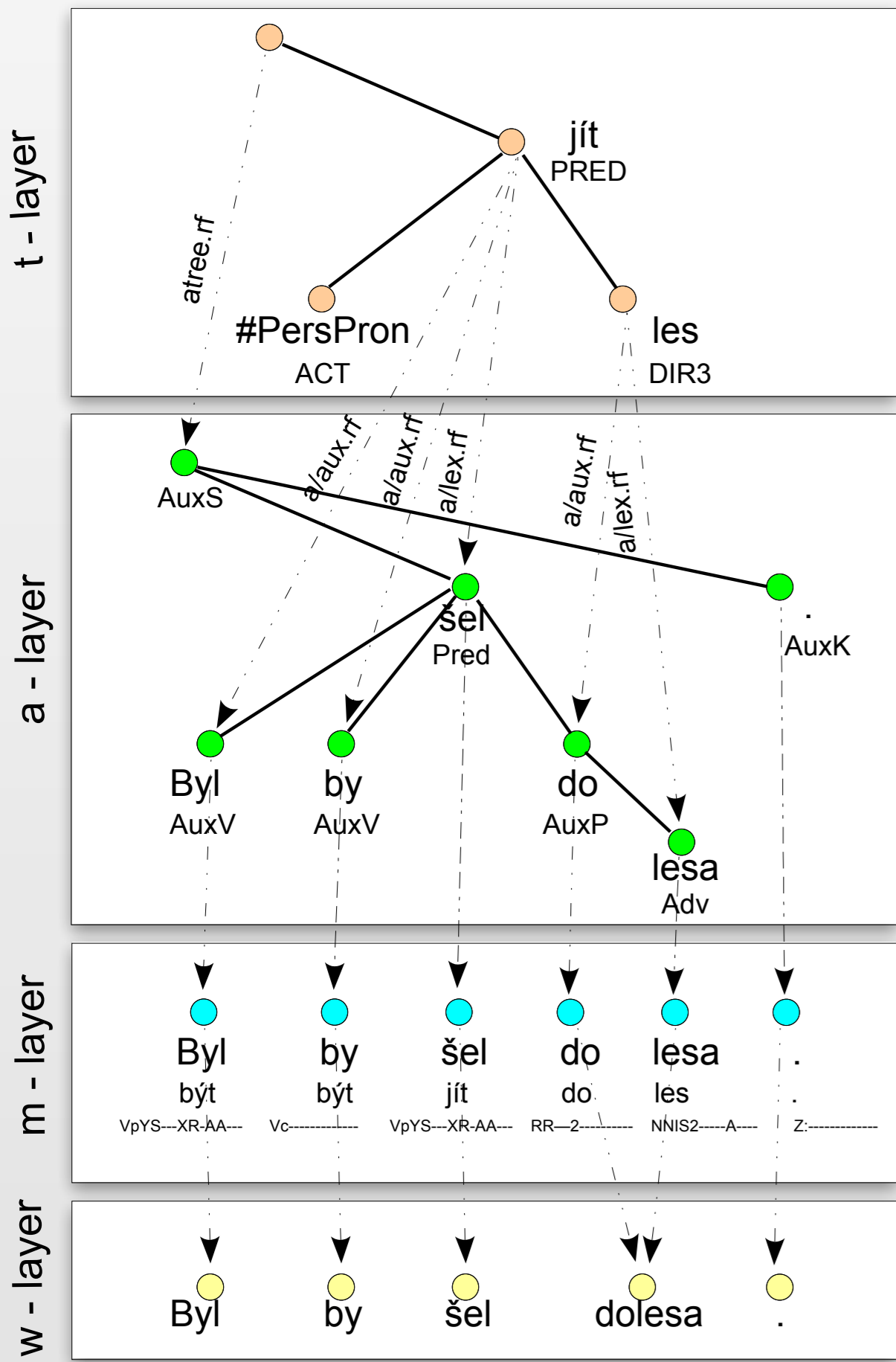# Annotation of Multiword Expressions in the Prague Dependency Treebank

