language Resources Association, Istanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 146–153.

Poláková L., Jínová, P., Zikánová, Š., Bedřichová, Z., Mírovský, J., Rysová, M., Zdeňková, J., Pavlíková, V. and Hajičová, E. (2012). Manual for Annotation of Discourse Relations in the Prague Dependency Treebank. *Technical report, UFAL MFF UK, Prague, Czech Republic.* Available at: http://ufal.mff.cuni.cz/techrep/tr47.pdf.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, B. (2007). The Penn Discourse TreeBank 2.0 Annotation Manual. Available at: http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf.

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

Prasad, R., Joshi, A. and Webber, B. (2010). Realization of Discourse Relation by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010),* Beijing, China, pp. 1023–1031.

Stede, M. (2004). The Potsdam Commentary Corpus.
In *Proceedings of the ACL-04 Workshop on Discourse Annotation*, Barcelona, Spain.

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