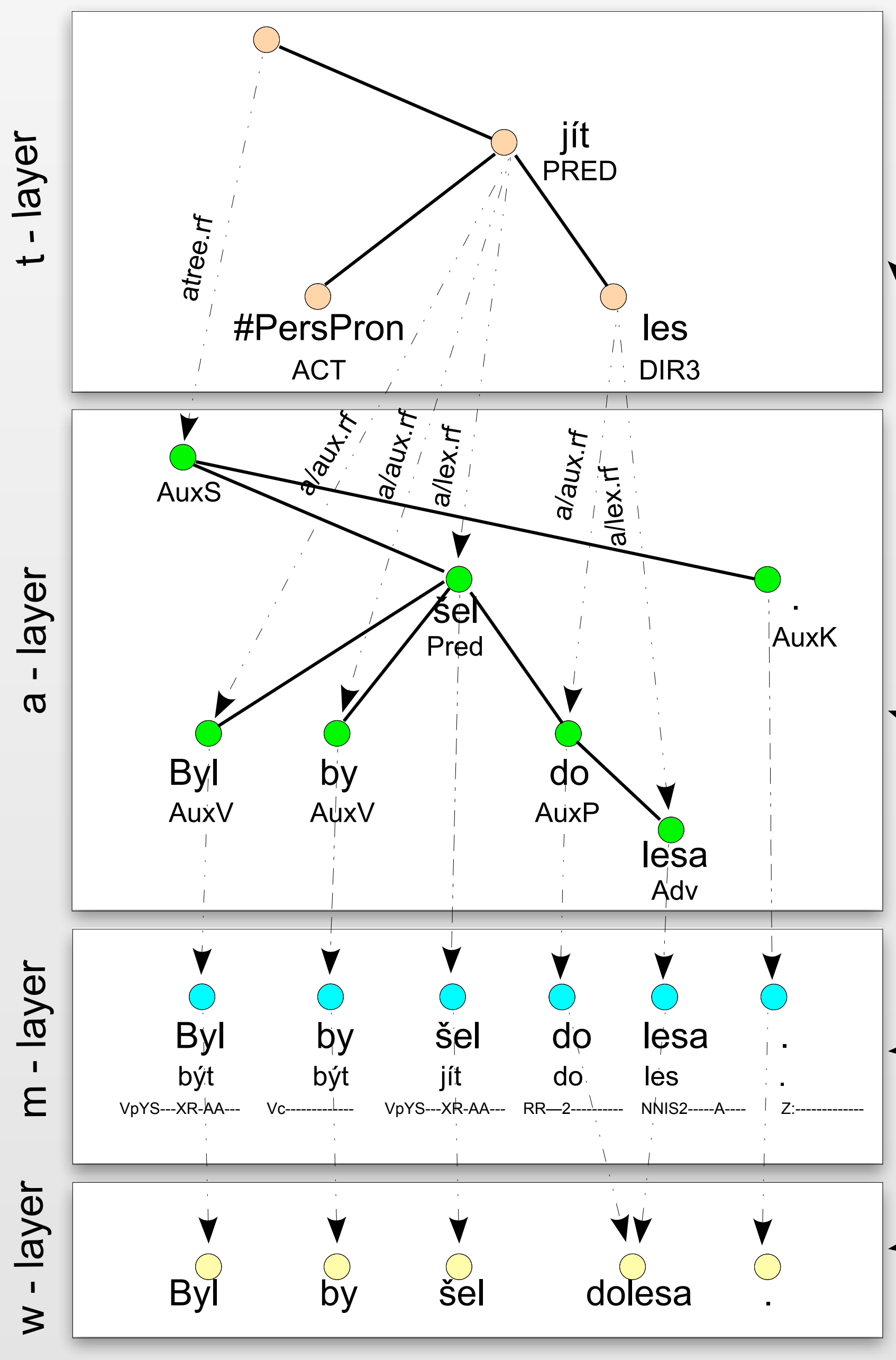
Annotation of Multiword Expressions in the Prague Dependency Treebank

