

# Annotation of Multiword Lexias in the PDT 2.0

Pavel Straňák

# Outline

---

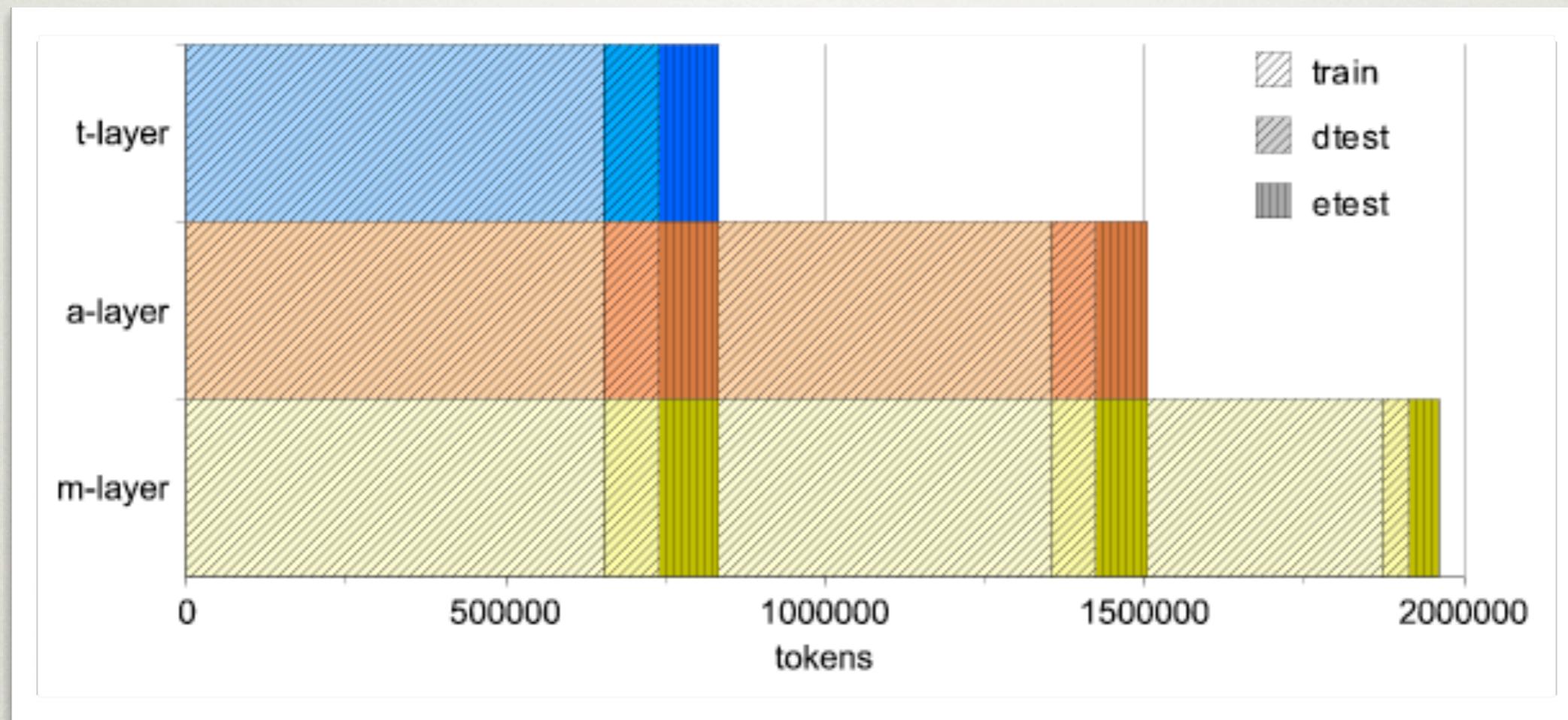
- Prague Dependency Treebank
  - Units of t-layer
  - Annotations of multiword units
- Our project
  - Multiword lexias (MWL)
  - Annotation of MWLs
  - Evaluating reliability of annotations

