Discourse Relations in the Prague Dependency Treebank 3.0

Jiří Mírovský, Pavlína Jínová, Lucie Poláková
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
mirovsky|jinova|polakova@ufal.mff.cuni.cz

Abstract

The aim of the demo is threefold. First, it introduces the current version of the annotation tool for discourse relations in the Prague Dependency Treebank 3.0. Second, it presents the discourse relations in the treebank themselves, including new additions in comparison with the previous release. And third, it shows how to search in the treebank, with focus on the discourse relations.

1 Introduction

The Prague Dependency Treebank 3.0 (Bejček et al., 2013) is the newest version of the Prague Dependency Treebank series, succeeding versions 1.0 (PDT 1.0; Hajič et al., 2001), 2.0 (PDT 2.0; Hajič et al., 2006), 2.5 (PDT 2.5; Bejček et al., 2012) and Prague Discourse Treebank 1.0 (PDiT 1.0; Poláková et al., 2012, 2013). It is a corpus of Czech, consisting of almost 50 thousand sentences annotated mostly manually on three layers of language description: morphological, analytical (surface syntactic structure), and tectogrammatical (deep syntactic structure). On top of the tectogrammatical layer, explicitly marked discourse relations, both inter- and intra-sentential ones, have been annotated. The discourse annotation first appeared in PDiT 1.0, and it was corrected and updated for the newest release of the Prague Dependency Treebank, PDT 3.0.

In Section 2, we present the annotation tool for discourse relations in PDT 3.0. In Section 3, we briefly introduce principles of discourse annotation in PDT 3.0. Section 4 is dedicated to searching in PDT 3.0, focusing on searching for discourse relations.

2 The Annotation Tool

The primary format of PDT since version 2.0 is called PML (http://ufal.mff.cuni.cz/jazz/PML/). It is an abstract XML based format designed for annotation of linguistic corpora, especially treebanks. Data in the PML format can be browsed and edited in TrEd, a fully customizable tree editor (Pajas and Štěpánek, 2008). TrEd is written in Perl and can be easily customized to a desired purpose by extensions that are included into the system as modules. The TrEd extension for discourse annotation in PDT was first described in Mírovský et al. (2010). Here we summarize the main features of the tool, including additions that have been made since the previous version. Also the data format of discourse relations in PDT 3.0 has undergone several changes.

The data format and the tool for annotation of discourse in PDT allow for:

- **Creation of a link** between arguments of a relation; the link is depicted with a thick orange arrow between nodes representing the arguments (see Figure 2 below).
- **Exact specification of the extent** of the arguments of the relation; it takes advantage of the tree structure of the tectogrammatical layer and specifies the range of an argument as a set of (sub)trees; in unclear cases the argument can be defined as an under-specified sequence of trees starting (or newly also ending) at a given point in the data. In rare cases, an arbitrary set of individual nodes can be specified as an argument as well.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 34–38, Dublin, Ireland, August 23-29 2014.

• **Assigning a connective** to the relation; the connective can be defined as a list of tectogrammatical nodes and, if needed, also by nodes from the lower (analytical) layer. Newly also extended connectives can be assigned to the relation, which is an addition required by the on-going annotation of so called AltLexes (alternative lexicalizations of connectives).
• **Setting additional information** to the relation (a type, a source, a comment etc.); newly also a flag for the AltLex can be indicated and also a flag for negation of a discourse type.
• **Assigning other discourse related information to nodes** at the tectogrammatical layer; article headings, table or figure captions and metatext can be indicated at the root node of the respective phrase.

3 Discourse Relations in PDT 3.0

Annotation of discourse relations in PDT 3.0 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 linguistics (see e.g. Sgall et al., 1986). While in PDTB approach, a list of possible discourse connectives was created and according to it, contexts for annotators were prepared, we only defined a connective theoretically and left annotators to go through the whole text and identify all such constructions with a connective function. In the first and second release of discourse annotation (PDiT 1.0 and PDT 3.0), only discourse relations indicated by overly present (explicit) discourse connectives, i.e. expressions like *but, however, as a result, even though* etc. have been annotated. Every discourse connective is thought of as a discourse-level predicate that takes two discourse units as its arguments. Only discourse relations connecting clausal arguments (with a predicate verb) have been annotated. The Prague discourse annotation also includes marking of list structures (as a separate type of discourse structure) and marking of smaller text phenomena like article headings, figure captions, metatext etc.

The annotation proceeded first manually for cases where the tectogrammatical layer did not allow for identifying a discourse relation automatically. Afterwards, using the information (mostly) from the tectogrammatical layer, we were able to identify and mark almost 10 thousand out of more than 12 thousand intra-sentential relations automatically – arguments (verbal phrases), types of relations and the connectives were identified using tree structures of the sentences, tectogrammatical functors (types of dependency or coordination), and morphological tags (details in Jínová et al., 2012).

