Introducing the Prague Discourse Treebank 1.0

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová,
Šárka Zikánová and Eva Hajičová
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
{polakova|mirovsky|nedoluzko|jinova|zikanova|hajicova}@ufal.mff.cuni.cz

Abstract

We present the Prague Discourse Treebank 1.0, a collection of Czech texts annotated for various discourse-related phenomena "beyond the sentence boundary". The treebank contains manual annotations of (1), discourse connectives, their arguments and senses, (2), textual coreference, and (3), bridging anaphora, all carried out on 50k sentences of the treebank. Contrary to most similar projects, the annotation was performed directly on top of syntactic trees (from the previous project of the Prague Dependency Treebank 2.5), benefiting thus from the linguistic information already existing on the same data. In this article, we present our theoretical background, describe the annotations in detail, and offer evaluation numbers and corpus statistics.

1 Introduction and Motivation

Large collections of gold standard language data are known to build an indispensable base for many NLP algorithms. Reliable morphological tagging and syntactic analysis (phrasal or dependency) are nowadays quite a standard information in language corpora released all over the world. With the gradually increasing interest in modeling discourse structure or using various discourse features[1] in different NLP tasks (anaphora resolution, summarization, MT), also the development of resources aimed at representing various discourse-related aspects has gained on importance. Moreover, both theoretical discourse research and NLP algorithms can benefit from a reliable **multi-dimensional** analysis of the data (Webber et al., 2003, Stede, 2004). There are already several elaborate theoretical concepts on discourse coherence brought to life in real-data annotation (see Sections 1.1 and 1.2). Still, it is only in recent years that large-scale corpora with manual annotations of sentential **and** discourse level phenomena have become available. Even fewer such corpora exist that combine more types of manual discourse-level annotations.

In this paper, we present a large-scale manual annotation project for Czech in which, apart from the "standard" analysis of a sentence (morphology, synctactic trees), several discourse phenomena are marked, all over the same data: pronominal, nominal and zero[2] coreference, discourse connectives (henceforth DCs) and the semantic relations they express, and the associative relations of the so-called bridging anaphora.

The paper is structured as follows: In Sections 1.1 and 1.2, brief overviews of recent projects concerning discourse relations and coreference + bridging anaphora are described, respectively. In Section 2, data and tools used in Prague Discourse Treebank (PDiT) are introduced. Section 3 describes the annotation scenario and is followed by evaluation of the project in comparison with similar projects (Section 4) and basic distribution numbers (Section 5). We conclude with discussion (Section 6).

1.1 Corpora of Discourse Relations

The first attempts in representing discourse structure date over a decade back. One of very first and most influential projects was the RST-Treebank (Carlson et al., 2001), an annotation project over the English texts of Wall Street Journal. In accordance with the Rhetorical Structure Theory of Mann and Thompson (1988), the whole document is represented as a single tree-like structure. Wolf and Gibson (2005) propose a less con-

[1] The term of *discourse* in this paper is used in two meanings. The broader interpretation is roughly equal to *text* (as in *discourse structure, discourse features* or *discourse coherence*) whereas the narrower sense denotes semantic relations between propositions (as in *discourse relations*).

[2] Czech is a pro-drop language. The restored ellipses in the underlying sentence analysis allow us to annotate zero forms as co-referential.

International Joint Conference on Natural Language Processing, pages 91–99,
Nagoya, Japan, 14-18 October 2013.

