Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT

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

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

Abstract

In the present paper, we describe in detail and evaluate the process of semi-automatic annotation of intra-sentential discourse relations in the Prague Dependency Treebank, which is a part of the project of otherwise mostly manual annotation of all (intra- and inter-sentential) discourse relations with explicit connectives in the treebank. Our assumption that some syntactic features of a sentence analysis (in a form of a deep-syntax dependency tree) correspond to certain discourse-level features proved to be correct, and the rich annotation of the treebank allowed us to automatically detect the intra-sentential discourse relations, their connectives and arguments in most of the cases.

Title and abstract in Czech

Poloautomatická anotace vnitrovětných diskurzních vztahů v PDT

Abstrakt

V tomto článku nabízíme detailní popis a evaluaci procesu poloautomatické anotace vnitrovětných textových vztahů v Pražském závislostním korpusu jako součást projektu jinak především manuální anotace všech (vnitro- a mezivětných) textových vztahů s explicitním konektorem v tomto korpusu. Potvrdil se náš předpoklad, že některé syntaktické vlastnosti analýzy věty (ve formě závislostního stromu hloubkové syntaxe) odpovídají jistým vlastnostem na úrovni analýzy textových vztahů (diskurzu). Bohatá anotace korpusu nám ve většině případů umožnila automaticky detekovat vnitrovětné vztahy, jejich konektory a argumenty.

Keywords : Tectogrammatics, PDT, Discourse annotation, Intra-sentential relations

Keywords in Czech : Tektogramatika, PDT, Anotace diskurzu, Vnitrovětné vztahy

Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA), pages 43–58,
COLING 2012, Mumbai, December 2012.

1 Introduction

Linguistic phenomena going beyond the sentence boundary have been coming into the focus of computational linguists in the last decade. Various corpora annotated with discourse relations appear, two of the first and most influential (for English) were the RST Discourse Treebank (Carlson, Marcu, Okurowski, 2002) and Penn Discourse Treebank (Prasad et al., 2008). For other languages we can mention discourse-annotated resources for Turkish (Zeyrek et al., 2010), Arabic (Al-Saif and Markert, 2010), and Chinese (Zhou and Xue, 2012). Most of these projects have raw texts as their annotation basis. In the discourse project for Czech, contrary to the others, discourse-related phenomena have been annotated directly on top of the syntactic (tectogrammatical) trees of the Prague Dependency Treebank 2.5 (henceforth PDT, Bejček et al., 2012), with the goal to make maximum use of the syntactico-semantic information from the sentence representation.

The annotation of discourse relations (semantic relations between discourse units) in PDT consisted of two steps – first, the inter-sentential discourse relations were annotated manually, second, the intra-sentential discourse relations were annotated semi-automatically. In both cases, only relations signalled by an explicit discourse connective have been annotated.

The main goal of this paper is to report in detail on the process of the semi-automatic annotation of intra-sentential discourse relations in PDT. As we assumed, some of the (not only) syntactic features already annotated in the treebank were very helpful and enabled us to perform automatic extractions and conversions.[1] Nevertheless, some manual work had to be done both before and after the annotation.

1.1 Layers of Annotation in PDT

The data in our project come from the Prague Dependency Treebank 2.5 (Bejček et al., 2012), which is a corrected and enhanced version of PDT 2.0 (Hajič et al., 2006). PDT is a treebank of Czech written journalistic texts (almost 50 thousand sentences) enriched with a complex manual annotation at three layers: the morphological layer, where each token is assigned a lemma and a POS tag, the so-called analytical layer, at which the surface-syntactic structure of the sentence is represented as a dependency tree, and the tectogrammatical layer, at which the linguistic meaning of the sentence is represented.

At the tectogrammatical layer, the meaning of the sentence is represented as a dependency tree structure. Nodes of the tectogrammatical tree represent auto-semantic words, whereas functional words (such as prepositions, auxiliaries, subordinating conjunctions) and punctuation marks have (in most cases) no node of their own. The nodes are labelled with a large set of attributes, mainly with a tectogrammatical lemma and a functor (semantic relation; e.g. Predicate (PRED), Actor (ACT), Patient (PAT),

[1] For details on the exploitation of the syntactic features during the manual annotation of the inter-sentential relations, please consult Mírovský et al. (2012).