The Prague discourse label set was inspired by the Penn sense tag hierarchy (Prasad et al., 2008) and by the tectogrammatical functors (Mikulová et al., 2005). 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 (see Poláková et al., 2013). The third level is captured by the direction of the discourse arrow. Within the four classes, the types of relations partly differ from the Penn senses and go closer to Prague tectogrammatical functors and/or are a matter of language-specific distinctions. The annotators, unlike in the Penn approach, were not allowed to only assign the major class, they always had to decide for a specific relation within one of the classes.

PDT 3.0 brings an update of the discourse annotation released in PDiT 1.0. It has been enriched with several newly annotated (or not yet released) discourse-related phenomena, namely genre specification of the corpus texts (see Poláková et al., 2014), annotation of some type of rhematizers (or foculizing particles) as discourse connectives, and annotation of second relations (discourse relations with more than one semantic type). Also a new attribute discourse_special for several special roles of phrases has been introduced.

We newly annotated foculizing particles in structures with conjunction, to see how these particles cooperate with other types of connectives in discourse. Second relations were annotated in the data already before the PDiT 1.0 release but only in the annotator's comment, which did not become a part of the official release. In PDT 3.0, each second relation has been captured as an additional full-fledged relation with its own type and connective. It means that the arguments in question are connected with two arrows representing two discourse relations.

The newly introduced attribute discourse_special captures three special roles of the phrase represented by a node and its subtree; the possible values are: “heading” (article headings; replaces attribute is_heading from PDiT 1.0), “metatext” (text not belonging to the original newspaper text, produced during the creation of the corpus), and “caption” (for captions of pictures, graphs etc.).[1]

[1] Metatext and caption were also annotated already before the PDiT 1.0 release in the annotator's comment (but not published there).

4 Searching in PDT 3.0

For searching in PDT, a client-server based system called PML-TQ has been developed (PML-Tree Query; Pajas and Štěpánek, 2009). It belongs to the most powerful systems for searching in treebanks. The server part is implemented either as a relational database or as a system in a command-line version of TrEd (btred). The client part uses either the tree editor TrEd along with a PML-TQ extension or a web browser. The web browser client has however some limitations, so (in the demo) we focus on TrEd with the PML-TQ extension.

Queries in PML-TQ can be created both in a textual form and (in the TrEd client) in a graphical environment. The query language allows to define properties of tree nodes and relations among them, inside or between sentences and also across the layers of annotation. Negation on the tree structure and Boolean expressions over the relations can be used. Results of the corpus search can be viewed along with the context or processed with output filters to produce statistical tables. A detailed documentation can be found at http://ufal.mff.cuni.cz/pmltq/doc/pmltq_doc.html, in the demo we will offer an introduction to the principal parts of the language along with a set of illustrative examples, from the basic queries to more complex ones, respecting requests from the audience.

The following example shows how to search for a discourse relation. The query defines two tectogrammatical nodes (t-nodes) connected with a special "member" node that represents a discourse relation between the two nodes. The required type of the discourse relation can be specified at the member node, in this example it is set to "reason". The query also specifies that the start and target nodes of the relation are not from the same tree, i.e. it looks for an inter-sentential discourse relation of the semantic type "reason".

Textual form of the query:

```
t-node
[ !same-tree-as $t,
     member discourse
     [ discourse_type = "reason",
 target_node.rf t-node $t := [  ] ] ];
```

Graphical form of the query:

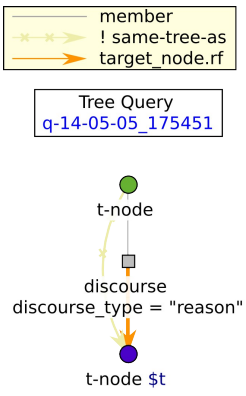

Figure 1: Graphical form of the query

The following two sentences represent one of the results of the query:

```
Pronikání do cizích počítačových systémů je podle našich zákonů beztrestné.
Policie tak jen bezmocně přihlíží, když v bankách řádí SLÍDILOVÉ.

[Infiltration of other computer systems is according to our laws not a criminal act.
Thus the police only helplessly watches, as SNOOPERS rage in banks.]
```

Figure 2 captures the tectogrammatical annotation of these two sentences, along with the discourse relation represented by the thick orange arrow connecting roots of the two respective propositions.
Results of queries in PML-TQ can be further processed using output filters. Thanks to an output filter, a result of a query does not consist of individual matching positions in the trees but of a tabular summary of all the matching positions, specified by the output filter.