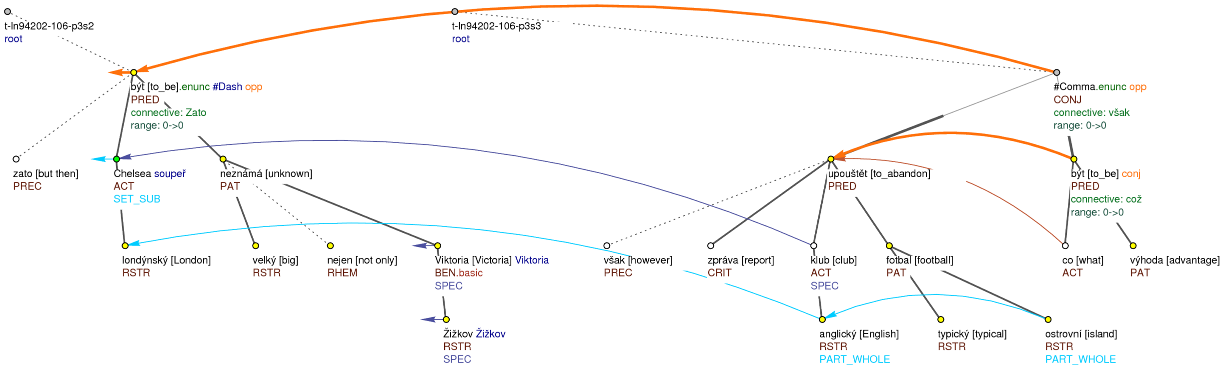

Figure 1. Annotation of two sentences. Discourse relations are represented by thick orange arrows, textual coreference by dark blue slim arrows, bridging anaphora by light blue slim arrows. Grammatical coreference (the only one in the figure is between nodes *co [what]* and *upouštět [to abandon]*) is represented by a brown slim arrow.

strained model in Discourse Graphbank by giving up the requirement of a tree-structure. These approaches are referred to as "deep discourse parsing" or modeling of global coherence (whole document = one connected structure) in contrast to the so-called "shallow discourse parsing" or local coherence modeling of the lexically grounded approaches, which are based on identification of discourse markers and relations they express. The most influential of the latter is the Penn Discourse Treebank (for English, PDTB, Prasad et. al., 2008) with several subsequent similarly aimed corpora for different languages, the project presented here being one of them.

Resources manually annotated for (some type of) discourse phenomena are already available or work-in-progress for various languages, including Chinese (Zhou and Xue, 2012), Arabic (Al-Saif and Markert, 2010), Turkish (Zeyrek et al., 2010), Hindi (Oza et al., 2009), French (Afantenos et al., 2012, Danlos et al., 2012), German (Stede, 2004, Gastel et al., 2011) and others. Additionally, the relevance of the PDTB annotation concept was further tested on specific domains, e.g. on spoken dialogs (Italian, Tonelli et al., 2010) and on biomedical texts (English, Prasad et al., 2011).

1.2 Corpora of Coreference and Bridging Relations

There is a number of different large-scale annotated corpora for coreference and anaphoric relations. The largest annotated corpora for English include MUC (Hirschman and Chinchor, 1997), ACE (Doddington et al., 2004), OntoNotes (Pradhan et al., 2007), GNOME (Poesio, 2004), ARRAU (Poesio and Artstein, 2008). The coreference annotations for other languages than English are more limited. The most well-known corpora including anaphoric information are AnCora (Recasens and Martí, 2009) for Spanish and Catalan, VENEX (Poesio et al., 2004a) for spoken and written Italian, the Italian Live Memories Corpus (Rodríguez et al., 2010), TüBA-D/Z (Hinrichs et al., 2004) and Postdam Commentary Corpus (Stede, 2004, Krasavina and Chiarcos, 2007) for German, and some others.

Early work on bridging relations dates back to the mid-70s. Clark (1975) documents several ways in which an inference is needed to understand the meaning intended by the speaker. Clark names several types of bridging relations such as set-membership, part-whole, roles, reasons and consequences. Bridging relations have been later investigated by Poesio et al. (1997, 2004b). The annotation of bridging relations in different projects includes different types of relations. In the GNOME corpus (Poesio, 2004), such bridging relations as set-membership, subset, and part-whole are annotated. The Copenhagen Dependency Treebank (Korzen and Buch-Kromann, 2011) has a very detailed annotation scheme based on general semantic roles. Another way to capture bridging relations is to define them vaguely, e.g. as a reference which is made to a subpart of an object that has already been mentioned in the discourse (Hendrickx et al., 2011) or to mark as bridging all non-coreferent anaphoric references. The last approach was used in Hou et al. (2013), providing a reasonably sized and reliably annotated corpus for English.