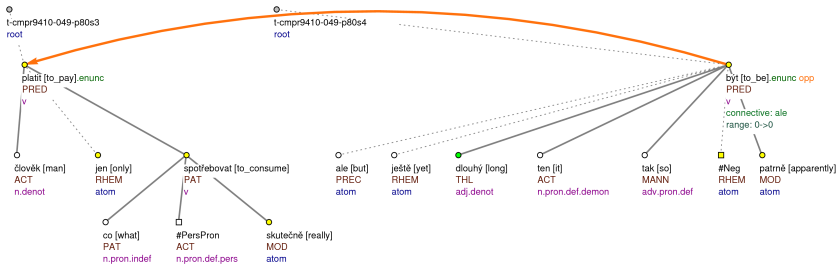

FIGURE 1 – An example of an inter-sentential discourse relation, represented by a thick arrow between roots of the arguments

Location (LOC))[2]. Additionally, the tectogrammatical layer includes the annotation of information structure attributes (sentence topic and focus, rhematizing expressions etc.).

1.2 Discourse Annotation in Two Steps

In the project of discourse annotation, 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 PDT. However, all the numbers reported in the paper refer to the training and development test parts of the whole data[3], i.e. 43,955 sentences (approx. 9/10 of the treebank).[4]

The annotation of discourse relations proceeded in two steps: First, the inter-sentential and some selected intra-sentential discourse relations were annotated manually, second, the remaining intra-sentential discourse relations were annotated (semi-)automatically, based on the information already annotated in PDT.[5]

The main theoretical principle of the annotation was the same for both phases. It was inspired partially by the lexical approach of the Penn Discourse Treebank project (Prasad et al., 2008), and partially by the tectogrammatical approach and the functional generative description (Sgall et al., 1986, Mikulová et al., 2005). A discourse connective in this view takes two text spans (verbal clauses or larger units) 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).[6]

[2] For a description of functors in PDT, see http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch07.html.
[3] as distinguished in the PDT project
[4] Thus the last tenth of the treebank, evaluation test data, remains (as far as possible) unobserved.
[5] The annotation had to proceed in this order. Our understanding what is possible to annotate automatically only formed during the manual annotation, as we got familiar with the data.
[6] For further information on the annotation guidelines, see http://ufal.mff.cuni.cz/discourse/.

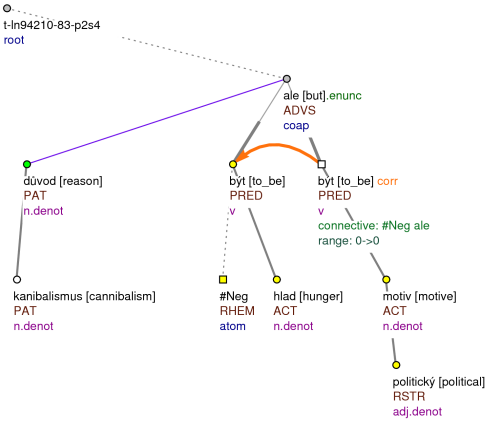

FIGURE 2 – An example of an intra-sentential discourse relation annotated during the first phase

1.2.1 Step 1: Manual Annotation (Mostly of the Inter-Sentential Relations)

The first phase of the annotation was a thorough manual processing of the treebank primarily focused on the inter-sentential relations (relations between sentences) signalled by explicit discourse connectives. Example 1 and Figure 1 show an inter-sentential discourse relation of type *opposition* with explicit connective *ale* (*but*).