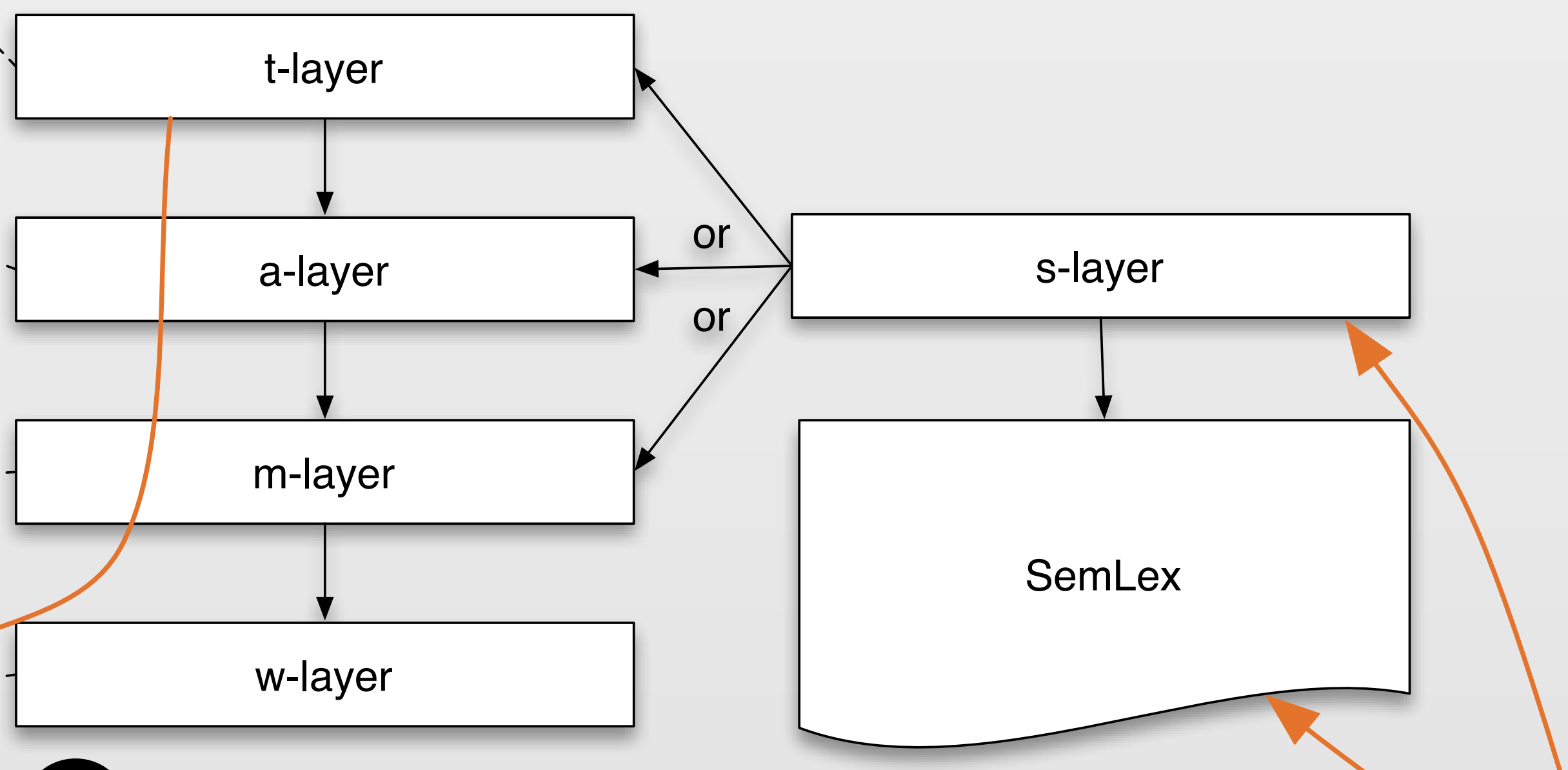
Eduard Bejček and Pavel Straňák

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

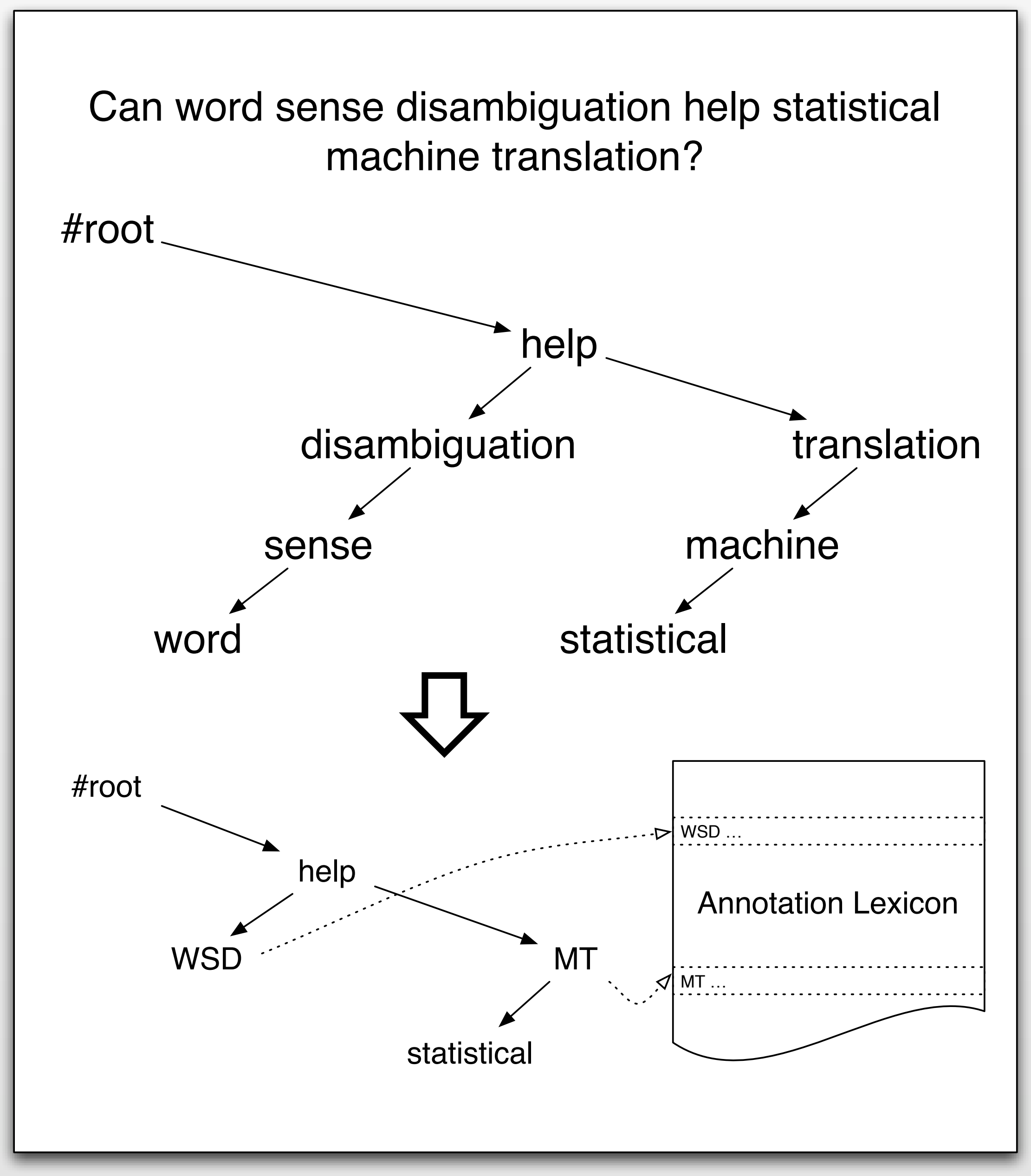
We annotate multiword expressions (MWEs) in Prague Dependency Treebank (❶).
We use PML scheme to capture the annotation of whole PDT as well as of our MWEs (❷).
Our annotators work with an annotation tool (❺) and mark occurences in sentences, which are generated from the deep syntactic layer (t-layer) of PDT: see ❶, ❷, ❹.
Our aim is to improve a representation of MWEs (incl. named entities) in t-trees (❸), because current approach is considered to be insufficient (❻).