To our knowledge, there are only few corpus projects portraying phenomena "beyond the sentence boundary" that gather different types of textual information, or, in other words, offer some kind of multi-dimensional discourse annotation. The texts of Wall Street Journal have undergone various annotations but they arose within different projects and frameworks – rhet-

TEMPORAL	CONTINGENCY	CONTRAST	EXPANSION
synchronous	reason – result	confrontation	conjunction
asynchronous	*pragmatic reason – result*	opposition	exemplification
	condition	*pragmatic contrast*	specification
	pragmatic condition	restrictive opposition	equivalence
	explication	concession	generalization
	purpose	correction	conjunctive alternative
		gradation	disjunctive alternative

Table 1: Distribution of discourse types in the data

orical structure analysis in RST-Treebank (385 WSJ articles), Discourse Graphbank (135 texts from AP Newswire and WSJ), Penn Discourse Treebank 2.0 (2,159 WSJ articles), OntoNotes (a substantial portion of the WSJ-Penn Treebank annotated for coreference) etc. A multi-dimensional analysis within a single project was conducted for French in AnnoDis (Afantenos et al. 2012, an intersection of all annotations on 13 articles), for German in the Potsdam Commentary Corpus (Stede, 2004, 170 texts), and lately in TüBa-D/Z (Gastel et al., 2011, 919 sentences in 31 articles). These projects include inter alia some particular version of a "global" discourse analysis, annotation of connectives and their senses, and coreference annotation.

2 Data and Tools

As the base data for the annotation, we used the Prague Dependency Treebank 2.5 (PDT, Bejček et al., 2012), which is an update of the Prague Dependency Treebank 2.0 (Hajič et al., 2006). It is a treebank of almost 50 thousand sentences of Czech newspaper texts, annotated manually on three levels of annotation: morphological, analytical and tectogrammatical. The annotation of a sentence at the highest, tectogrammatical layer captures the deep syntax and the information structure of a sentence and is represented by a dependency tree.

For the annotation of discourse relations, textual coreference and bridging anaphora, we used several extensions to a highly customizable tree editor TrEd (Pajas and Štěpánek, 2008). Technically, each of the annotated relations is represented as an arrow connecting two tectogrammatical nodes. The two nodes represent the two arguments of the relation, i.e. typically the subtrees of the nodes. All information about the relation is kept in a set of dedicated attributes at the initial node of the relation, containing a unique identifier of the target node of the relation, type of the relation, and other pieces of information (depending on the relation, e.g. a connective for the discourse relation). The relation is depicted as a curved arrow between the nodes, see Figure 1. For details on the annotation tool for discourse, see Mírovský et al. (2010a), for details on the annotation tool for textual coreference and bridging anaphora, see Mírovský et al. (2010b).

3 Annotation

The following subsections 3.1 and 3.2 describe the annotation principles for the two subprojects in PDiT, the annotation of discourse relations and the annotation of textual coreference and bridging anaphora. Detailed descriptions of the annotation guidelines can be found in annotation manuals (Poláková et al., 2012a, Nedoluzhko et al., 2011). Figure 1 shows the annotation of two sentences in Example 1 in all these aspects.

(1) *Zato londýnská Chelsea je velkou neznámou nejen pro Viktorii Žižkov. Podle zpráv **však** anglický klub upouští od typického ostrovního fotbalu, což by mohlo být výhodou.*

*But then London Chelsea is a big unknown not only for Victoria Žižkov. According to reports, **however**, the English club abandons the typical island football, which could be an advantage.*

3.1 Discourse

Annotating discourse relations in PDiT is inspired by the PDTB lexical approach of connective identification (Prasad et. al., 2008) but it also takes advantage of the Prague tradition of dependency treebanking. This means in practice that some discourse information (intra-sentential) could have been extracted from the previous rich annotation of syntax, with only minor enhancements (Jínová et al., 2012b). In the first release of PDiT, we only focused on discourse relations indicated by overly present (explicit) discourse connectives, i.e. expressions like *but, however, as a result, even though* etc.[3] Every DC is thought of as a discourse-level predicate that

[3] Some remarks on annotation of the implicit DCs and of the so-called alternative lexicalizations of connectives (AltLex) are added in the discussion in Section 6.

takes two discourse units as its arguments. Only discourse relations connecting clausal arguments (with a predicate verb), i.e. not those between nominalizations or deictic expressions were annotated in version 1.0. Additionally, the Prague discourse annotation includes marking of list structures (as a separate type of discourse structure) and marking of some smaller text phenomena: article headings, figure captions, non-coherent texts like collections of news etc.

The annotation of discourse relations consisted of two phases, first being manual and the subsequent including automatic extraction of relevant syntactic features. For the manual part, the annotators had at their disposal both plain text and the tree structures, the annotation itself was carried out on syntactic (tectogrammatical) dependency trees, as we did not want to lose connection with and information from the analyses of previous levels. Intra-sentential discourse relations, i.e. those that had already been captured within the syntactic (tectogrammatical) analysis, were only to be newly annotated if their discourse semantics differed from the tectogrammatical interpretation (Jínová et al., 2012b), otherwise they were automatically extracted and mapped onto the discourse annotation.