(1) *Lidé chtějí platit jen to, co skutečně spotřebovali.*
Ještě dlouho tomu tak ***ale*** *patrně nebude.*

People only want to pay for what they really consumed.
But *apparently, it will not be so yet for a long time.*

Intra-sentential relations (within a sentence) were during the first phase only marked manually in cases where the discourse type could not be determined unambiguously by the tectogrammatical label (functor) and the actual discourse type was not prevailing for the given functor. For instance, the tectogrammatical label (functor) ADVS (the *adversative* relation, in our case clausal) is too general and corresponds to several finer discourse types, namely the types of *opposition*, *restrictive opposition*, *correction*, *confrontation*, and *concession*. *Opposition* is predominant among the discourse types for the functor ADVS, so it was not annotated in the first phase (and was left for the second phase)[7]. All the other discourse types for the functor ADVS were annotated manually in the first phase. The situation is illustrated by Example 2 and Figure 2; on the tectogrammatical layer, the relation between the two clauses was labelled as ADVS

[7] See Table 1 for predominant discourse types for various functors.

(functor of the coordinative node in Figure 2); the discourse type is *correction* (the relation is marked by the arrow with label *corr* in Figure 2).

(2) *Důvodem kanibalismu nebyl hlad, **ale** politické motivy.*

*The reason for the cannibalism was not hunger **but** political motives.*

For a more detailed description of the manual annotation of the treebank including the annotation evaluation see e.g. Jínová et al. 2012.

1.2.2 Step 2: Automatic Annotation of the Intra-Sentential Relations

The second phase of the annotations consisted predominantly of an automatic procedure that extracted mostly tectogrammatical features and used them directly for the annotation of intra-sentential discourse relations. The main goal was to find and mark all so far unmarked intra-sentential discourse relations.

This is the main topic of the present paper and we describe it in detail in the following sections. Section 2 briefly describes the manual preparatory work preceding the automated part of the extraction. Section 3 is devoted to the automatic annotation itself and to some practical issues connected to it. In Section 4, we mention two necessary manual corrections performed after the automatic annotation, and we evaluate our results in Section 5, which is followed by a conclusion.

2 Pre-Annotation

Two manual steps preceded the automatic annotation of the intra-sentential discourse relations: completely manually annotated selected intra-sentential relations and partially manually annotated temporal relations.

2.1 Manual Work

As explained in Subsection 1.2.1 (Example 2, Figure 2), some of the intra-sentential discourse relations were annotated manually during the first phase of the annotations. It was 510 vertical (subordinate) relations and 1,681 horizontal (coordinate)[8] intra-sentential relations. Other cases of intra-sentential relations, where the tectogrammatical annotation was adequate for the discourse interpretation, were left to the second phase. As an example, if we follow the sub-classification of the ADVS tectogrammatical label for discourse semantics mentioned above in 1.2.1, except for the relations marked previously in the manual phase, the remaining cases were all automatically set to discourse type *opposition (opp)*, see Table 1 and Section 3.1 for details.

2.2 Semi-Automatic Annotation

Finite verbs with the type of dependency being one of the temporal relations (functors TFHL, THL, THO, TSIN, TTILL, TWHEN) were pre-processed manually. For each of

[8] In dependency trees of PDT, root nodes of coordinated phrases are captured as siblings (direct children of the coordinating node), hence "horizontal" relations.

them, the type of the discourse relation was set by a human annotator, along with the direction of the relation (whether from the dependent node to its governor or the other way)[9] and the exact position of the arguments (the nodes themselves or possibly their coordinating nodes (if present)). All this information was annotated in a table and passed to the automatic script to create the discourse relations and to find and set the appropriate connective to each relation automatically. Altogether, it was 491 relations.

3 Automatic Annotation

After the manual annotation described in Subsection 2.1 and the manual preprocessing of temporal relations described in Subsection 2.2, an automatic script went through the tectogrammatical layer of the whole data of PDT, document by document, sentence by sentence and node by node.

If the node represented

- a finite verb with one of the temporal functors (TFHL, THL, THO, TSIN, TTILL, TWHEN), it was annotated using the information from the manually created table (Subsection 2.2 above).
- a finite verb with functor CAUS, COND, CNCS, AIM, CONTRD or SUBS, it became a candidate for an automatically detected vertical discourse relation.
- a coordination node with functor REAS, CSQ, ADVS, CONFR, GRAD, CONJ or DISJ, coordinating (directly or transitively) finite verbs or non-finite-verbal nodes with functor PRED[10], it became a candidate for a horizontal relation.

