Prague Dependency Treebank

The PDT contains 49,431 sentences of Czech journalistic texts from two daily newspapers, one business weekly and one scientific journal (see also Chapter 3 in the PDT guide (http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch03.html)). The length of the documents varies from a single sentence to 231 sentences, with an average length of 15.6 sentences. In the section Data (http://ufal.mff.cuni.cz/discourse/data.php), we present two example files with discourse annotation (see below), containing 40 and 105 sentences, respectively. The full version of the Prague Discourse Treebank 1.0 can be downloaded from the LINDAT-Clarin repository (https://ufal-point.mff.cuni.cz/xmlui/handle/11858/00-097C-0000-0008-E130-A) (see the Licence (http://ufal.mff.cuni.cz/discourse/licence.php)).

- Discourse relations (http://ufal.mff.cuni.cz/discourse/documentation.php#discourse)
- Coreference and bridging relations (http://ufal.mff.cuni.cz/discourse/documentation.php#coreference)

Discourse relations

Methodology and annotation procedure

Discourse connectives and their arguments

The primary goal of the discourse annotation in the PDT was to capture all discourse relations (both inter-sentential and intra-sentential ones) signaled by discourse connectives. The connectives play the most important role in identifying and describing these relations since they are the most apparent pointers to the structure of a discourse on the surface, both for humans and machines. This principle, namely the identification of discourse connectives and the text spans (or arguments) they connect, is also the underlying idea of the approach to discourse annotation in the Penn Discourse Treebank (Prasad et al. 2008) (http://www.seas.upenn.edu/~pdtb/), one of the most influential projects in this field.

Following this lexically-grounded approach, we have developed a unified methodology for description of discourse relations in Czech. A discourse connective (DC), both in Penn and Prague approaches, is defined as a predicate of a binary relation; it takes two text spans (mainly clauses or sentences) as its arguments. It connects these units to larger ones while signaling a semantic relation between them at the same time. Secondly, DCs are morphologically inflexible and they never act as grammatical constituents of a sentence. Like modality markers, they are “above” or “outside” of the proposition. They are represented by coordinating conjunctions (e.g. and, but), some subordinating conjunctions (e.g. because, if, while), some particles (e.g. also,only) and sentence adverbials (e.g. afterwards), and marginally also by some other parts-of-speech – mainly in case of fixed compound connectives like in other words or on the contrary.

Annotation process

The annotation in PDT only focused on discourse relations:

1. indicated by overly present (explicit) discourse connectives – the relations not indicated by a discourse connective were not annotated in this stage of the project,
2. discourse relations between clausal arguments, i.e. not between nominalizations or deictic expressions.

Additionally, the discourse annotation includes also marking of list structures (as a separate type of discourse structure) and marking of some other text phenomena like article headings, alternatively lexicalized discourse connectives, figure captions, non-coherent texts like collections of news etc.

Although the annotators had at their disposal both plain text and the tree structures (the underlying, so called tectogrammatical analysis), the annotation itself was carried out on syntactic (tectogrammatical) trees, since we did not want to lose connection with the analyses of previous levels.

The annotation consisted of two major steps:

1. All inter-sentential relations (relations between sentences) and a small part of intra-sentential relations (relations in one sentence) were annotated completely manually. Intra-sentential relations were only annotated manually in cases when their discourse semantics differed from the tectogrammatical interpretation (as it is the case for pragmatic interpretations, finer subcategorization of adversatives etc.).
2. The remaining intra-sentential relations (the interpretation of which on the deep-syntactic (tectogrammatical) layer was adequate for discourse level analysis) were automatically extracted and mapped onto the discourse annotation.

The manual part of annotation proceeded in three steps:

1. a connective was identified,
2. its two arguments (i.e. their extent) set, and
3. to each relation represented by a connective a discourse-semantic label was assigned (see Figure 1).

The automatic part of annotation was based on extracting relevant information (presence of the relation, scope of the arguments, the connective(s), a discourse-semantic label) from the deep-syntactic layer of PDT.

Both parts of the annotation (the manual and the automatic subparts) underwent consistent checking procedures.