# Prague Dependency Treebank (PDT)

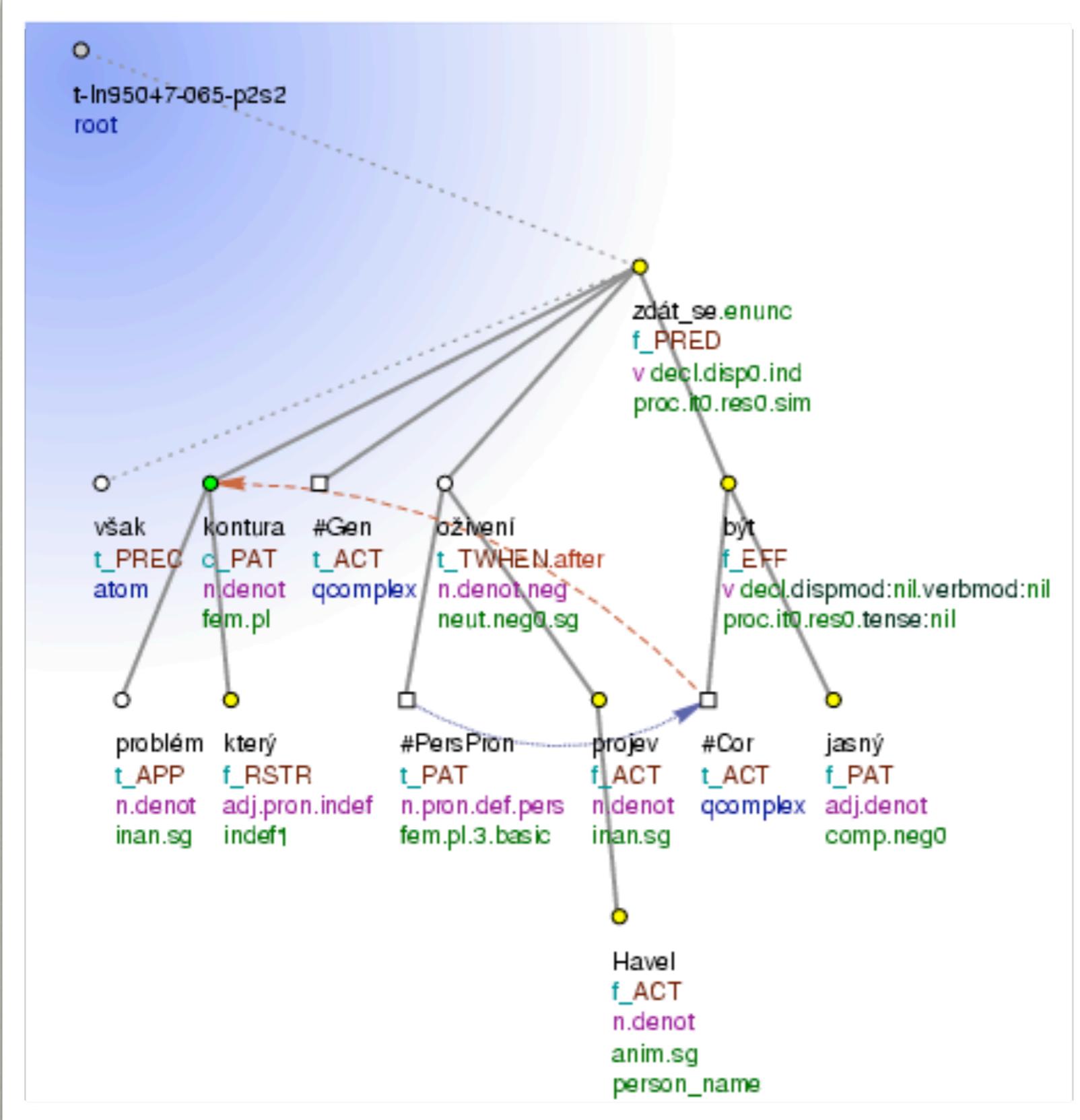
---

- 3 layers of annotation:
  - w-layer: segmentation and tokenization
  - m-layer: lemmas and morphological tags
  - a-layer: analytical (surface) dependency trees
  - t-layer: tectogrammatical (deep) dependency trees