Automatic Extraction of Syntactic Features

An automatic procedure was designed to extract discourse-relevant features from the syntactic level of description, i.e. the intra-sentential discourse relations. As mentioned earlier, the tectogrammatical tree structures offer some types of information that can be transferred to the discourse-level annotation. In general, this concerns subordinate syntactic relations between clauses with labels like causality, conditionality, temporality, concession etc.; and coordinate syntactic relations between clauses of one sentence with selected coordinative labels like conjunction, disjunction, opposition or contrast, confrontation etc. These relations were semi-automatically mapped onto the discourse annotation. (Jínová et al., 2012b).

Semantic labels

The Prague discourse label set was inspired by the tectogrammatical functors (Mikulová et al., 2005) and also by Penn sense tag hierarchy (Miltsakaki et al., 2008). Table 1 shows the discourse-semantic label set used for PDiT 1.0. The four main semantic classes, Temporal, Contingency, Contrast (Comparison) and Expansion are identical to those in PDTB but the hierarchy itself is only two-level. The third level is captured by the direction of the discourse arrow. The annotators, unlike in the Penn approach, were not allowed to only assign the major class, they always had to decide for a single relation within one of the classes.[4] Within these four classes, the types of the relations partly differ from the Penn types and go closer to Prague tectogrammatical functors and/or are a matter of language-specific distinctions. Compared to the PDTB label set, we added the categories of *purpose* and *explication* in the Contingency group and *restrictive opposition* and *gradation* to the Contrast group. In the PDTB, four pragmatic meanings are distinguished and annotated: *pragmatic cause, condition, contrast* and *concession*. In the Prague scenario, three pragmatic senses were annotated, pragmatic concession and pragmatic contrast joined to one group, for the lack of reliable distinctive features.[5]

Post-annotation checks and fixes

After the manual annotation of discourse relations was finished, some checks turned up to be necessary, especially for relations whose nature revealed to be more complicated in real data than we had expected on the basis of linguistic handbooks. After having collected all examples of these relations (namely *specification*, *explication*, *generalization*, *exemplification* and *equivalence*) in our data and established more complex definitions of their nature, annotation of these relations was manually unified in the whole data. Also some DCs required unification via post-annotation. Additionally, the part of the data which was annotated first was fully re-annotated at the end since we expected it might have suffered from initial inexperience of the annotators.

Results of the automatic extraction were checked randomly on several hundreds of examples. All discrepancies found were integrated in an automatic script (treatment of multiple DCs, multiple coordinations etc.). Only two situations required manual checks and fixes: i) Due to a complicated situation in a tree, the automatic extraction failed in 23 cases of DC identification (opposed to 10,482 cases with correct identification). ii) Solely manual treatment was necessary for constructions with a discourse-relevant clause dependent on a complex predicate structure with

[4] In special cases, they had the option to assign an additional secondary relation.

[5] It may be that different text types require slightly different sets of semantic labels. For instance, some discourse projects use a more fine-grained set of pragmatic senses (e.g. for spoken dialogs).

an infinitive or a noun phrase. In such cases only semantics allowed to distinguish if the clause is related to the whole structure or only to the infinitive or noun phrase.[6]

3.2 Coreference and Bridging Relations

In PDiT 1.0, two types of coreference (grammatical and textual) and six types of bridging relations are marked. The **grammatical coreference** typically occurs within a single sentence, the antecedent being able to be derived on the basis of grammatical rules of a given language (Czech). It includes relative pronouns, verbs of control, reflexive pronouns, reciprocity and verbal complements (Mikulová et al., 2005). **Textual coreference** marks coreferential relations between language expressions referring to the same discourse entity when the reference is not expressed by grammatical means alone, but also via context. Anaphoric (occasionally cataphoric) relations are expressed by various linguistic means (pronouns, synonyms, generalizing nouns etc.). Textual coreference has been annotated in two time periods. First, the so-called pronominal textual coreference was manually annotated. It was restricted to cases in which a demonstrative *this* or an anaphoric pronoun of the 3rd person, also in its zero form, are used (Kučová and Hajičová, 2004). Afterwards, the annotation of textual coreference was extended to cases where the anaphoric expression is represented by other means such as full noun phrases, adverbs (*there, then* etc.) and some types of numerals and pronouns left out during the first stage (Nedoluzhko et al., 2013).

