Use of Coreference in Automatic Searching for Multiword Discourse Markers in PDT

Magdaléna Rysová
Jiří Mírovský

Charles University in Prague
Institute of Formal and Applied Linguistics
Overview

- **Prague Dependency Treebank**
  - PDT 1.0, PDT 2.0, PDT 2.5, PDiT 1.0, PDT 3.0

- **Discourse in Prague Dependency Treebank**
  - PDiT 1.0, PDT 3.0

- **AltLexes in Prague Dependency Treebank**
  - for the next version of PDT
Prague Dependency Treebank

- 3,165 documents, **49,431 sentences**, 833,195 tokens
- (mostly) **manually annotated** Czech journalistic texts from 1990's
  - morphological layer
  - analytical (surface syntax) layer
  - tectogrammatical (deep syntax) layer
  - discourse phenomena
The Tectogrammatical Layer

- Tectogrammatical layer
  - sentences represented as *dependency trees*
  - semantic labels called *functors* (approx. 44 labels) represent type of dependency (PRED, ACT, ADDR, ...)
  - some of them further specified by *subfunctors* (functor LOC “where” is further specified by subfunctors *in, behind, under, along* etc.)
The Tectogrammatical Layer

- Tectogrammatical layer (cont.)
  - coreference
    - grammatical (given by grammatical rules)
    - pronominal textual
    - nominal textual (since PDiT 1.0)
    - bridging anaphora (since PDiT 1.0)
  - topic-focus articulation – (contrastively) contextually bound and contextually non-bound elements, comunicative dynamism expressed in node order
Můžete to vysvětlit na příkladu?
[Could you explain it on an example?]
Prague Dependency Treebank

Updates

- PDT 1.0 – published in 2001 (LDC)
- PDT 2.0 – published in 2006 (LDC)
- PDT 2.5 – published in 2011 (downloadable from Lindat/Clarin repository, Creative Commons License)
- PDiT 1.0 – published in 2012 (downloadable from Lindat/Clarin repository, Creative Commons License)
- PDT 3.0 – published in December 31\textsuperscript{st}, 2013 (…)}
Annotation of Discourse in Prague Dependency Treebank

Stage 1 → PDiT 1.0 (2012)

- discourse relations with explicit connectives between verbal arguments, 23 discourse types (senses)
- formal definition of connectives (not a list), annotators examined the whole text
- inter-sentential manually, intra-sentential semi-auto
- AltLexes only marked in an annotator's comment
- (nominal textual coreference, bridging anaphora)
AltLexes in Prague Dependency Treebank

Connectives vs. AltLexes
(Alternative Lexicalizations of Connectives)

proto (therefore) vs. z toho důvodu (for that reason)
proto (therefore) vs. za tím účelem (for that purpose)
protože (because) vs. důvodem je (the reason is)
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.
Annotation of Discourse in Prague Dependency Treebank

Stage 2  PDT 3.0 (2013)

- 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 captions
- annotation of focalizing particles in structures with conjunction
- (textual coreference of 1st and 2nd person)
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.]
AltLexes in Prague Dependency Treebank

Stage 3 → a future PD(i)T issue (2015?)

- annotation of AltLexes
- annotated partially manually, partially automatically based on the annotator's comment from Stage 1
AltLexes in Prague Dependency Treebank

49,431 sentences in PDT, over 20 thousand discourse relations with explicit connectives

but:

only 306 cases of annotator's comment “AltLex”
AltLexes in Prague Dependency Treebank

however:

AltLex *díky* (*thanks to*) appears in PDT in 14 instances although it was marked in the annotator's comment just in one case.
AltLexes in Prague Dependency Treebank

Three types

1) AltLexes with a **key word** occurring in several collocations

2) **Fixed collocations** functioning as AltLexes only in given combinations and forms

3) AltLexes obligatorily containing an **anaphoric reference** to previous context (on the surface layer)
AltLexes in Prague Dependency Treebank

1) AltLexes with a key word occurring in several collocations

- **důvod (reason):** důvodem je (the reason is), jako důvod uvádí (he gives us this reason)...

