PDT → UMR

Jan Štěpánek

PDT → UMR

Two Steps

1. PDT → Treex
2. Treex → UMR

Treex?

What is Treex?

- Perl framework originally designed for machine translation.
- Main difference to PML: All the layers (and even languages) in one file.

Why Treex?

Ask Dan and Mišo 😇

PDT → Treex

Treex can already read PDT files.

```
Read::PDT schema_dir=$UFAL_UMR/data/stepanek language=cs from=$file
Write::Treex
```

Schemata

- pdt-c/tred-extension/pdt_c_m/resources/mdata_36_schema.xml
- pdt-c/tred-extension/pdtc10/resources/adata_c_schema.xml
- pdt-c/tred-extension/pdtc10/resources/mdata_c_schema.xml
- pdt-c/tred-extension/pdtc10/resources/tdata_c2_schema.xml
- pdt-c/tred-extension/pdtc10/resources/wdata_c_schema.xml
- tred/extensions.git/pdt20/resources/adata_schema.xml
- tred/extensions.git/pdt20/resources/mdata_schema.xml
- tred/extensions.git/pdt20/resources/tdata_schema.xml
- tred/extensions.git/pdt20/resources/wdata_schema.xml
- tred/extensions.git/pdt30/resources/wdata_30_schema.xml

Treex → UMR

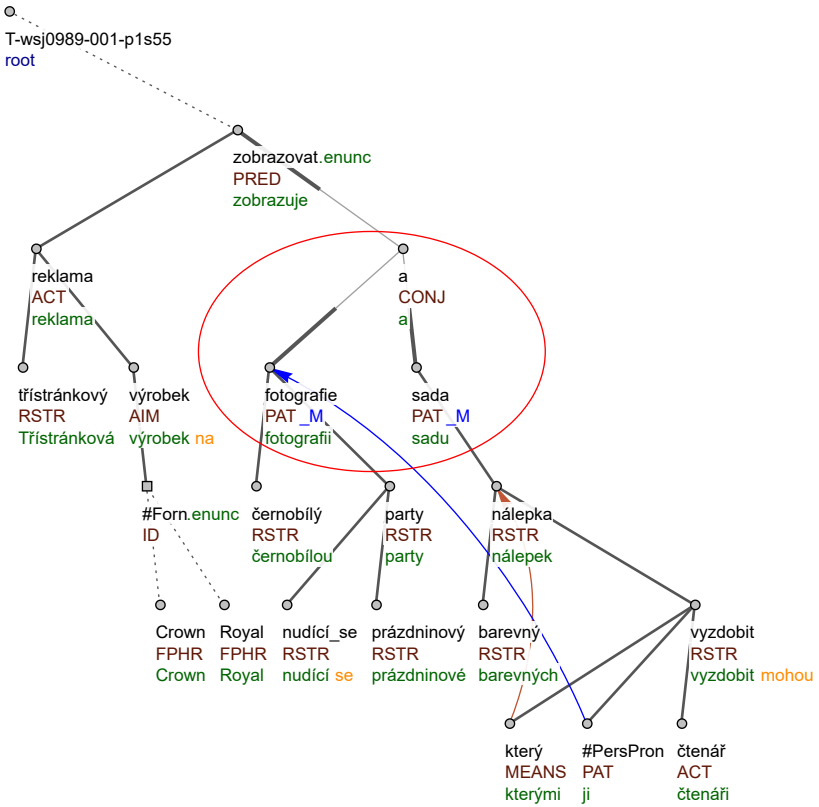
```
T2U::BuildUtree csv=$UFAL_UMR/data/pdt2pb/v5d.csv vallex=/net/data/PDT-C-2.0/dictionaries/pdtvallex-4.5.xml
T2U::ConvertCoreference
T2U::AdjustStructure
Write::UMR
```

```
"""blikat-001""",blikat (v41fxA),vec01173,,,,,,,disagree,both,,,,,0
,ACT: 1,ACT->Protagonist,,,,,,,,,ACT->ARG1/1,"ACT->Protagonist(ARG0/1,ARG1/1)","ACT->Protagonist(ARG0/1,ARG1/1)","ACT->Protagonist(ARG0/1,ARG1/1)",0
,,,,,,,,,,,,,,,,
"""blikat-001""",blikat (v41fxA),vec01605,,,,,,,disagree,both,,,,,0
,ACT: 1,ACT->Source,,,,,,,,,ACT->ARG1/1,"ACT->Source(ARG0/194,ARG1/4)",,,0
,,,,,,,,,,,,,,,,
"""blokovat-001""",blokovat (v41geA),vec00174,,,,,,,disagree,both,,,,,0
,ACT: 1,ACT->Cause,,,,,,,,,"ACT->ARG0/19,ARG1/52,ARG3/5","ACT->Cause(ARG0/183,ARG1/1,ARG3/6)","ACT->Cause(ARG0/183,ARG1/1,ARG3/6)","ACT->Cause(
,PAT: 4,PAT->Event,,,,,ARG1,,,,"PAT->ARG0/1,ARG1/58","PAT->Event(ARG0/1,ARG1/452,ARG2/2)","PAT->Event(ARG0/1,ARG1/452,ARG2/2)","PAT->Event(ARG0/1
,,,,,,,,,,,,,,,,
"""bloudit-001""",bloudit (v41ggA),vec01623,,,,,,,,,,,,,0
,ACT: 1,ACT->Protagonist,,,,,,,,,ACT->Protagonist(),,,0
,,,,,,,,,,,,,,,,
"""bloumat-001""",bloumat (v41ghA),vec01025,,,,,,,,ssc,,,,,0
,ACT: 1,ACT->Mover,ARG0,,,ARG0,,,,,,ACT->Mover(ARG0/24),ACT->Mover(ARG0/24),ACT->Mover(ARG0/21),0
,DIR2: *,DIR2->Path,ARG1,,,ARG1,,,,,,DIR2->Path(ARG1/7),DIR2->Path(ARG1/7),DIR2->Path(ARG1/4),0
,,,,,,,,,,,,,,,,
"""bláznit-001""",bláznit (v41ftB),,,,,,,,,,,,,,,0
,ACT: 1,,,,,,,,,,,,,,,0
,,,,,,,,,,,,,,,,
"""bláznit-002""",bláznit (v41ftA),,,,,,,,,,,,,,,0
,ACT: 1,,,,,,,,,,,,,,,0
```