The textual coreference is further classified into two types – coreference of noun phrases with specific (type SPEC) or generic (type GEN) reference. Compare examples (2) and (3):

> (2) *Mary and John went together to Israel, but Mary* [type SPEC] *had to return because of the illness.*
>
> (3) *Dogs bark. This is the way how they* [type GEN] *express their emotions.*

Discourse deixis (reference to a non-nominal antecedent) is annotated as a textual coreference link when referring to a clause or a sentence. If a noun phrase endophorically refers to a discourse segment that is larger than one sentence or it is understood by inferencing from a broader context, the antecedent is not specified.[7]

A specifically marked link for **exophora** denotes that the referent is "out" of the co-text, it is known only from the actual situation. In the same way as for segments, the new nominal and adverbial links were added.

For the **bridging relations**, the following types are distinguished: part-of relation (*room - ceiling*), set – subset (*students – some students*) and FUNCT (*trainer – football team*) traditional relations, CONTRAST for coherence relevant discourse opposites (e.g. *this year – next year*), ANAF for explicitly anaphoric relations without coreference (*second world war – at that time*) and the further underspecified group REST, which is mainly used to capture such types of bridging relations as location – inhabitants or event – argument. A more detailed description of the types can be found in Nedoluzhko and Mírovský (2011).

Automatic Preannotation

For the textual coreference, only a limited preannotation was carried out: We used a list of pairs of words that with a high probability form a coreferential pair in texts. Most of the pairs in the list consist of a noun and a derived adjective, which are different in Czech, e.g. Praha – pražský (in English: Prague – Prague, like in the sentence: *He arrived in Prague and found the Prague atmosphere quite casual*). The rest of the list is formed by pairs consisting of an abbreviation and its one-word expansion, e.g. ČR – Česko (similarly in English: USA – States). The whole list consists of more than 6 thousand pairs obtained automatically from the morphological synthesizer for Czech, manually checked and slightly extended.

4 Inter-Annotator Agreement

Several annotators annotated the data but (for obvious reasons of limited resources) each part of the data has only been annotated by one of them. Only 4% of the data (44 documents, 2,084 sentences) have been annotated in parallel by two annotators of discourse relations, and 3% (39 documents, 1,606 sentences) have been annotated in parallel by two annotators of textual coreference and bridging anaphora. We used the parallel (double) annotations for measuring the inter-annotator agreement, and for analyzing the most common errors, i.e. difficult parts of the annotation.

[6] For more details, see Jínová et al. (2012b).

[7] This decision is considered to be provisional. The antecedents are supposed to be specified in further phases of the annotation.

To evaluate the inter-annotator agreement on texts annotated in parallel by two annotators, we used several measures. The connective-based F1-measure (Mírovský et al., 2010c) was used for measuring the agreement on the recognition of a discourse relation, the chain-based F1-measure was used for measuring the agreement on the recognition of a coreference or bridging relation. A simple ratio and Cohen's κ were used for measuring the agreement on the type of the relations in cases where the annotators recognized the same relation.[8]

In the connective-based measure, we consider the annotators to be in agreement on recognizing a discourse relation if the two connectives they mark (each of the connectives marked by one of the annotators) have a non-empty intersection (technically, a connective is a set of tree nodes). For details, see Jínová et al. (2012a).

In the chain-based measure, we consider the annotators to be in agreement on recognizing a coreference or a bridging relation if two nodes connected by an arrow by one of the annotators have also been connected by the other annotator; coreference chains are taken into account, i.e. it is sufficient for the agreement if the arrow starts in or goes to a node that is coreferentially connected (possibly transitively) with the node used for the relation by the other annotator.

Table 2 shows the results of the inter-annotator agreement measurements.

relation	F1	agreement on types	Cohen's κ
discourse	0.83	0.77	0.71
text. coref.	0.72	0.90	0.73
bridging	0.46	0.92	0.89

Table 2: Inter-annotator agreement

Comparison of the inter-annotator agreement with other similar projects is difficult, as the projects usually use different annotation schemes and different scores. Nevertheless, some comparisons can be done:

The simple ratio agreement on types in discourse relations (0.77 on all parallel data, the third column of Table 2) is the closest measure to the way of measuring the inter-annotator agreement used on subsenses in the Penn Discourse Treebank 2.0, reported in Prasad et al. (2008). Their agreement was 0.8.