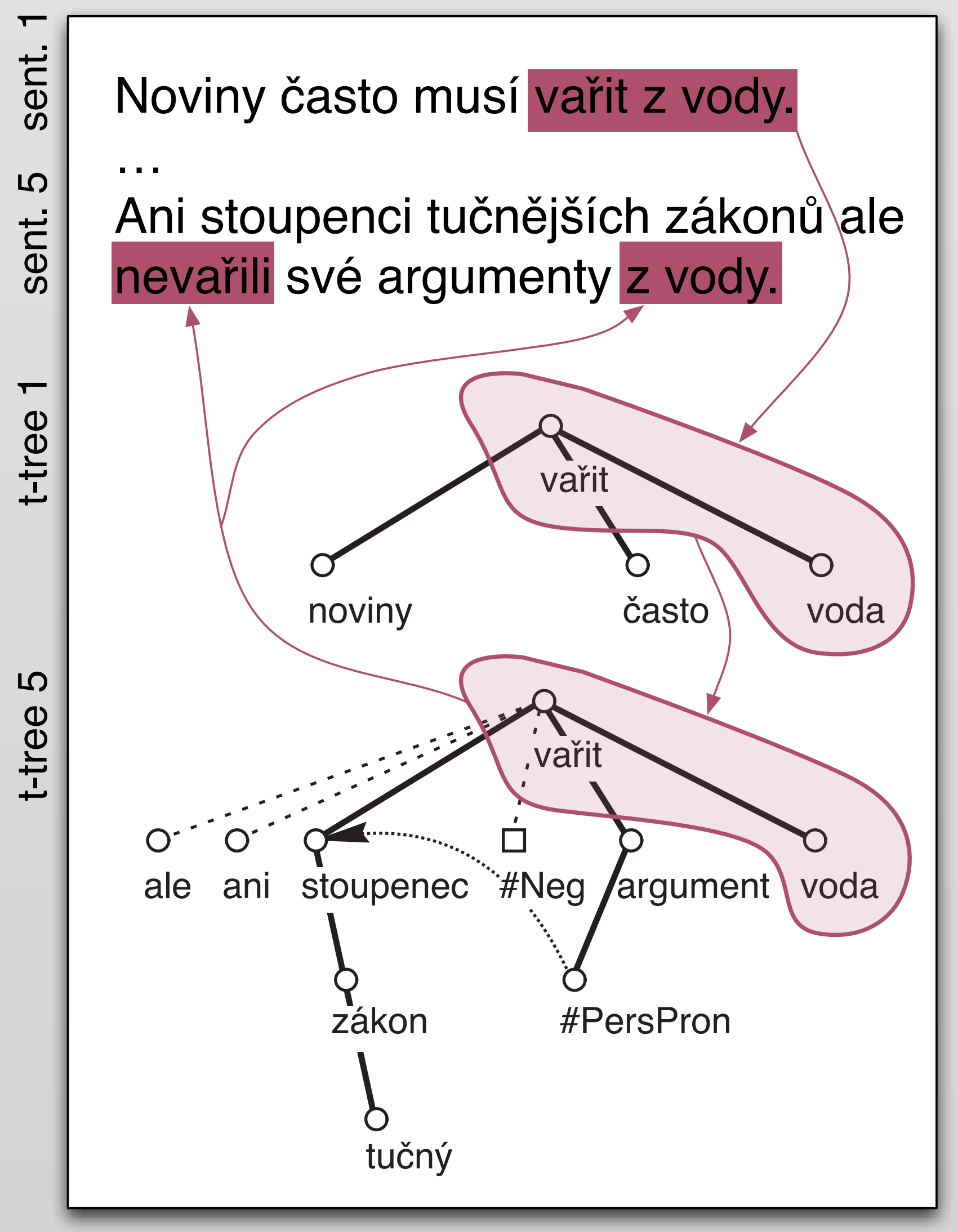
❶ PDT 2.0



❷ PML files and SemLex



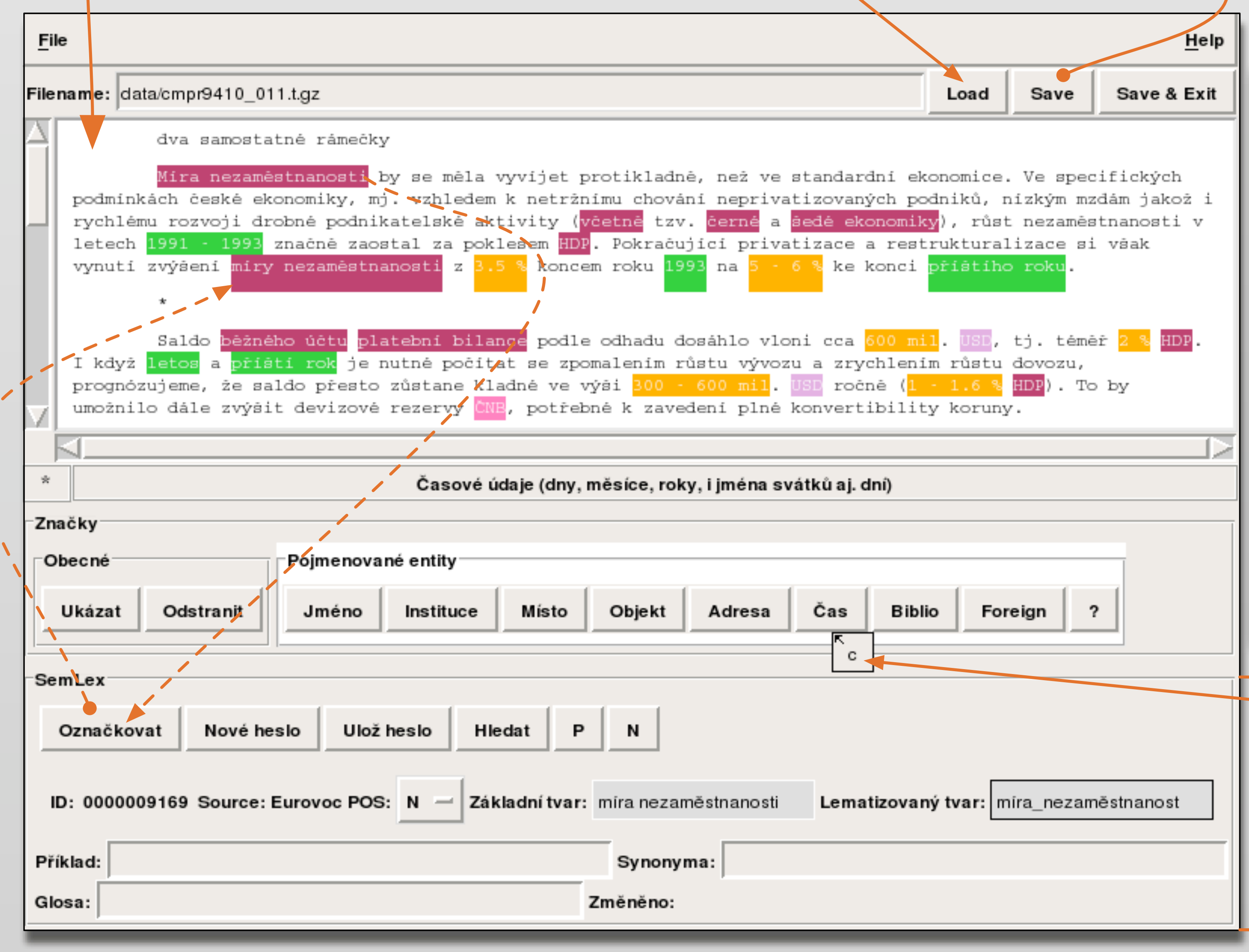
❸ Scheme of changes



❹ Use of t-trees in the pre-annotation

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

Load + automatic preannotation (from SemLex)


