Discourse Relations in Prague Dependency Treebank 3.0

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

Prague Dependency Treebank series

PDT 1.0 (2001, LDC) – 50 thousand sentences in Czech annotated on morphological and analytical layer, dependency syntax
PDT 2.0 (2006, LDC) – 50 thousand sentences in Czech annotated on three layers, newly incl. tectogrammatical (deep syntax) layer
PDT 2.5 (2011, Lindat/Clarin) – multiword expressions, clause segmentation, pair/group meaning
PDiT 1.0 – Prague **Discourse** Treebank 1.0 (2012, Lindat/Clarin) – **inter- and intra-sentential discourse relations** marked by explicit connectives
PDT 3.0 (2013, Lindat/Clarin) – **extension of discourse relations**, **genres of documents,** coreference of 1st and 2nd person, grammatemes revised

Extension of discourse annotation in PDT 3.0

- **second relations** (fully annotated as two independent relations)
- annotation of **genres of documents** (20 genres for 3,165 documents)
- attribute **discourse_special** for *article headings*, *metatext*, and *caption*
- annotation of **focalizing particles** in structures with conjunction

Distribution of discourse types (senses) in PDT 3.0

	total	intra-sent.	inter-sent.	intra-sent.	inter-sent.
conjunction	7,498	6,109	1,389	81%	19%
opposition	3,196	1,396	1,800	44%	56%
reason-result	2,632	1,601	1,031	61%	39%
condition	1,369	1,350	19	99%	1%
concession	880	617	263	70%	30%
precedence	840	591	249	70%	30%
confrontation	653	345	308	53%	47%
specification	630	519	111	82%	18%
correction	445	322	123	72%	28%
gradation	445	241	204	54%	46%
purpose	414	413	1	100%	0%
disj. alternative	272	257	15	94%	6%
restriction	269	97	172	36%	64%
explication	230	100	130	43%	57%
synchrony	226	174	52	77%	23%
exemplification	148	28	120	19%	81%
generalization	106	9	97	8%	92%
equivalence	105	41	64	39%	61%
conj. alternative	90	69	21	77%	23%
pragm. opposition	50	23	27	46%	54%
pragm. reason-result	40	12	28	30%	70%
pragm. condition	16	15	1	94%	6%
other	2	1	1	50%	50%

Second relations

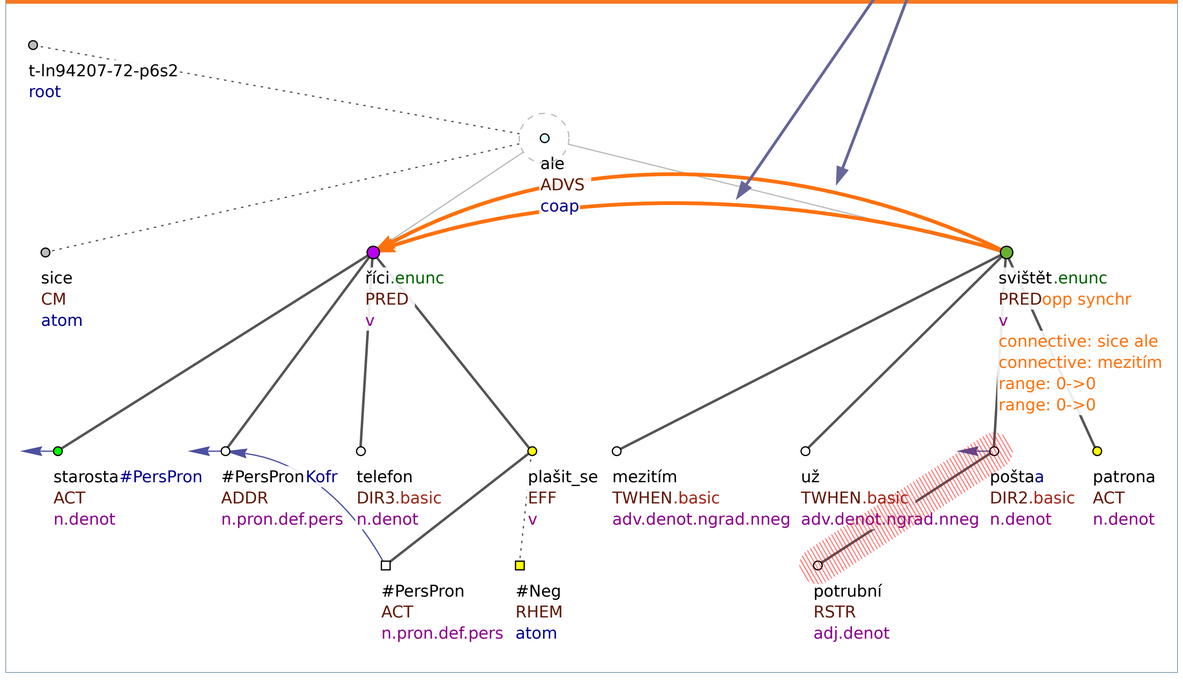
A sentence in Czech (original from PDT) and English (translation)