Eduard Bejček and Pavel Straňák
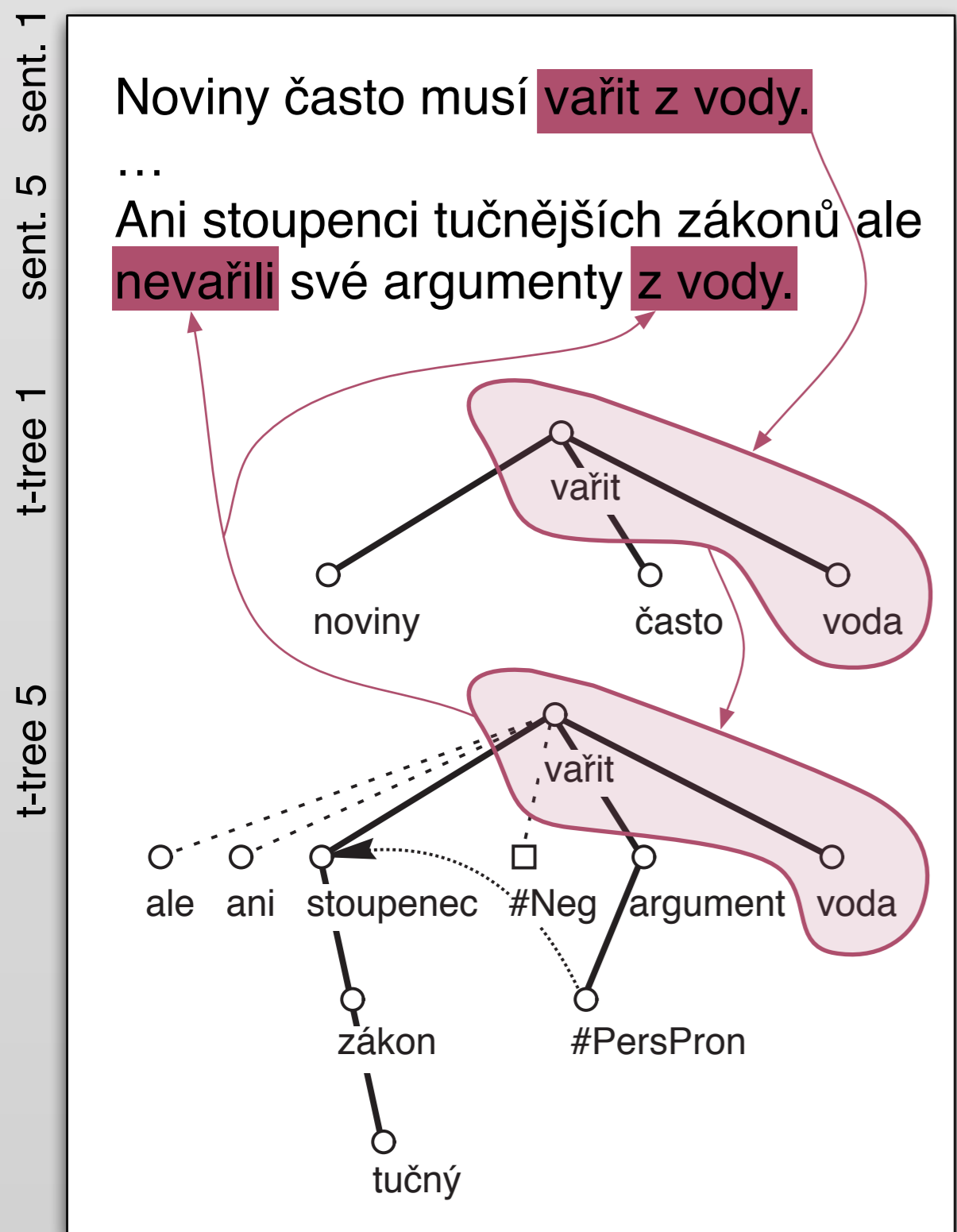
ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

We annotate multiword expressions (MWEs) in Prague Dependency Treebank (❶).

We use PML scheme to capture the annotation of whole PDT as well as of our MWEs (❷). Our aim is to improve a representation of MWEs (incl. named entities) in t-trees (❸).

Our annotators work with an annotation tool (❺) and mark occurences in sentences, which are generated from the deep syntactic layer (t-layer) of PDT: see ❶, ❷, ❹. A MWE lexicon SemLex is produced (❻).