File Help

Filename: data/cmpr9410_011.t.gz   Load   Save   Save & Exit

```
        dva samostatné rámečky

        Míra nezaměstnanosti by se měla vyvíjet protikladně, než ve standardní ekonomice. Ve specifických
podmínkách české ekonomiky, mj. vzhledem k netržnímu chování neprivatizovaných podniků, nízkým mzdám jakož i
rychlému rozvoji drobné podnikatelské aktivity (včetně tzv. černé a šedé ekonomiky), růst nezaměstnanosti v
letech 1991 - 1993 značně zaostal za poklesem HDP. Pokračující privatizace a restrukturalizace si však
vynutí zvýšení míry nezaměstnanosti z 3.5 % koncem roku 1993 na 5 - 6 % ke konci příštího roku.

        *

        Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 mil. USD, tj. téměř 2 % HDP.
I když letos a příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu,
prognózujeme, že saldo přesto zůstane kladné ve výši 300 - 600 mil. USD ročně (1 - 1.6 % HDP). To by
umožnilo dále zvýšit devizové rezervy ČNB, potřebné k zavedení plné konvertibility koruny.
```

* Časové údaje (dny, měsíce, roky, i jména svátků aj. dní)

Značky

Obecné: Ukázat | Odstranit

Pojmenované entity: Jméno | Instituce | Místo | Objekt | Adresa | Čas | Biblio | Foreign | ?

SemLex: Označkovat | Nové heslo | Ulož heslo | Hledat | P | N

ID: 0000009169 Source: Eurovoc POS: N Základní tvar: míra nezaměstnanosti Lematizovaný tvar: míra_nezaměstnanost