In the annotation of coreference relations in OntoNotes, the inter-annotator agreement on English was 80.9 for newspaper texts and 78.4 for magazine texts. On Chinese, the agreement was 73.6 for newspaper texts and 74.9 for magazine texts (reported in Pradhan et al. 2012). These numbers can be compared with our chain-based F1 measure (0.72 in the second column of Table 2), as it is similar to the MUC-6 score they used.

As to the bridging anaphora, we can compare our chain-based F1 score (0.46 in the second column of Table 2) to F1 score on recognition of bridging relations reported for the annotation of the COREA corpus (Dutch texts); their agreement on newspaper texts was 0.39 (reported in Hendrickx et al., 2011).

5 The Corpus in Numbers[9]

Table 3 shows total numbers of annotated relations in the whole data of PDiT.

relation	count
discourse relations	20,542
- discourse inter-sentential	6,195
- discourse intra-sentential	14,347
textual coreference	87,299
grammatical coreference[10]	23,272
bridging anaphora	33,154

Table 3: Total numbers of annotated relations in PDiT

bridging type	count
ANAF	847
CONTRAST	2,305
FUNCT_P	516
PART_WHOLE	2,017
P_FUNCT	1,743
REST	2,226
SET_SUB	13,106
SUB_SET	5,885
WHOLE_PART	4,509
total	**33,154**

Table 4: Distribution of bridging types in PDiT

In addition to the numbers in Table 3, there have been annotated 445 members of lists, 4,188 headings, 1,505 coreference relations to segment and 689 references out of the text (exophora).

[8] In all our measurements, only inter-sentential discourse relations have been counted, as the intra-sentential relations were mostly annotated automatically.

[9] Please note that 1/10 if the PDT/PDiT data has been designated to evaluation tests. Numbers presented in this section include also this part of the data. Therefore, these numbers should not be used in any experiments tested on the evaluation test data of PDT/PDiT!

[10] mostly annotated already in PDT

Table 4 shows a distribution of bridging types annotated in PDiT. Table 5 shows the total number of individual discourse types annotated in PDiT.

discourse type	full name	count
conc	concession	878
cond	condition	1,369
confr	confrontation	654
conj	conjunction	7,551
conjalt	conj. alternative	90
corr	correction	440
disjalt	disj. alternative	270
equiv	equivalence	104
exempl	exemplification	142
explicat	explication	225
f_cond	pragm. condition	16
f_opp	pragm. contrast	50
f_reason	pragm. reason	40
gener	generalization	106
grad	gradation	430
opp	opposition	3,209
preced	asynchronous	808
purp	purpose	414
reason	reason-result	2,626
restr	restr. opposition	269
spec	specification	627
synchr	synchronous	222
other	*other*	2
total		**20,542**

Table 5: Distribution of discourse types in PDiT

6 Discussion

In the first release of PDiT, the annotation of discourse relations is limited to relations expressed by explicit DCs (coordinating conjunctions, particles, adverbs etc.), other tags between adjacent sentences were not inserted, unlike in some similar projects. Alternative lexicalizations (AltLex) are not annotated in PDiT, their thorough analysis is a recent work in progress. Entity-based relations (EntRel) are, in our view, a matter of coreference and bridging annotation.

Implicit connectives

Annotation of implicit connectives has been in all known attempts a problematic task, as the IAA numbers are rather low. For implicit connectives (not present on the surface, a DC must be "inferred" from the context), we conducted an experimental annotation of 100 sentences, trying to remove factors known as repeatedly disturbing.[11] The annotators agreed in 49% on type of the relation. If only the distinction between *any* discourse relation on one side and coref + bridging relation on the other side was taken into consideration, the agreement was slightly higher – 58%. The most problematic issue revealed to be distinguishing between elaborative relations and relations based only on coreference. The restriction of the annotation only to slots between adjacent sentences was found useful for simplifying the annotation but it did not always match the annotators' intuition where the argument borders should be (e.g. if only the sentence-last dependent clause relates to the following sentence). Although the annotators were able to agree in most cases after discussion, the results convinced us to reconsider the annotation setting for implicit DCs before any future annotation.

Another phenomenon not present in PDiT in comparison with PDTB is attribution. We believe that this information can be at least partially obtained from syntactic features of the syntactic layers of PDT (e.g. attributes for direct speech, parentheses, verbal valency etc.).