In all cases, the connective was detected automatically (see below in Subsection 3.4).

Vertical Relations

Candidates for a vertical relation were checked for a presence of a previously manually annotated relation; if there was none, an automatic discourse relation was created, in the basic case directly between the dependant and governing verbal nodes. If one of the nodes was a member of a coordination, more complex procedure was used to set the exact position of the arguments (see below Subsections 3.2 and 3.2.1). The discourse type and direction of the discourse arrow were set based on the tectogrammatical functor of the dependant node, see Subsection 3.1 below for details. Finally, the connective was found and set – see Subsection 3.4 for the procedure.

Horizontal Relations

Similarly, candidates for a horizontal relation were checked for a presence of a previously manually annotated relation; if there was none, an automatic discourse

[9] There is a rich variety of connectives, and also verbal aspect values and negation play a role. These features in combination determine the discourse type and also the direction of the discourse arrow (i.e. the nature of the discourse arguments: *precedence – succession*). However, as the occurrences in the data were not so many, it was faster to decide on the type of the relation and the order of arguments manually.

[10] PRED – a tectogrammatical predicate; for a list and description of all functors, please see the tectogrammatical manual: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch07.html

relation was created among the members of the coordination. A special case of multiple coordinations is discussed in 3.2.2 below. The discourse type and direction of the arrow were established based on the tectogrammatical functor of the coordinating node, again see Subsection 3.1 below for details. Subsection 3.4 describes the procedure of searching for the connective of the horizontal relation.

3.1 Functor to Discourse Type Conversion

Table 1 shows a list of tectogrammatical functors and their corresponding prevailing discourse types. After the manual annotation, the table could be (and was) used to identify the discourse type of the remaining relations. Note that it is still not a 1-1 relation, for example the discourse type *confrontation* can be signalled by two different functors (CONTRD and CONFR), as we give up the syntactic distinction of hypotactic (CONTRD) vs. paratactic (CONFR) in this respect. The transformation table was used for all automatically annotated horizontal relations (7,392 cases) and all automatically annotated vertical relations (2,599 cases).

Functor	Functor (long name)[11]	Discourse type	Discourse type (long name)
AIM	purpose	purp	purpose
CAUS	cause	reason	reason-result
CNCS	concession	conc	concession
COND	condition	cond	condition
CONTRD	confrontation	confr	confrontation
SUBS	substitution	corr	correction
ADVS	adversative relation	opp	opposition
CONFR	confrontation	confr	confrontation
CONJ	conjunction	conj	conjunction
CSQ	consequence	reason	reason-result
DISJ	disjunction	disjalt	disjunctive alternative
GRAD	gradation	grad	gradation
REAS	causal relation	reason	reason-result

TABLE 1 – Functor to discourse type automatic translation table; the first six rows represent vertical relations, the last seven rows represent horizontal relations.

3.2 Arguments with Coordinations

In PDT, coordinating expressions are represented as separate nodes and technically they are not different from other nodes representing content words. In the detection of discourse arguments, two situations needed to be treated in a special way, as described in the following two subsections.

[11] taken from the tectogrammatical manual: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch07.html

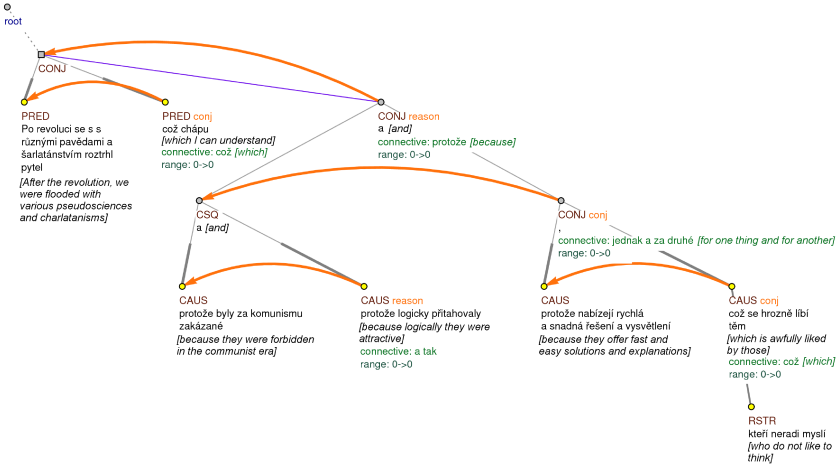

FIGURE 3 – An example of mixed coordinations in a folded mode.