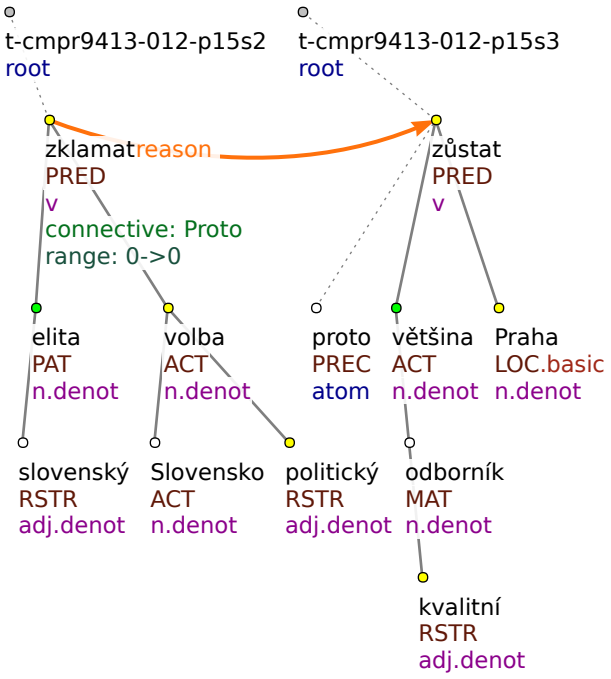

Figure 1: Example of discourse annotation: Slovenská elita byla zklamána politickou volbou Slovenska. Proto většina kvalitních odborníků zůstala v Praze. (The Slovak elite were disappointed by the political choice of Slovakia. That's why most of the quality specialists stayed in Prague.)

Basic features of the annotation on trees are illustrated by Figure 1. A discourse relation between subtrees is marked with a thick orange arrow, the type of the relation is displayed next to the tectogrammatical (deep-syntactic) lemma of the starting node (`reason` in Figure 1). A connective assigned to the relation shows in green (`Proto` in Figure 1). The attribute `range` determines the extent of the respective arguments (`start range -> target range`); the value of this attribute (in most cases a number) defines the number of trees to the right from the tree, where the arrow starts/ends (`0 -> 0` in Figure 1 means that the arguments are only the subtrees of the starting/target node). For arguments not fulfilling a form of a (sub)tree, the attribute `group` is used. Arguments of arrows within one tree are understood not to include one another (although one might be in the subtree of the other).

Benefiting from the syntactic annotation in PDT

In its present shape, the multilayer annotation of the Prague Dependency Treebank 2.5 already marks some of the phenomena relevant for discourse analysis and modeling. As was already mentioned, in the discourse annotation project we take advantage of these. The discourse layer of annotation adopts a part of the underlying syntactic annotation – namely some of the dependency relations, the coordinating relations between clauses (not those between lower units) and expressions marked with the semantic label PREC (reference to PREceding Context – assigned mainly to sentence-initial connectives like and, next, further, however etc., see Mladová et al. 2008).

Semantics of discourse relations

For the task of determining the semantic type of a discourse relation, a set of discourse semantic labels was developed as a result of comparison of the sense hierarchy used in Penn (Miltsakaki et al. 2008) and the set of Prague tectogrammatical labels called functors (Mikulová et al. 2005). For further information see the section Publications (http://ufal.mff.cuni.cz/discourse/publications.php).

Inter-annotator agreement

We had several annotators but each part of the data has been annotated by one annotator only, with the exception of a small overlap (4% of the data) for studying and measuring the inter-annotator agreement. To evaluate the agreement, we have used the connective-based F1-measure (Mírovský et al., 2010), a simple ratio, and Cohen's κ (Cohen, 1960). The connective based F1-measure has been used for measuring the agreement on the recognition of discourse relations, a simple ratio and Cohen's κ have been used for measuring the agreement on the type of the relations in cases where the annotators recognized the same relation.

In the connective-based measure, we consider the annotators to be in agreement on recognizing a discourse relation if two connectives they mark (each of them marked by one of the annotators) have a non-empty intersection (technically, a connective is a set of tree nodes). For example, if one of the annotators marks two words a proto [and therefore] as a connective, and the other annotator only marks the (same) word proto [therefore], we take it as agreement – they both recognized the presence of a discourse relation. (They still may disagree on the type of the relation.)

Table 1: The inter-annotator agreement on all parallel discourse data

measurement	F1	agreement on types	kappa on types
all parallel data	0.83	0.77	0.71

Table 1 shows the inter-annotator agreement measured at once on all the parallel data. Altogether, there have been 44 (parallelly annotated) documents, 2,084 sentences and 33,987 words. Only the inter-sentential discourse relations (not relations representing lists) have been taken into account.

The simple ratio agreement on types from Table 1 (0.77) is the closest measure to the way of measuring the inter-annotator agreement on subsenses in the annotation of discourse relations in the Penn Discourse Treebank 2.0, reported in Prasad et al. (2008). Their agreement was 0.8.

For more information about the measurements of the inter-annotator agreement, see publications about the inter-annotator agreement in the section Publications (http://ufal.mff.cuni.cz/discourse/publications.php).

Coreference and bridging relations

The primary goal of the coreference and bridging annotation was to capture all coreference relations (including zero anaphora) and some types of bridging relations in the PDT.

Two or more expressions are considered to be coreferential if they refer to the same extralinguistic entity, equivalence of the head nouns not being necessary a precondition to call the expressions coreferential.

According to syntactic properties of anaphoric elements, two types of coreference relations are distinguished:

- grammatical coreference (a kind of coreference in which it is possible to identify the antecedent on the basis of grammatical rules, e.g. coreference with reflexive pronouns, relative elements, controls, etc.)
- textual coreference (pronoun, zero and nominal coreference)

Non-coreferential association relations are annotated as bridging relations if they stand in one of specific types (described below) of semantic, lexical or conceptional relations to their antecedents.