- **příklad (example):** příkladem je (the example is), uvádí několik příkladů (he states several examples), příkladem toho bylo (the example of this was)...
2) Fixed collocations functioning as AltLexes only in given combinations and forms

- *jak je vidět* (as seen)
- *rozumějme* (to understand)
- *krátce/jednoduše řečeno* (shortly/simply speaking)
- ...
3) AltLexes obligatorily containing an anaphoric reference to previous context (in the surface layer)

- prepositional phrases:
  - *díky tomu* (thanks to *this*)
  - *kvůli tomu* (because of *this*)
  - *kromě toho* (apart from *this*)
  - *navzdory tomu* (despite *this*)
  - ...
AltLexes in Prague Dependency Treebank

Only interested in anaphoric reference to a verbal argument

*Itálie šetří.*

*Kvůli tomu tam přestanou vycházet některé deníky.*

*[Italy saves.]*

*Because of this, some journals will no longer be published.*
AltLexes in
Prague Dependency Treebank

Using all available anaphoric links

- grammatical coreference
- textual coreference
- bridging anaphora
- coreference to segment
PML-TQ Query

1 t-node $t :=
2  [ (1+x coref_gram.rf t-node
3      [ gram/sempos = "v" ] or
4    1+x coref_text/target-node.rf t-node
5      [ gram/sempos = "v" ] or
6    1+x bridging/target-node.rf t-node
7      [ gram/sempos = "v" ] or
8    1+x coref_gram.rf t-node
9      [ nodetype = "coap", t-node
10      [ gram/sempos = "v" ] ] or
11    1+x coref_text/target-node.rf t-node
12      [ nodetype = "coap", t-node
13      [ gram/sempos = "v" ] ] or
14    1+x bridging/target-node.rf t-node
15      [ nodetype = "coap", t-node
16      [ gram/sempos = "v" ] ] or
17    coref_special = "segm"),
18   a/lex.rf|a/aux.rf a-node
19    [ m/form ~ "^[Vv]inou$" ] ];
20
21 >> give $t.id
AltLexes in Prague Dependency Treebank

Results

1,482 cases of prepositions selected on the basis of Stage 1 (annotator's comment)

89 instances automatically selected as AltLexes
(9 of them marked as AltLex in Stage 1 in annotator's comment)
AltLexes in Prague Dependency Treebank

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Instances as AltLexes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>díky (thanks to)</td>
<td>14</td>
<td>191</td>
</tr>
<tr>
<td>kromě (in addition to)</td>
<td>44</td>
<td>309</td>
</tr>
<tr>
<td>kvůli (due to)</td>
<td>5</td>
<td>130</td>
</tr>
<tr>
<td>na rozdíl od (unlike)</td>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td>na základě (on the basis of)</td>
<td>7</td>
<td>167</td>
</tr>
<tr>
<td>navzdory (despite)</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>přes (in spite of)</td>
<td>9</td>
<td>389</td>
</tr>
<tr>
<td>vinou (due to)</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>vzhledem k (considering)</td>
<td>6</td>
<td>157</td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
<td>1,482</td>
</tr>
</tbody>
</table>
AltLexes in Prague Dependency Treebank

Reliability

191 instances of **díky (thanks to)** checked manually
- **14** with reference to a verbal node (OK)
- **21** with reference to a non-verbal node (OK)
- **156** without a reference
  - 3 disputable cases where a coreference could be annotated
Discourse in PDT
Future Plans

- **finish** the annotation in *AltLexes* (so far >700)
- study a possibility to annotate **implicit relations**
- study a possibility to **automatically** (based on the current annotation of discourse, coreference and bridging anaphora):
  - mark places with no discourse relation (**NoRel**)  
  - mark places with entity based relation (**EntRel**)
Use of Coreference in Automatic Searching for Multiword Discourse Markers in PDT

Thank you for your attention!

Magdaléna Rysová
Jiří Mírovský
I) Discourse connectives vs. their alternative lexicalizations (= AltLex’s)

**Connectives** = expressions with connecting function at the level of discourse description

a) Coordinating conjunctions: *and (a), but (ale), therefore (proto)*;
b) subordinating conjunctions: *although (ačkoliv)*;
c) particle expressions (including rhematizers): *even (dokonce), too (také)*;
d) adverbs: *then (potom)*;
e) certain uses of pronouns: *except for this (kromě toho)*;
f) idiomatic multiple-word connective means formed by linking of different expressions: *on the one hand (na jedné straně)*;
g) elements formed by letters or numbers expressing enumeration: *a), b), 1., 2.*;
h) two punctuation marks: colon and dash.

Expressions with the same function but from other classes = **alternative lexicalizations of discourse connectives (= AltLex’s)**