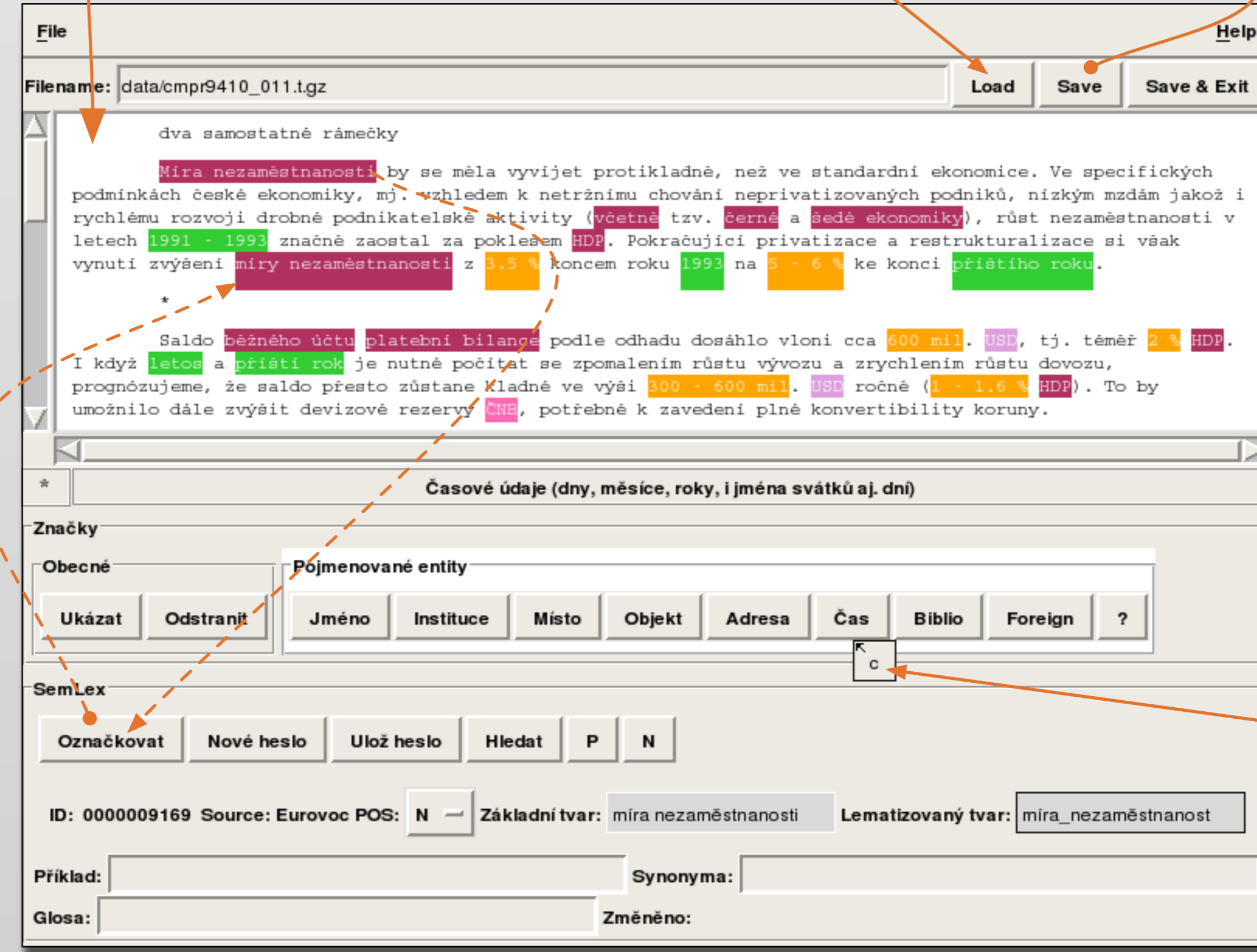We use PML-TQ (❼) to query