The annotation of coreference and bridging relations is based on the tectogrammatical level. The annotation has been carried out on tectogrammatical trees and many elements of tectogrammatical level, such as functors, node types, grammatemes etc. have been made use of.

The annotation of coreference and bridging relations over PDT 2.0 data was published in 2011 (Nedoluzhko et al. 2011b). The present release presents an adaptation of the annotation to the PDT 2.5 data and contains a number of fixes.

Annotation procedure

The annotation proceeded in two stages:

1. Grammatical coreference and pronoun textual coreference (including the annotation of zero anaphora). This annotation was completed for PDT 2.0 (Kučová et al. 2003).
2. Extended textual coreference and bridging relations, new for PDT 2.5 in 2012

The annotation of extended textual coreference and bridging anaphora consists of the following actions:

- automatic pre-annotation (e.g. linking some named entities),
- automatic useful tools which help annotators to find the correct antecedents (highlighting already linked items in the trees, underlying the same lemmas, etc.),
- manual annotation,
- automatic check of some aspects of coreference links (finding the nearest antecedent, preserving coreferential chains, bridging long coreferential chains)

The relations of the following types are annotated:

- grammatical coreference
- textual coreference for NPs with specific (type SPEC) and generic reference (type GEN)
- bridging relations of types PART–WHOLE, SET–SUBSET, FUNCTION–OBJECT, CONTRAST, ANAPHora without coreference and REST for some other specific cases.

Figure 2 shows basic features of the coreference and bridging annotation. Coreference/bridging relation between subtrees are marked by arrows of different colors (dark-red arrow for grammatical coreference, dark-blue arrows for textual coreference and light-blue arrows for bridging reference), the arrow pointing from an anaphor to an antecedent. If an antecedent is in one of the preceding sentences, its lemma is written in dark-blue near its anaphor.

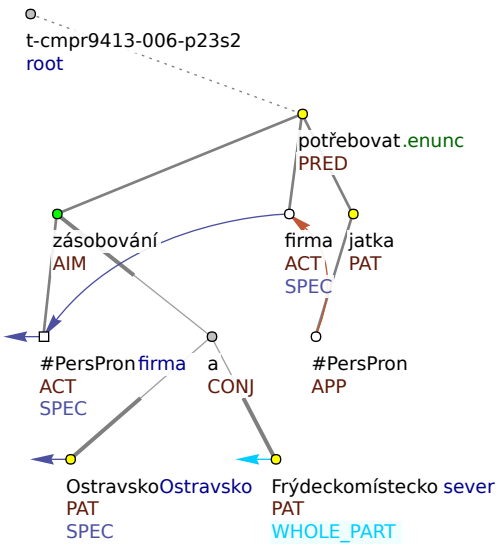

Figure 2: Example of coreference annotation: Pro zásobování Ostravska a Frýdeckomístecka potřebuje firma svá jatka. (The company needs its slaughterhouse in order to supply the Ostrava and Frydek-Mistek regions.)

Annotation principles

While annotating coreference and bridging relations, a number of basic principles and preferences were followed, e.g.:

- Chain principle (coreference relations in text are organized in ordered chains; the most recent mention of an entity is marked as the antecedent),

- Principle of the maximum length of coreferential chains (in case of a multiple choice, we prefer to continue the existing coreference chain, rather than to begin a new one),
- Principle of maximal size of an anaphoric expression (it is always the whole subtree of the antecedent/anaphor, which is the subject to the annotation),

etc.

Inter-annotator agreement

Similarly to the measurement of the inter-annotator agreement in the annotation of discourse relations, we have measured the inter-annotator agreement in the annotation of coreference and bridging anaphora on a small part of the data that had been annotated in parallel by two annotators. To evaluate the agreement, we have used the chain-based F1-measure, a simple ratio, and Cohen's κ (Cohen, 1960). The chain-based F1-measure has been used for measuring the agreement on the recognition of a coreference or bridging relation, a simple ratio and Cohen's κ have been used for measuring the agreement on the type of the relations in cases where the annotators recognized the same relation.

In the chain-based measure, we consider the annotators to be in agreement on recognizing a coreference or bridging relation if the 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: The inter-annotator agreement on all parallel coreference and bridging data

measurement	F1	agreement on types	kappa on types
all parallel data - coreference	0.72	0.90	0.73
all parallel data - bridging anaphora	0.46	0.92	0.89

Table 2 shows the inter-annotator agreement measured at once on all the parallel data. Altogether, there have been 39 (parallelly annotated) documents, 1,606 sentences and 26,620 words.

Malostranské náměstí 25
118 00 Praha
Czech Republic
+420 951 554 278 (phone)
+420 257 223 293 (fax)
ufal@ufal.mff.cuni.cz (mailto:ufal@ufal.mff.cuni.cz)