*Starosta mu **sice** do telefonu řekl, ať se neplaší, **ale mezitím** už potrubní poštou svištěly patrony.*

*[The mayor **may have** told him on the phone not to freak out **but in the meantime** bullets already whistled through the tubular post.]*

two arrows represent two discourse relations (opposition and synchrony) expressed by two connectives:
sice – ale [may have – but] and mezitím [in the meantime]

Graphical representation in tree editor TrEd

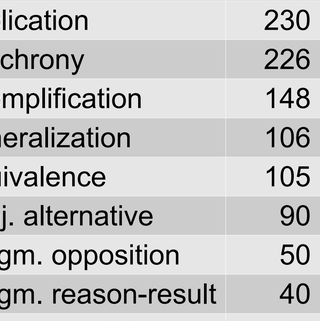


PML Tree Query (PML-TQ) – A Search Tool

A Query

Tree Query
q-14-05-05_175451

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



a textual form of the query

a graphically created query searching for an inter-sentential discourse relation of discourse type *reason*

Focalizing particle *také [also]* in structures with conjunction

As a part of a connective:

*Taková odměna může mít skutečně silný motivační účinek pro účastníky **a** může být **také** užitečným přínosem pro firmu, která náklady plně hradí.*

*[Such a reward can really have a strong motivational effect for the attendees **and** can **also** be a useful contribution for the company that fully pays the costs.]*

Not a connective:

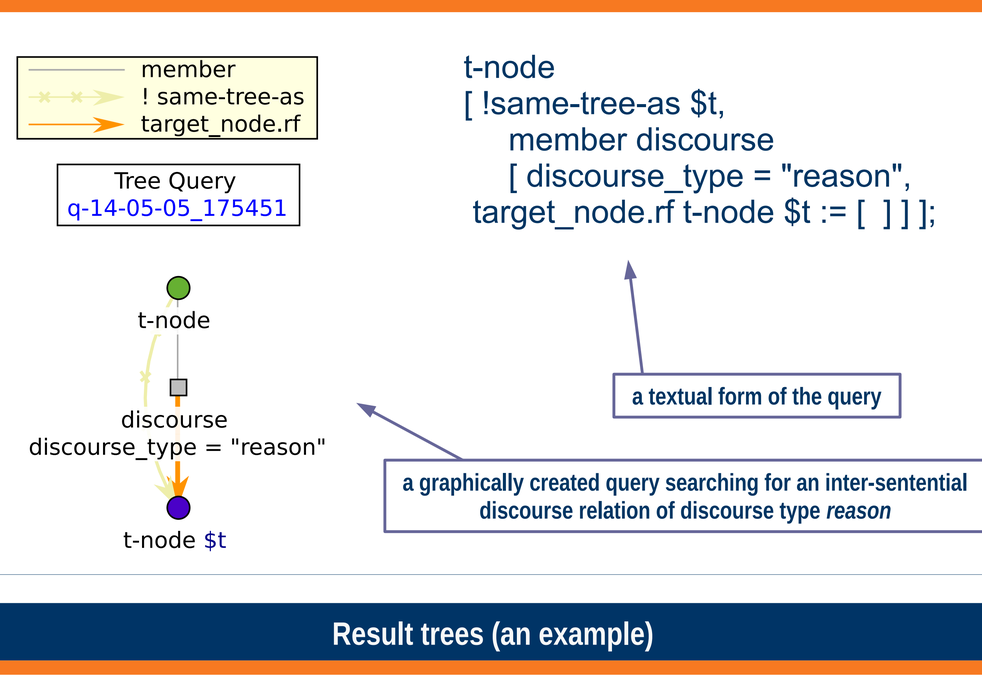
*"Japonci chtějí křeslo v Radě bezpečnosti, to není žádné tajemství, **a** my chceme **také** jedno," prohlásil Kinkel před novináři.*

*[Japanese want a seat on the Security Council, it is no secret, **and** we want to have **also** one," Kinkel said to journalists.]*

Result trees (an example)

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

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



Result of the query with an output filter

```
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
```

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

a similar query, this time with an output filter generating a distribution table of discourse types in the data

(selected) results of the query with the output filter

Final remarks

The Prague Dependency Treebank 3.0 was published in December 2013 under the Creative Commons License and it is available to download from the LINDAT-Clarin repository.

Tree editor TrEd is available under the GNU GPL licence for Linux, MacOS, and MS Windows up to ver. 7.

PML-TQ is an extension of TrEd and has been used for many different treebanks.

The presented work was supported by the Grant Agency of the Czech Republic (projects n. P406/12/0658 and P406/2010/0875), and by the Ministry of Education, Youth and Sports of the Czech Republic (project n. LM2010013 – LINDAT/CLARIN).