PDT including the MWE data.
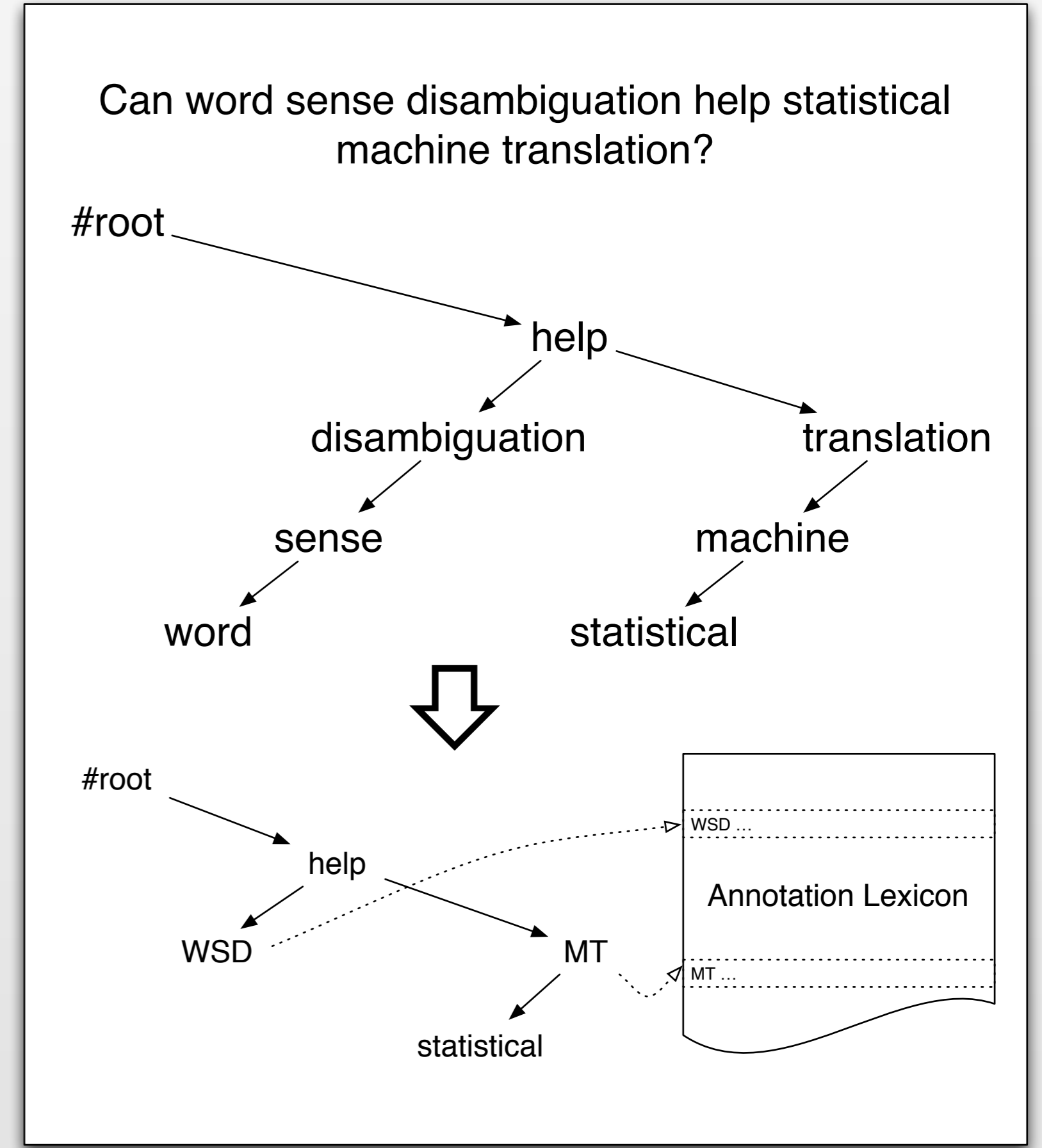