3.2.1 Coordinated Structures in the Detection of the Argument Position

In many cases, an argument of a discourse relation is represented by a coordination of verbal nodes, not by the verbal nodes individually. In such cases, the position of the argument was shifted from the verbal nodes to the coordinating node. It could even happen transitively, so the topmost suitable coordination was always searched for.

Example 3 demonstrates a complex case of coordinated arguments. The situation is depicted in Figure 3, which is a tectogrammatical tree in a folded mode (nodes of the tree represent individual clauses or coordinations)[12]. All discourse annotation in the tree is a result of the automatic procedure.

(3) *Po revoluci se s různými pavědami a šarlatánstvím roztrhl pytel,* ***což*** *chápu,* ***protože jednak*** *byly za komunismu zakázané,* ***a tak*** *logicky přitahovaly,* ***a za druhé*** *nabízejí rychlá a snadná řešení a vysvětlení,* ***což*** *se hrozně líbí těm, kteří neradi myslí.*

After the revolution, we were flooded with various pseudosciences and charlatanisms, ***which*** *I can understand,* ***because for one thing****, they were forbidden in the communist era* ***and so*** *logically they were attractive,* ***and for another****, they offer fast and easy solutions and explanations,* ***which*** *is awfully liked by those who do not like to think.*

[12] For all features of the annotation tool for discourse, see Mírovský et al. (2010).

In this example sentence, five discourse relations along with their types and connectives have been automatically detected. Four of them are horizontal relations:

i. a horizontal relation of type *conj* between clauses *"Po revoluci se … roztrhl pytel"* (*"After the revolution, we were flooded … charlatanisms"*), and *"chápu"* (*"I can understand"*), with the connective *což* (*which*),

ii. a horizontal relation of type *reason* between clauses *"logicky přitahovaly"* (*"logically they were attractive"*) and *"byly za komunismu zakázané"* (*"they were forbidden in the communist era"*), with the connective *"a tak"* (*"and so"*),

iii. a horizontal relation of type *conj* between clauses *"nabízejí … vysvětlení"* (*"they offer … explanations"*) and *"se hrozně líbí … neradi myslí"* (*"is awfully liked … do not like to think"*), with the connective *což* (*which*),

iv. and a horizontal relation of type *conj* between coordinations of clauses in (ii) and (iii), with the connective *"jednak a za druhé"* (*"for one thing and for another"*).

One of them is a vertical relation:

v. a vertical relation of type *reason* between the coordination of the coordinations in (iv) and the coordination of clauses in (i), with the connective *protože* (*because*).

Cases (i), (ii) and (iii) are simple cases where the arguments are represented directly by the coordinated verbal nodes.

Case (iv) is also a relatively simple case, only a presence of a coordinated[13] finite-verb in the subtree of both the coordinated clauses needed to be checked (transitively in general).

Case (v) is a vertical discourse relation represented by an arrow between the two coordinating nodes. The relation was however signalled by four occurrences of functor CAUS, marking a linguistic (effective) dependency[14] between each of the transitively coordinated finite verbs with this functor[15] and each of their linguistic parents (finite verbs *"roztrhnout se"* (*"be flooded"*) and *chápat* (*"to understand"*)), which are also coordinated. The arguments of the relation(s) needed to be lifted to the topmost suitable coordinating nodes.[16] Thus, instead of eight discourse relations that could be created directly between the individual verbal nodes, only one overall discourse relation was created, which is a more comprehensible solution, without a loss of any information.