BuildUtree

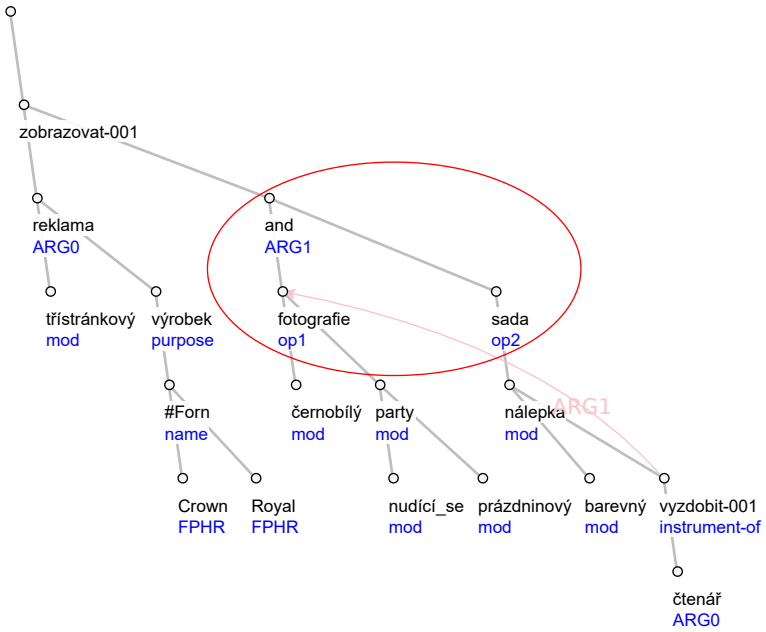
- Walks a t-tree recursively, copies nodes to a new **u-tree**.
- #Neg.RHEM nodes are not copied, parent is negated (polarity).
- Translate valency frame.
- Translate non-valency functors to relations.
- Set alignment.
- Deduce aspect (state verbs listed, the rest is taken from the m-tag and m-lemma).
- Set polarity based on grammatemes (negation or indeftype or m-tag if grammatemes are missing).
- Resolve coordination.

Coordination (PDT)



Třístránková reklama na výrobek Crown Royal zobrazuje černobílou fotografii nudící se prázdninové party - a sadu barevných nálepek, kterými ji mohou čtenáři vyzdobit.

Coordination (UMR)



Třístránková reklama na výrobek Crown Royal zobrazuje černobílou fotografii nudící se prázdninové party - a sadu barevných nálepek, kterými ji mohou čtenáři vyzdobit.

What if different functors are coordinated?

We use the **most frequent functor**.

What if different functors are most frequent?