❶ PDT 2.0

❷ PML files and SemLex

Can word sense disambiguation help statistical machine translation?

❸ Scheme of changes

Load + automatic preannotation (from SemLex)

❺ Annotation interface

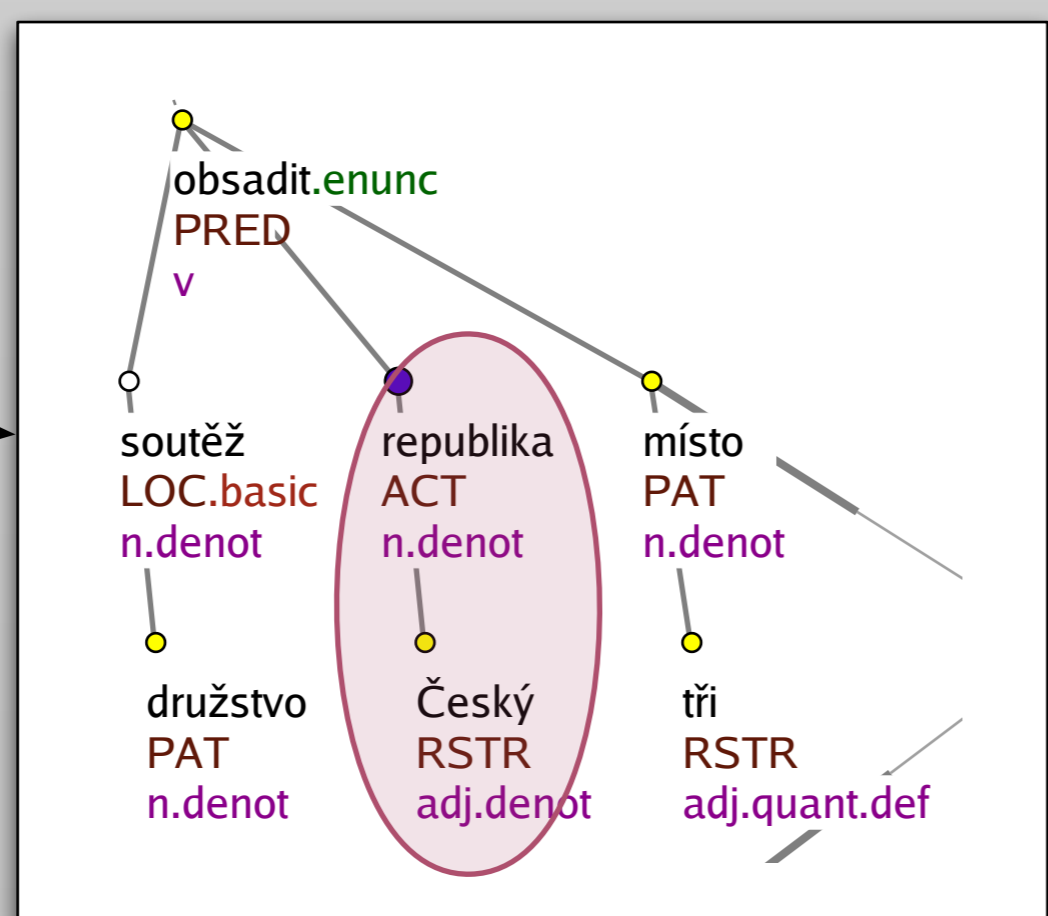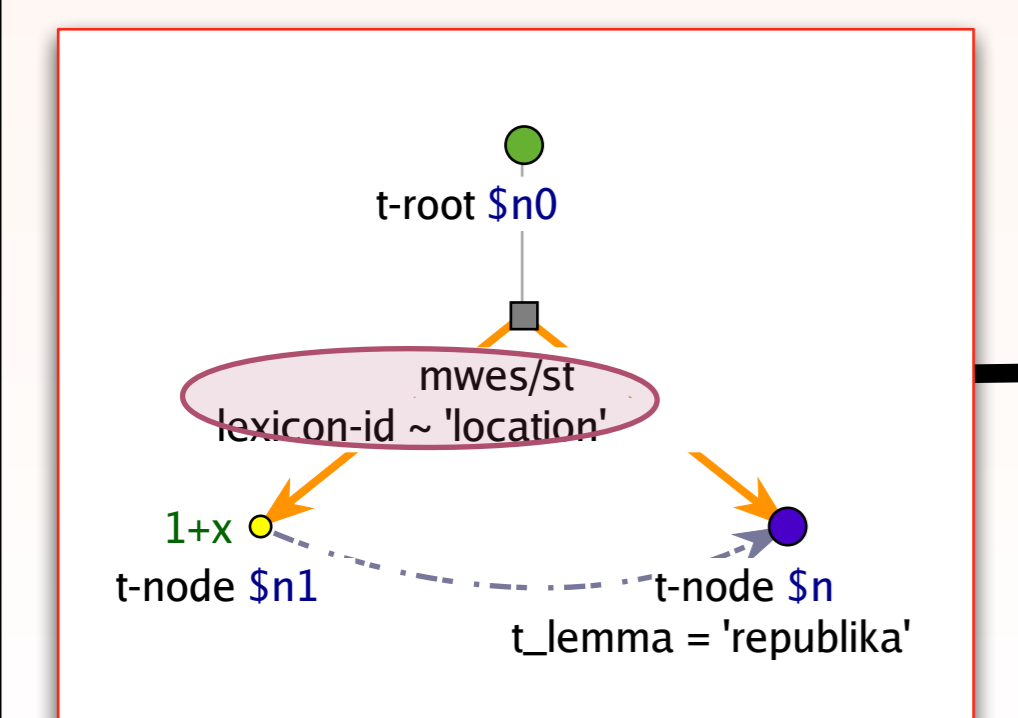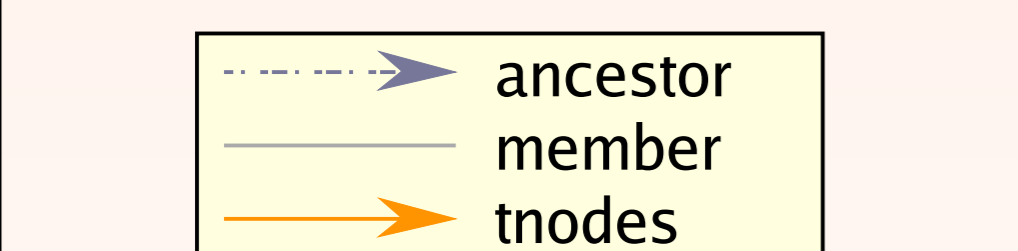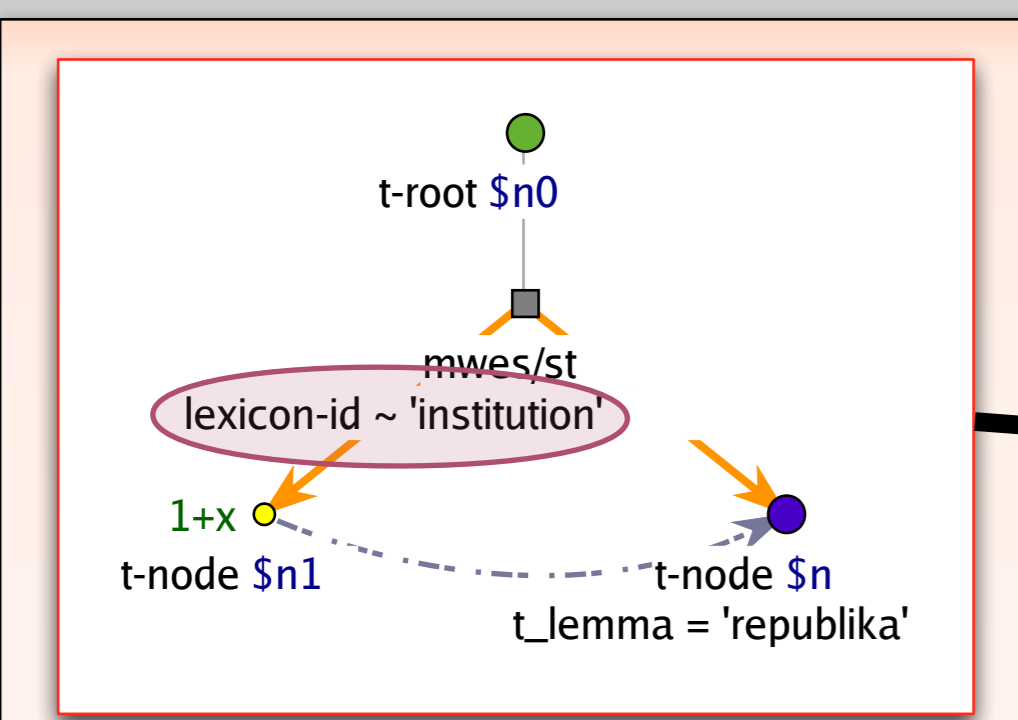Annotate a *new* MWE + automatic pre-annotation of all other occurrences (see ❹)

```
- !!perl/hash:SemLex_heslo
  BASIC_FORM: vykročit ve stopách
  CREATED: 090820221315
  EXAMPLE: "vykročila ve stopách svého otce"