In all detected vertical relations, the effective parent was shifted by one coordination level 263 times, resulting in 110 discourse relations, and by two coordination levels 8 times, resulting in 3 discourse relations. The effective child was shifted by one

[13] The tectogrammatical attribute is_member serves to distinguishing coordinated and non-coordinated children of a coordinating node.

[14] The effective dependency is a linguistic dependency between nodes representing content words, taking all effects of coordinations etc. into account.

[15] verbal nodes *"být (zakázaný)"* (*"to be (forbidden)"*), *přitahovat* (*"to be attractive"*), *nabízet* (*"to offer"*), and *"líbit se"* (*"to be liked"*)

[16] Again, the tectogrammatical attribute is_member was used.

coordination level 634 times, resulting in 314 discourse relations, and by two coordination levels 61 times, resulting in 25 discourse relations.

3.2.2 Multiple Coordinations

In case of multiple coordinations (coordinations with more than two members) with only a comma as the conjunction of the first members of the coordination and a connective (often *a* (*and*)) as the conjunction of the last two members of the coordination, only the last two members form a discourse relation with an explicit connective (as we do not consider a comma to be a discourse connective). Example 4 demonstrates such a case:

(4) *Pozoroval jsem jednou jednu slečnu: seděla u PC, měla prst zabořen do klávesnice* ***a*** *evidentně se nudila.*

I watched a young lady once: she was sitting at a PC, had her finger buried in the keyboard ***and*** *evidently was bored.*[17]

Here, a discourse relation was only created between clauses "*evidentně se nudila*" ("*evidently was bored*") and "*měla prst zabořen do klávesnice*" ("*had her finger buried in the keyboard*"), with *a* (*and*) as a connective. The other discourse relations in these coordinations are considered implicit and will be annotated in the future, during the annotations of implicit discourse relations.

Multiple coordinations of this type occur 501 times in the data.

3.3 Scope of Arguments

In all intra-sentential relations, the scope of a discourse argument is defined as the effective subtree[18] of the root node of the argument (the root node of the argument can either be a finite verb or a node coordinating[19] finite verbs or another type of node with functor PRED), 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.[20]

3.4 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.

[17] The presence of a subject in a Czech clause is irrelevant for the decision whether to annotate a discourse relation or not, as Czech is a pro-drop language. Hence, the English translation of the example sentence with no subject in the last two clauses is not to be treated as a VP coordination, which would not be annotated in some projects for English like the PDTB (see Prasad, 2007)

[18] 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.

[19] possibly transitively, i.e. through other coordinating nodes

[20] For the 2,191 manually annotated intra-sentential 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.

Connectives of the vertical relations can be found among nodes from the analytical layer that correspond to the verbal root of the discourse argument on the tectogrammatical layer. All auxiliary analytical counterparts (not the lexical counterpart) of the verbal node except for auxiliary verbs and reflexive particles (*se*, *si*) become a part of the connective.

Connectives of the horizontal relations can be found on the tectogrammatical layer at the coordinating node (all its analytical counterparts, e.g. *a* (*and*), *buď – nebo* (*either – or*), etc.) or its modifiers (functor CM (conjunction modifier), e.g. *dokonce* (*even*), *přesto* (*despite of that*), or negation).

With the exception of 23 atypical cases (which were fixed manually, see Subsection 4.1), discourse connectives could be detected automatically for all 10,482 intra-sentential discourse relations. In the rest of this subsection, we point out three special cases of the connective detection.

3.4.1 Connectives with *tak*, *pak*, *potom*

For vertical relation, connectives like *jestliže – pak* (*if – then*), the second part (*pak* (*then*)) needed to be found among the effective children of the effective parent(s) of the given verbal node. They were filtered using the tectogrammatical lemma (only *tak, pak, potom* (*so, then, then*)) and the functor (only PREC or one of the temporal relations). It happened 93 times in the data.