7 Conclusion

We described the Prague Discourse Treebank 1.0, PDiT 1.0, a large collection of Czech texts that offers a rare combination of manual annotations of discourse relations, textual coreference and bridging anaphora. PDiT 1.0 is an extension of PDT 2.5 and all the annotation presented in this paper was carried out on the dependency trees of the tectogrammatical (deep syntax) layer. It was released in November 2012 under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License and it is available at the LINDAT-Clarin repository[12] (Poláková et al., 2012b).

Recently, we focus on extensions of the annotation for the upcoming release of PDT 3.0. A genre classification of the corpus texts for the purposes of data clustering in automatic experiments has been finished. Annotation of alternative lexicalizations (AltLex) and anaphoric expressions of 1st and 2nd person are in progress.

Acknowledgment

We gratefully acknowledge support from the Grant Agency of the Czech Republic (projects n.

[11] The annotation was carried out by two most experienced annotators, the chosen text types were from an accessible domain (cultural event description), the texts were short, up to 35 sentences each. Another option would be to underspecify the sense hierarchy but we did not do that. Instead, we allowed for labels coref, bridging (=EntRel) and NoRel.

[12] http://hdl.handle.net/11858/00-097C-0000-0008-E130-A

P406/12/0658 and P406/2010/0875), the LINDAT-Clarin project (LM2010013) and SVV of the Charles University (267 314).

References

S. D. Afantenos, N. Asher, F. Benamara et al. 2012. An empirical resource for discovering cognitive principles of discourse organization: the ANNODIS corpus. In: *Proceedings of LREC 2012*, Istanbul, Turkey.

A. Al-Saif, K. Markert. 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.

E. Bejček, J. Panevová, J. Popelka et al. 2012. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, Mumbai, India, pp. 231–246.

L. Carlson, D. Marcu, M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue,* Eurospeech 2001.

H. H. Clark. 1975. Bridging. In: *The Conference on Theoretical Issues in NLP*, pp. 169–174.

L. Danlos, D. Antolinos-Basso, C. Braud et al. 2012. Vers le FDTB: French Discourse Tree Bank In: *Actes de la conférence conjointe JEP-TALN-RECITAL*, Grenoble, France, volume 2 : TALN, 2, pp. 471–478.

G. Doddington, A. Mitchell, M. Przybocki et al. 2004. The Automatic Content Extraction (ACE) program – tasks, data, and evaluation. In: *Proceedings of LREC 2004*, Lisbon.

A. Gastel, S. Schulze, Y. Versley et al. 2011. Annotation of Explicit and Implicit Discourse Relations in the TüBa-D/Z Treebank. In: *Multilingual Resources and Multilingual Applications, Proceedings of the German Society of Computational Linguistics and Language Technology (GSCL) 2011*. Hamburg, pp. 99–104.

J. Hajič, J. Panevová, E. Hajičová et al. 2006. *Prague Dependency Treebank 2.0.* Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, http://www.ldc.upenn.edu, Jul 2006.

I. Hendrickx, O. De Clercq, V. Hoste. 2011. Analysis and Reference Resolution of Bridge Anaphora across Different Text Genres. In: *Anaphora Processing and Applications.* Lecture Notes in Computer Science Volume 7099, pp. 1–11.

E. Hinrichs, S. Kübler; K. Naumann et al. 2004. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*. Tübingen.

L. Hirschman, N. Chinchor. 1997. *MUC-7 Coreference Task Definition – Version 3.0*.

Y. Hou, K. Markert, M. Strube. 2013. Integrating semantics and saliences for brodging resolution using Markov logic. In *NAACL 2013* to appear.

P. Jínová, J. Mírovský, L. Poláková. 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.

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

I. Korzen, M. Buch-Kromann. 2011. Anaphoric relations in the Copenhagen dependency treebanks. In: *Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena.* DGfS Workshop, pp. 83–98.

O. Krasavina, Ch. Chiarcos. 2007. PoCoS –Potsdam Coreference Scheme. In *Proceedings of the Linguistic Annotation Workshop*, Prague.

L. Kučová, E. Hajičová. 2004. Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium*, S. Miguel.

W.C. Mann, S. A. Thompson. 1988. Rhetorical structure theory. Toward a functional theory of text organization. In: *Text, 8(3)*:243–281.

