PARSEME Shared Task and Prague Dependency Treebank: Comparison and Data Conversion

WG4

Eduard Bejček, Pavel Straňák, Zdeňka Urešová
Charles University in Prague, MFF,
Institute of Formal and Applied Linguistics
{bejcek,stranak,uresova}@ufal.mff.cuni.cz

WORK IN PROGRESS

PARSEME Shared Task

- 2015 -- 2017
- MWEs study + their annotation in treebanks
- automatic detection of verbal MWEs
- set of standardized annotation principles
- 20 languages: Czech among others
- 5 groups of MWEs + language specific

Verbal MWE types for Czech

- ID: idiom
- LVC: light verb construction
- IReflV: inherently reflexive verbs
- OTH: other types
- *neither VPC (verb-particle combination) nor language specific category for Czech*

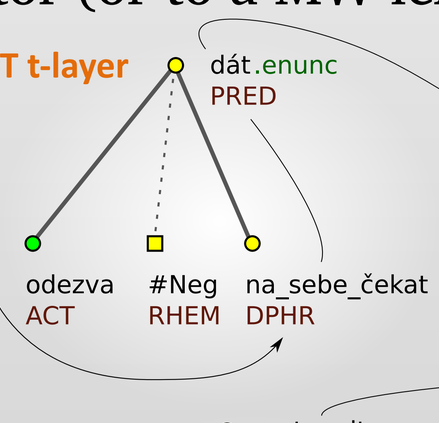
To minimize annotation work, it should be transformed from PDT data.

ID ...corresponds to a DPHR functor (or to a MW lexeme)

1. Input text

Odezva **na** sebe nedala čekat.
Reaction on itself not-gave wait.
The reaction didn't keep us waiting.

2. PDT t-layer



3. Output annotation

<MWE category="ID">
Odezva na sebe nedala čekat.
didn't keep us waiting

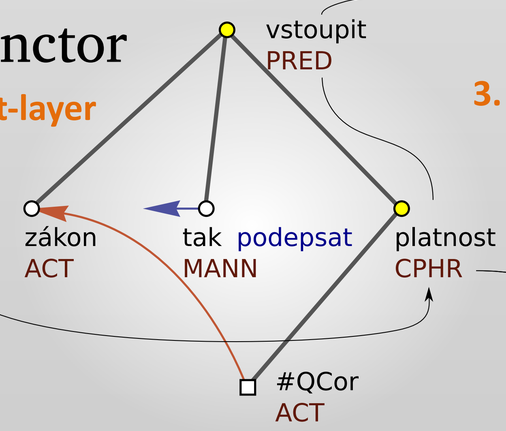
Condition: DPHR node
Range: all words mentioned in DPHR node + its governing verb node

LVC ...corresponds to a CPHR functor

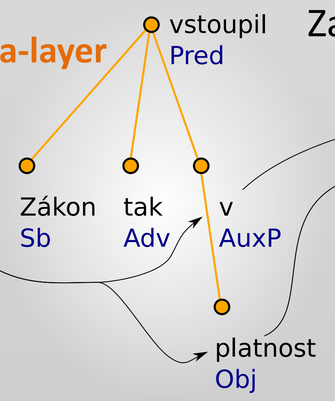
1. Input text

Zákon tak vstoupil v **platnost**.
Law so came into force.
By that the law has come into force.

2. PDT t-layer



3. PDT a-layer



4. Output annotation

<MWE category="LVC">
Zákon tak vstoupil v platnost.
has come into force

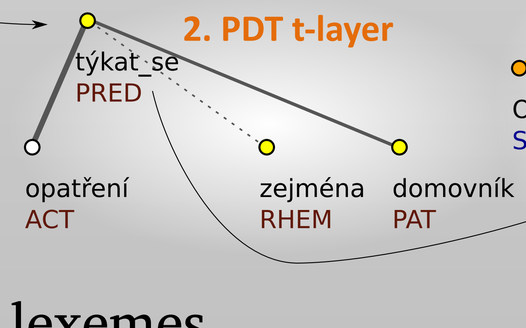
Condition: CPHR node
Range: noun in DPHR node + its governing verb node + a preposition, if exists with the noun

IReflV ...corresponds to a reflexive t-lemma

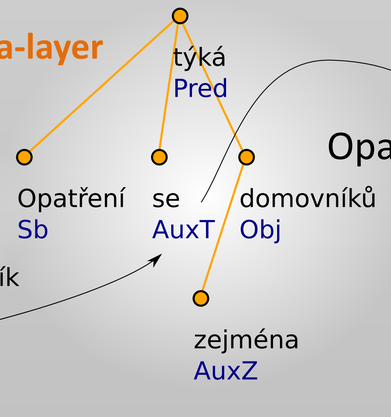
1. Input text

Opatření **se** týká zejména domovníků.
The meassure involves chiefly housekeepers.

2. PDT t-layer



3. PDT a-layer



4. Output annotation

<MWE category="IReflV">
Opatření se týká zejména domovníků.
involves

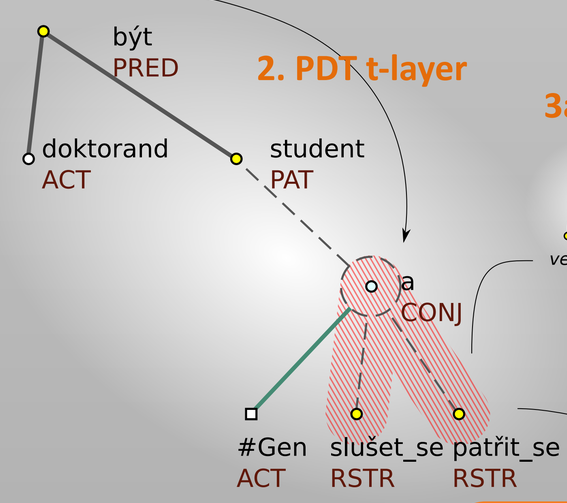
Condition: t-lemma with a reflexive particle + the particle's type is AuxT
Range: verb + reflexive particle

OTH ...corresponds to **verbal** MW lexemes

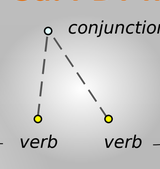
1. Input text

Doktorand je studentem, jak se **sluší** a patří.
PhD-student is student, as <REFL> suits and befits.
A PhD student is a student, as he should be.

2. PDT t-layer



3a. PDT m-layer



3b. SemLex

```
...
BASIC_FORM:     jak se sluší a patří
LEMMATIZED:     jak se slušet a patřit
MORPHO_TAGS:    RB  RFL  VBZ  CC  VBZ
PDT_FREQ:       1
POS:            verb
...
```

4. Output annotation

<MWE category="OTH">
Doktorand je studentem, jak se sluší a patří.
as he should be

Condition: MWE, type 'lexeme'
+ neither idiom nor LVC + head is not a verb
+ it contains verb or is marked as verb in lexicon
Range: contents words are listed in MWE + auxiliary words -- to be done

LINDAT CLARIN

This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

This research has been supported by grant COST CZ LD14117 of the Ministry of Education, Youth and Sports of the Czech Republic.

PARSEME

7th and Final General Meeting,
Dubrovnik, Croatia,
26 – 27 September, 2016