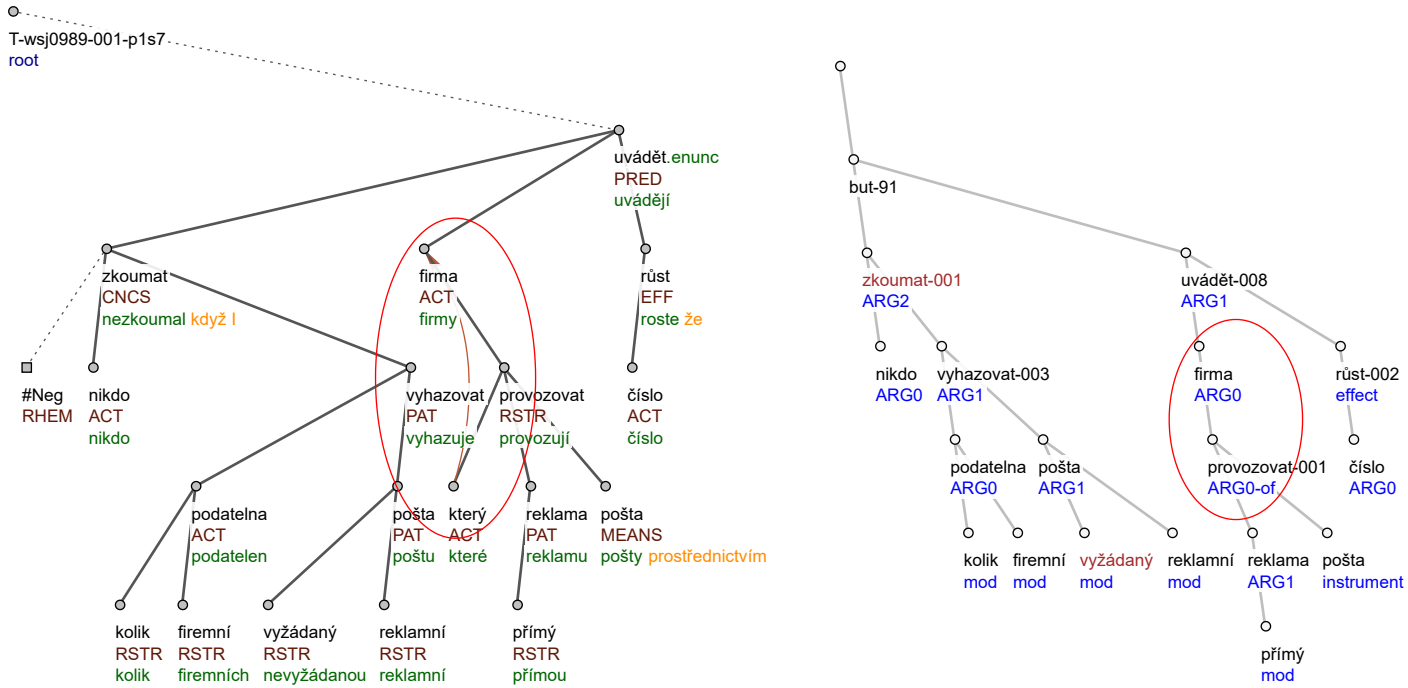
We select a **random one**.

ConvertCoreference

Create a graph from all the textual and grammatical coreference links and process the nodes topologically sorted.

- Remove INTF nodes.
- Keep the link if it leads to a different sentence.
- For #Cor, #QCor, #PersPron, and relative pronouns, remove the node and "join" the chain.
- For relative pronouns in RSTR sentences, try to "reverse" the relation to *-of.