M. Mikulová et al. 2005. *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. The Annotation Guidelines.* Prague: UFAL MFF. Available at: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

E. Miltsakaki, L. Robaldo, A. Lee et al. 2008. Sense Annotation in the Penn Discourse Treebank. In: *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics*.

J. Mírovský, L. Mladová, Z. Žabokrtský. 2010a. Annotation Tool for Discourse in PDT. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Tsinghua University Press, Beijing, China, ISBN 978-7-302-23456-2, pp. 9–12.

J. Mírovský, P. Pajas, A. Nedoluzhko. 2010b. Annotation Tool for Extended Textual Coreference and Bridging Anaphora. In: *Proceedings of the 7th International Conference on Language Resources* *and Evaluation (LREC 2010)*, Valletta, Malta, ISBN 2-9517408-6-7, pp. 168–171.

J. Mírovský, L. Mladová, Š. Zikánová. 2010c. Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Tsinghua University Press, Beijing, China, pp. 775–781.

A. Nedoluzhko, J. Mírovský, M. Novák. 2013. A Coreferentially annotated Corpus and Anaphora Resolution for Czech. To appear in *Computational Linguistics and Intellectual Technologies.* Papers from the Annual International Conference "Dialogue 2013". Moskva.

A. Nedoluzhko, J. Mírovský. 2011. *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank.* Technical report no. 2011/44, ÚFAL MFF UK, Prague, Czech Republic, 69 pp.

U. Oza, R. Prasad, S. Kolachina et al. 2009. The Hindi Discourse Relation Bank. In: *Proc. Linguistic Annotation Workshop*, pp.158–161.

P. Pajas, J. Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),* Manchester, pp. 673–680.

M. Poesio, R. Vieira, S. Teufel. 1997. Resolving bridging references in unrestricted text. In: *ACL Workshop on Robust Anaphora Resolution*, pp. 1–6.

M. Poesio. 2004. The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In: *Proceedings of The 5th SIGdial Workshop on Discourse and Dialogue*, Boston.

M. Poesio, R. Delmonte, A. Bristot et al. 2004a. *The Venex corpus of anaphora and deixis in spoken and written Italian*. Manuscript.

M. Poesio, R. Mehta, A. Maroudas et al. 2004b. Learning to resolve bridging references. In: *42nd Meeting of the Association for Computational Linguistics (ACL 2004)*, pp. 143–150.

M. Poesio, R. Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC 2008,* Marrakech.

L. Poláková, P. Jínová, Š. Zikánová et al. 2012a. *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.

L. Poláková, P. Jínová, Š. Zikánová et al. 2012b. *Prague Discourse Treebank 1.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://ufal.mff.cuni.cz/discourse/, Nov 2012.

S. Pradhan, E. Hovy, M. Marcus et al. 2007. Ontonotes: A unified relational semantic representation. In: *Proceedings of the International Conference on Semantic Computing*, Washington DC.

S. Pradhan, A. Moschitti, N. Xue et al. 2012. *CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes.* Jeju, South Korea, Jul 2012.

R. Prasad, N. Dinesh, A. Lee et al. 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.

R. Prasad, S. McRoy, Nadya Frid et al. 2011. *The Biomedical Discourse Relation Bank, BMC 1*, 12:188

M. Recasens, A. M. Martí. 2009. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. In: *Language Resources and Evaluation*.

K. Rodríguez, F. Delogu, Y. Versley et al. 2010. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In *Proceedings of LREC 2010*, Valletta, Malta.

M. Stede. 2004. The Potsdam Commentary Corpus. *Proc. of the ACL 2004 Workshop on Discourse Annotation*, pp. 96–102.

S. Tonelli, G. Riccardi, R. Prasad et al. 2010. Annotation of Discourse Relations for Conversational Spoken Dialogs. 2010. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010),* pp. 2084–2090. Valletta, Malta.

B. Webber, A. Knott, M. Stone et al. 2003. Anaphora and Discourse Structure. *Computational Linguistics 29(4)*, pp. 545–588.

F. Wolf, E. Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2).

D. Zeyrek, I. Demirşahin, A. Sevdik-Çalli et al. 2010. The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations. In: *Proceedings of the Fourth Linguistic Annotation Workshop.* Pages 282–289. Uppsala, Sweden.

Y. Zhou, N. Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.* pp. 69–77. Jeju, Republic of Korea. July 2012.