3.4.2 Connectives with Expression *což*

The expression *což* (*which*) can represent an intra-sentential connective with the *conjunctive* meaning even though it can be inflected and plays a role of a participant of the clause structure (including a valence participant). To make it possible to distinguish the connective role of this expression automatically, grammatical coreference[21] was used. If the annotated anaphoric link from the expression *což* referred to the coordinated verbal phrase (or in a more complex case to a coordination of verbal phrases), *což* became a part of the connective. See Example 5, where *což* (*which*) refers (via the grammatical coreference) to *stal se* (*became*):

(5) *Pavlov se pak stal předsedou vlády,* ***což*** *se Klausovi přihodilo nakonec také.*

Pavlov then became the prime minister, ***which*** *after all happened to Klaus as well.*

In the data, 220 occurrences of the expression *což* have a grammatical coreference link to a finite-verb node, 11 occurrences have this link to a coordination of finite-verb nodes. Altogether, 231 discourse relations were created with *což* (*which*) as a part of the connective.

[21] Grammatical coreference was annotated in PDT for expressions where it is possible to identify the coreferred part of the text on the basis of grammatical rules (see Mikulová et. al, 2005).

3.4.3 Double Connectives

In some cases of a vertical relation where dependant finite verbal nodes are coordinated, the coordinated clauses begin with separate or different connectives, like *protože – protože* (*because – because*) in Example 6. Both the connectives become a part of the connective of the discourse relation.

(6) *… je škodlivý a ideologicky zavádějící,* ***protože*** *odráží nedůvěru v racionalitu chování každého z nás a* ***protože*** *implikuje falešnou víru ve schopnosti některých z nás vytvořit pro nás ostatní lepší, dokonalejší svět.*

… is harmful and ideologically misleading ***because*** *it reflects the mistrust in the behaviour rationality of each of us and* ***because*** *it implicates a false faith in the ability of some of us to create for the rest of us a better, more perfect world.*

This happened 69 times in our data.

4 Manual Corrections

After the automatic annotation, a few manual checks and corrections were needed. They are described in the following two subsections.

4.1 Failures in the Connective Identification

After having run the script, some manual correction turned up to be necessary in cases where the automatic search for connectives failed (23 cases in sum). These failures arose from two types of situation. First, connectives were placed on a non-typical position in the tree. Second, connectives were not present in the sentence at all. This situation is illustrated by Example 7: the last clause (*he did not pay for this*) is interpreted as a causal sentence on the tectogrammatical layer, but no connective signals this relation.

(7) *… vůbec nejhorší posádka v safari busu je smíšená: Angličan si zapomene kameru v hotelu a chce se vrátit, Francouz zuří, za tohle neplatil!*

… the absolutely worst crew in a safari bus is a mixed one: the Englishman forgets his camera in the hotel and wants to go back, the Frenchman is furious, he did not pay for this!

In the first type of situation, the connective was added manually (we count these relations under the manually annotated ones), in the second type (as in Example 7), the whole relation was deleted for violation of the surface-present connective rule.

4.2 Clauses Depending on a Noun Phrase or an Infinitive

Solely manual treatment required those types of constructions where the dependent clause with discourse semantics was related to a complex predicate structure containing a noun phrase or an infinitive. Only semantics allows to distinguish cases where the dependent clause is related to the whole predicate structure from those related only to an infinitive or a noun phrase. Consider Examples 8 and 9. In both structures, the dependent clause is a child-node of the infinitive, but only in Example 8 it is

semantically related to the whole predicate structure *"je ochoten povolit"* (*"is willing to permit"*). In Example 9 the dependent clause is semantically related only to the noun phrase *"připravenost odpovědět silou"* (*"readiness to respond with force"*). As we only annotate discourse relations between text spans with finite verbs, only in Example 8 a discourse relation was annotated.

(8) *Srbský prezident Slobodan Miloševič je ochoten povolit mezinárodní kontrolu své blokády bosenských Srbů,* ***pokud*** *bude obdobná kontrola uplatněna i na hranicích Chorvatska a Bosny.*

The Serbian president Slobodan Milosevic is willing to permit an international inspection of his blockade of the Bosnian Serbs ***if*** *a similar control is applied also on borders of Croatia and Bosnia.*

(9) *Zdůraznili však také připravenost odpovědět silou,* ***pokud*** *opozice bude trvat na použití zbraní.*

However, they also emphasised their readiness to respond with force ***if*** *the opposition will insist on the use of weapons.*

There were 146 cases with such a dependent clause related to the whole predicate structure and 73 occurrences where it was not the case.