“Reversed” relations

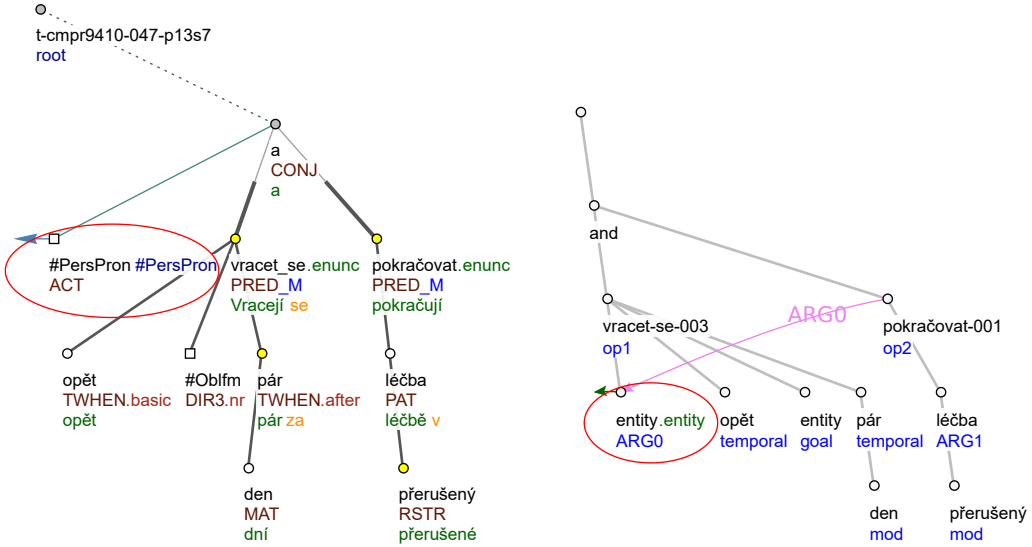


I když nikdo nezkoumal, kolik firemních podatelen nevyžádanou reklamní poštu vyhazuje, firmy, které provozují přímou reklamu prostřednictvím pošty, uvádějí, že číslo roste.

AdjustStructure

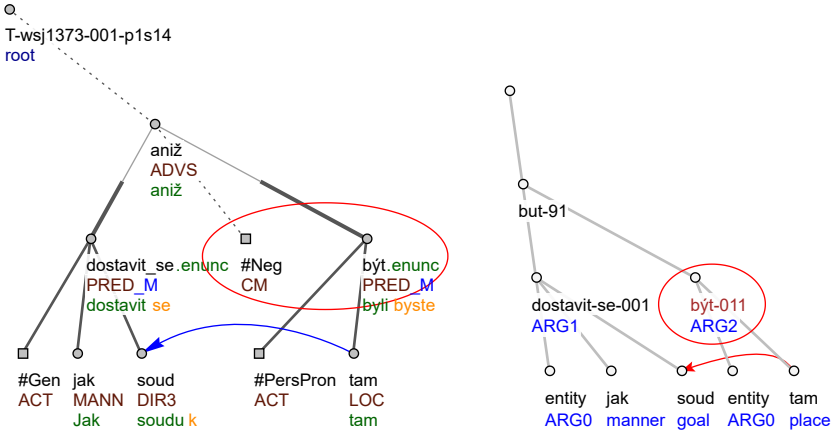
- Translate COMPL and its secondary dependency.
- For CONTRD and CNCS, restructure the subordinate structure to a coordination.
- Move common dependent nodes to the first coordination member, use reentrancy to mark their relation to the remaining members.
- Remove duplicate edges (fallout of "reversed" relations).
- Negate siblings of negative RHEMs and CMs (special handling of coordination).

Common Dependent Nodes



Vracejí se opět za pár dní a pokračují v přerušené léčbě.

Negation in Coordination



Jak se dostavit k soudu, aniž byste tam byli

Write::UMR

- Node identifiers (like s3p1) are created here.
- Cataphoric links are reverted.

Rewrite Needed?

- Using t-layer in later steps: sometimes, the correspondence is broken by removed nodes or changed structure.
- Non-determinism (topological sort).
- Tricks and hacks, increasingly difficult to extend.

Thank you

```
#################################################################################
# sent_id = u_tree-cs-s7-root
# :: snt7
Index: 1  2     3     4  5   6       7     8  9  10     11 12 13    14   15        16  17        18        19          20       21      22      23
Words: Na druhém místě s  13.9 procenty hlasů (  18 mandátů )  je zatím blok Spolehlivý dům dosavadního nejvyššího představitele republiky Arnolda Rüütela .

# sentence level graph:
(s7b1 / být-011
      :place (s7m1 / místo
         :mod (s7d1 / dva))
      :companion (s7p1 / procento
         :mod (s7x1 / 13.9)
         :MAT (s7h1 / hlas)
         :PAR (s7m2 / mandát
            :mod (s7x2 / 18)))
      :temporal (s7z1 / zatím)
      :ARG0 (s7b2 / blok
         :ID (s7d2 / dům
            :mod (s7s1 / Spolehlivý))
         :poss (s7r1 / Rüütel
            :mod (s7p2 / představitel
               :mod (s7d3 / dosavadní)
               :mod (s7v1 / vysoký)
               :ARG1 (s7r2 / republika
                  :refer-number singular))
            :mod (s7a1 / Arnold)))
      :aspect activity
      :modal-strength full-affirmative)

# alignment:
s7b1: 12-12
s7m1: 1-1,3-3
s7d1: 2-2
s7p1: 4-4,6-6
s7x1: 5-5
s7h1: 7-7
s7m2: 10-10
s7x2: 9-9
s7z1: 13-13
s7b2: 14-14
s7d2: 16-16
s7s1: 15-15
s7r1: 22-22
s7p2: 19-19
s7d3: 17-17
s7v1: 18-18
s7r2: 20-20
s7a1: 21-21

# document level annotation:
(s7s0 / sentence
   :coref ((s7r2 :same-entity s5e1)))
```