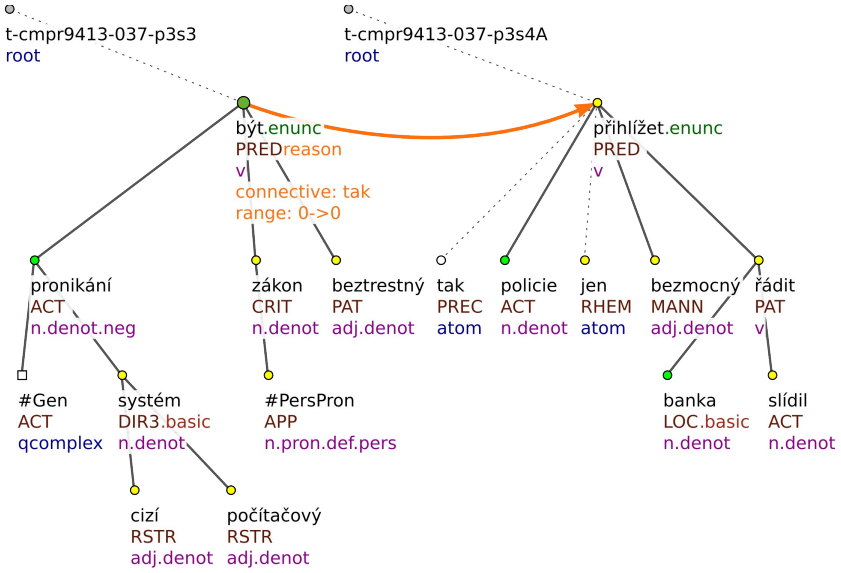

Figure 2: The tectogrammatical representation of the two result sentences of the query

If we modify the previous query by deleting the definition of the discourse type (`discourse_type = "reason"`), naming the member node (`$d :=`) and adding an output filter (the last line with prefix >>):

```
t-node
[ !same-tree-as $t,
     member discourse $d :=
         [ target_node.rf t-node $t := [  ] ] ];
>> for $d.discourse_type give $1, count() sort by $2 desc
```

...the query will search for all inter-sentential discourse relations in the data and – thanks to the output filter – produce the following distribution table of the discourse types, sorted in the descending order by the number of occurrences (only a few selected lines are printed here to save space):

opp	1,800
conj	1,389
reason	1,031
...	
grad	204
restr	172
explicat	130
...	

Table 1: (Selected) results of the output filter

5 Conclusion

A good annotation tool, well designed annotation guidelines, and a powerful search tool are necessary parts of a well managed project of any linguistic annotation. In the demo, we present all these parts for the current version of annotation of discourse relations in the Prague Dependency Treebank 3.0.

The PML as a data format, the annotation tool TrEd and the search system PML-TQ can be and have been extensively used for many other annotation tasks in PDT and also for many other treebanks, see for example a project of harmonizing various treebanks in HamleDT (Zeman et al., 2012).

Acknowledgment

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (projects P406/12/0658 and P406/2010/0875). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

Bejček, Eduard, Hajičová, Eva, Hajič, Jan et al. (2013). *Prague Dependency Treebank 3.0.* Data/software, Charles University in Prague, MFF, ÚFAL. Available at: http://ufal.mff.cuni.cz/pdt3.0/.

Bejček, Eduard, Panevová, Jarmila, Popelka, Jan et al. (2012). *Prague Dependency Treebank 2.5.* Data/software, Charles University in Prague, MFF, ÚFAL. Available at: http://ufal.mff.cuni.cz/pdt2.5/.

Hajič, Jan, Vidová Hladká, Barbora, Panevová, Jarmila et al. (2001). *Prague Dependency Treebank 1.0* (Final Production Label). In: CDROM, CAT: LDC2001T10., ISBN 1-58563-212-0.

Hajič, Jan, Panevová, Jarmila, Hajičová, Eva et al. (2006). *Prague Dependency Treebank 2.0*. Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4.

Jínová, Pavlína, Mírovský, Jiří, & Poláková, Lucie (2012). Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT. In: *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*, Mumbai, India, pp. 43-58.

Mikulová, Marie 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.

Mírovský, Jiří, Mladová, Lucie, & Žabokrtský, Zdeněk (2010). 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.

Pajas, Petr, & Štěpánek, Jan (2009). System for Querying Syntactically Annotated Corpora. In: *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, Association for Computational Linguistics, Suntec, Singapore, ISBN 1-932432-61-2, pp. 33-36.

Pajas, Petr, & Štěpánek, Jan (2008). Recent Advances in a Feature-Rich Framework for Treebank Annotation. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, ISBN 978-1-905593-45-3, pp. 673-680.

Poláková, Lucie, Jínová, Pavlína, & Mírovský, Jiří (2014). Genres in the Prague Discourse Treebank. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, ISBN 978-2-9517408-8-4, pp. 1320-1326.

Poláková, Lucie, Mírovský, Jiří, Nedoluzhko, Anna et al. (2013). Introducing the Prague Discourse Treebank 1.0. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing,* Asian Federation of Natural Language Processing, Nagoya, Japan, ISBN 978-4-9907348-0-0, pp. 91-99.

Poláková, Lucie, Jínová, Pavlína, Zikánová, Šárka et al. (2012). *Prague Discourse Treebank 1.0.* Data/software, Charles University in Prague, MFF, ÚFAL. Available at: http://ufal.mff.cuni.cz/pdit/.

Prasad, Rashmi, Dinesh. Nikhil, Lee, Alan 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.

Sgall, Petr, Hajičová, Eva, & Panevová, Jarmila (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel Publishing Company and Prague: Academia.

Zeman, Daniel, Mareček, David, Popel, Martin et al. (2012). HamleDT: To Parse or Not to Parse? In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012*, European Language Resources Association, İstanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 2735-2741.