5 Summary

Table 2 shows the summary of all relations annotated during both phases of the project, and gives detailed numbers of various "types" of the intra-sentential relations. The last row of the table presents the whole number of all annotated discourse relations of any type.[22]

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

TABLE 2 – Overview of discourse relations annotated in PDT

We were able to automatically convert 9,991 (2,599 vertical and 7,392 horizontal) tectogrammatical dependencies into discourse relations, along with all properties of the relations (i.e. the position of arguments, the discourse type and the connective). For

[22] Let us emphasize again: although everything was done on the whole PDT data, all reported numbers only refer to the training and development test parts of the data (9/10 of the treebank, 43,955 sentences).

another 491 vertical dependencies, the discourse type, the order of arguments and their position according to possible coordinations were set manually, as explained in Subsection 2.2, while the rest of the work with these relations was also done automatically; we count these relations as semi-automatic. Mostly during the first phase of the annotation, 2,191 (510 vertical and 1,681 horizontal) intra-sentential discourse relations were annotated completely manually. After the automatic procedure, non-typical connectives needed to be fixed in 23 cases, and 146 relations between a dependent clause and a complex predicate structure needed to be manually added, as explained in Section 4.

Conclusion

In the paper, we have presented in detail the second phase of the discourse annotation project in the Prague Dependency Treebank 2.5, namely the semi-automatic annotation of intra-sentential discourse relations marked by an explicit connective. In the preceding first phase of the project, the whole treebank was processed manually and all inter-sentential relations were marked by a human annotator. Also all intra-sentential relations were assessed manually and those relations whose discourse semantics was not unambiguously inferable from the tectogrammatical information were annotated. After the manual annotation, the tectogrammatical interpretation of the remaining relations conveyed the discourse semantics properly and, in the second phase of the project, all these remaining intra-sentential relations were annotated semi-automatically or automatically. During the automatic part of the annotation, the presence of a discourse relation, the exact position of its arguments, its discourse type and the connective were automatically detected, using the annotation of the deep-syntax dependency trees at the tectogrammatical layer of PDT. As a final step, a few manual checks and corrections were performed.

We have also discussed interesting theoretical observations revealed during the semi-automatic annotation, namely to what extent a syntax-based discourse analysis is automatically processible and what are the special (and so linguistically interesting) cases that require more attention.

The annotated data (both intra- and inter-sentential relations) was published in the autumn of 2012 under the same licence as the underlying PDT 2.5, i.e. the Creative Commons licence[23]. It is available (downloadable) from the repository of LINDAT-Clarin – Centre for Language Research Infrastructure in the Czech Republic[24].

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).

[23] http://creativecommons.org
[24] http://www.lindat.cz

References

Al-Saif, A; Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, pp. 2046–2053.

Bejček, E., Panevová, J., Popelka, J., Smejkalová, L., Straňák, P., Ševčíková, M., Štěpánek, J., Toman, J., Žabokrtský, Z., Hajič, J. (2012). Prague Dependency Treebank 2.5. *Data/software, Charles University in Prague,* Czech Republic, http://ufal.mff.cuni.cz/pdt2.5/.

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.

Mírovský, J., Jínová, P., Poláková, L. (2012). Does Tectogrammatics help the Annotation of Discourse? In *Procedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December 2012.

Jínová, P., Mírovský, J., Poláková, L. (2012). 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),* Lisboa, Portugal.

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.

Mírovský, J., Mladová, L., Žabokrtský, Z. (2010). Annotation Tool for Discourse in PDT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, pp. 9-12.

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, pp. 2961–2968.*

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

Zeyrek, D., Demirşahin Işın, Çallı A. B. S., Balaban H. Ö., Yalçınkaya İ., & Turan Ü. D. (2010). The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*, Uppsala, Sweden, pp. 282–289.

Yuping Zhou and Nianwen Xue. (2012). PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*ACL 2012),* Jeju, Republic of Korea, pp. 69–77.