  GLOSS: "následovat, započít stejnou životní dráhu"
  ID: 0000030967
  LEMMATIZED: vykročit ve stopa
  MODIFIED: 090820221334
  MODIFIER: vimmrova
  ORIGID: ~
  POS: V
  SOURCE: vimmrova
  SYNONYMS: ~
```
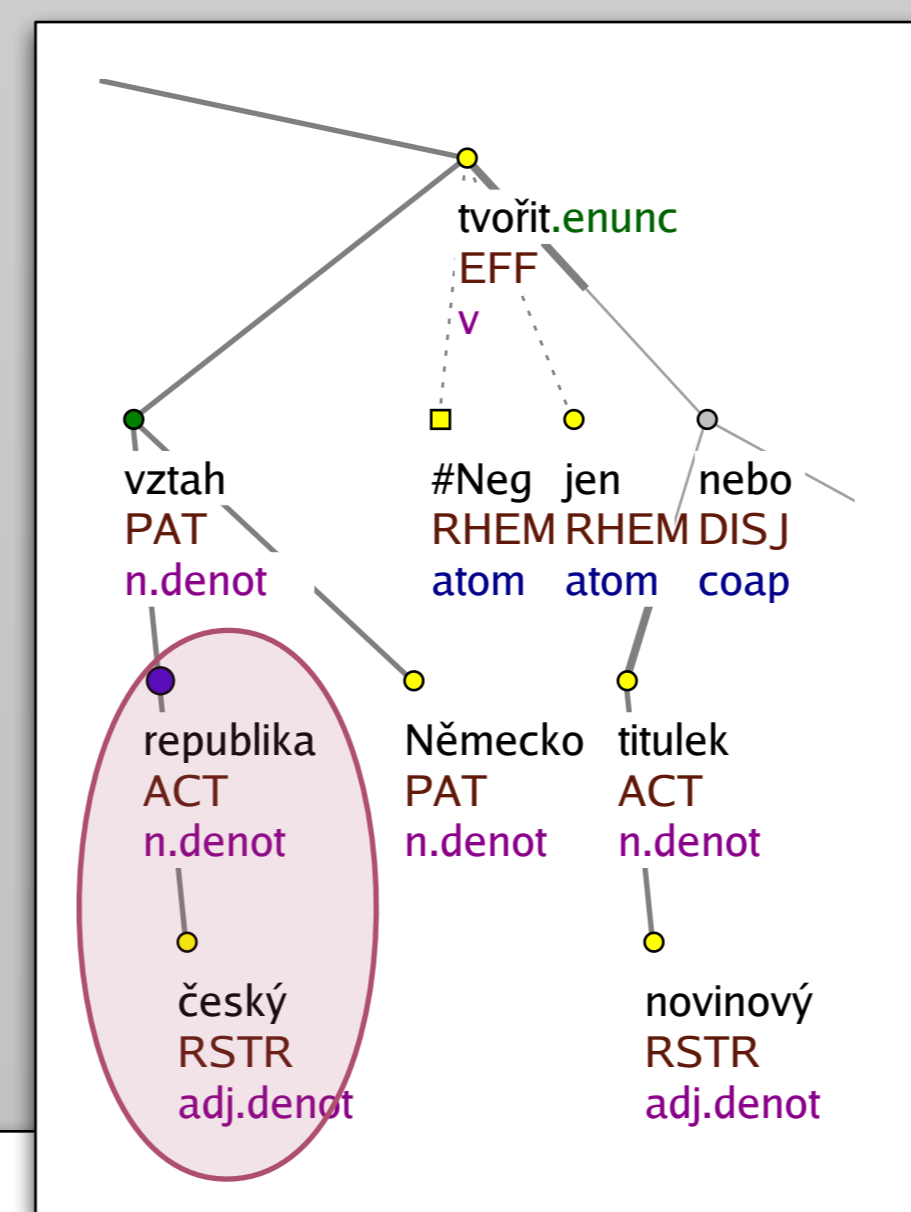
❻ A SemLex entry ("follow in one's footsteps")

one-key shortcuts for annotation

searchable SemLex browser and editor

❹ Use of t-trees in the pre-annotation

Noviny často musí vařit z vody.
…
Ani stoupenci tučnějších zákonů ale nevařili své argumenty z vody.

❼ PML–TQ queries over MWEs in the PDT

ancestor
member
tnodes

In the competition [of] teams [the] Czech republic took the 3rd place …

In the 4th quarter of the year … came into [the] Czech Republic …

Only 19 people in [the] Czech Republic …

```
t-root $n0 := ...
[ member mwes/st
    [ lexicon-id ~ 'institution',
        tnodes t-node $n1 :=
        [ ancestor $n ],
        tnodes t-node $n :=
        [ t_lemma = 'republika' ] ] ];

>> give distinct $n0.id,concat($n1.id, ' ' over $n sort by
$n1.deepord),$n.id,concat($n1.t_lemma, ' ' over $n sort by
$n.deepord),$n.t_lemma sort by $1
```

An output filter can produce various lists from results of a query