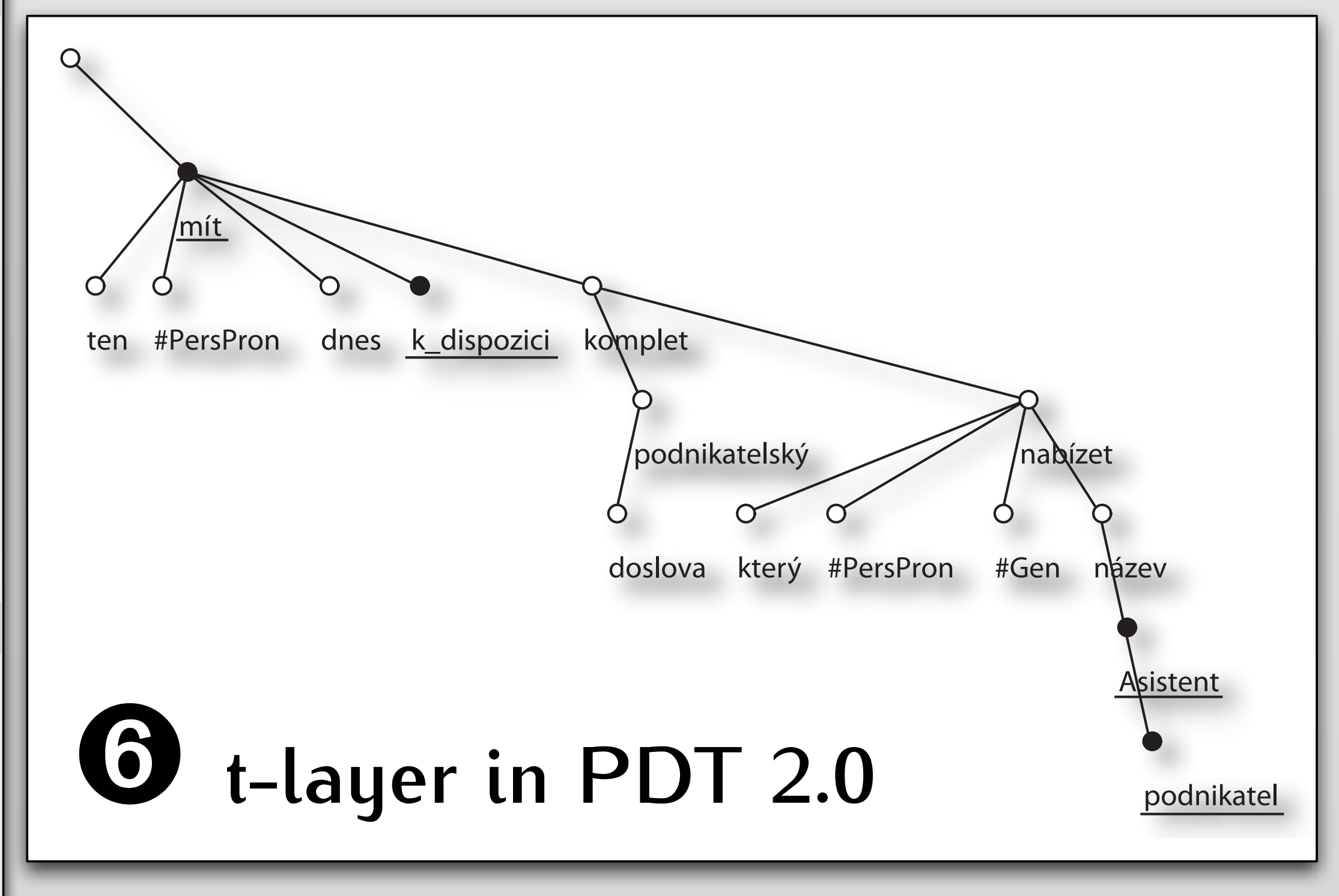
Příklad: Synonyma:

Glosa: Změněno:

one-key shortcuts for annotation

searchable SemLex browser and editor

❺ Annotation interface



❻ t-layer in PDT 2.0

Inter-annotator Agreement

Each t-node may be: (i) annotated with a SemLex entry (ii) one of nine types of named entities or (iii) not annotated. This yields a scale between full agreement and none. Each type of agreement is assigned a weight according to approximate amount of information it provides (❼).

Then we use the slightly modified pi measure (❽) on these weighted values to compute inter-annotator agreement.

	Agreement				Disagreement
	Annotated			Not annot.	
	Agreement on NE / lexia				
	Full agr.	Disagr.			
class c	1	2	3	4	5
# of t-nodes n	10,527	2,365	389	83,287	3,988
weight w	1	0.5	0.25	0.052	0

❼ Annotated t-nodes

$$\pi_w = \frac{A_o - A_e}{\widehat{U} - A_e}$$

$$\widehat{U} = \frac{n_{A \cup B}}{N} + 0.052 \cdot \frac{N - n_{A \cup B}}{N} = 0.215$$

$$\pi_w = \frac{A_o - A_e}{\widehat{U} - A_e} = \frac{0.160 - 0.047}{0.215 - 0.047} = 0.676$$

❽ Agreement measure

Current State

- We use tectogrammatical tree structures of MWEs for automatic pre-annotation.
  - The richer the tectogrammatical annotation the better the possibilities for automatic pre-annotation, which minimises human errors
- Weighted measure that accounts for partial agreement as well as estimation of maximal agreement
  - The resulting $\pi_w = 0.676$ is statistically significant
  - Agreement should gradually improve as:
    - we clean up the annotation lexicon
    - more entries are pre-annotated automatically
    - and further types of pre-annotation are employed.