# Layers of Annotation



# Tectogrammatical tree



# Units of the t-layer (ideal)

---

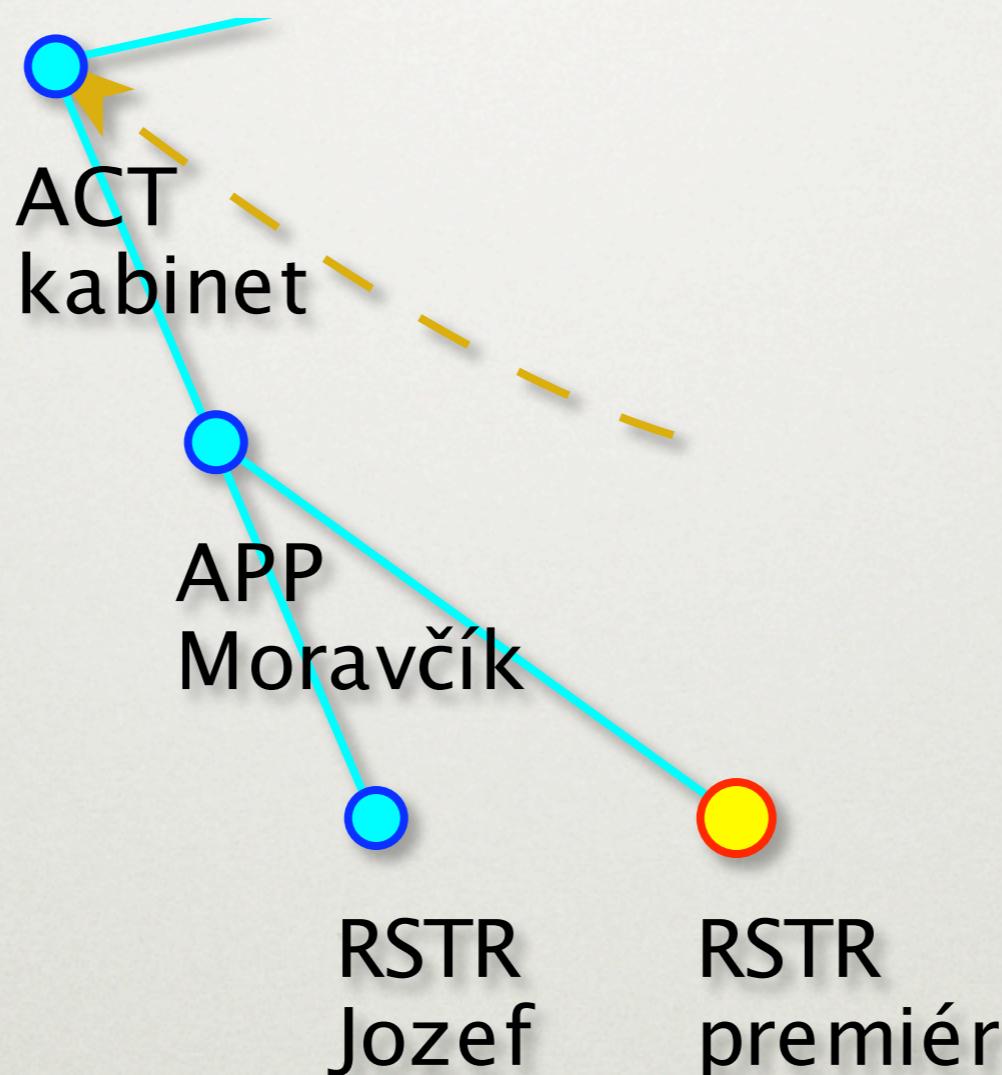
- *Lexia* (Filipčík and Čermák, 1986) is equivalent to a “monosemic lexeme” of Filipčík (1994) or a “lexical unit” of Cruse (1986): “*a pair of a single sense and a basic form (plus its derived forms) with relatively stable semantic properties*”.

compare

- *semanteme* – Melčuk (Dep. Syn.) – unit of DSyntR: “*a language-specific semantic unit which corresponds to one particular word-sense or, more precisely, to the signifié of a (desambiguated) lexical unit.*”

# Units of the t-layer (PDT 2.0)

---



# Information in the PDT

---

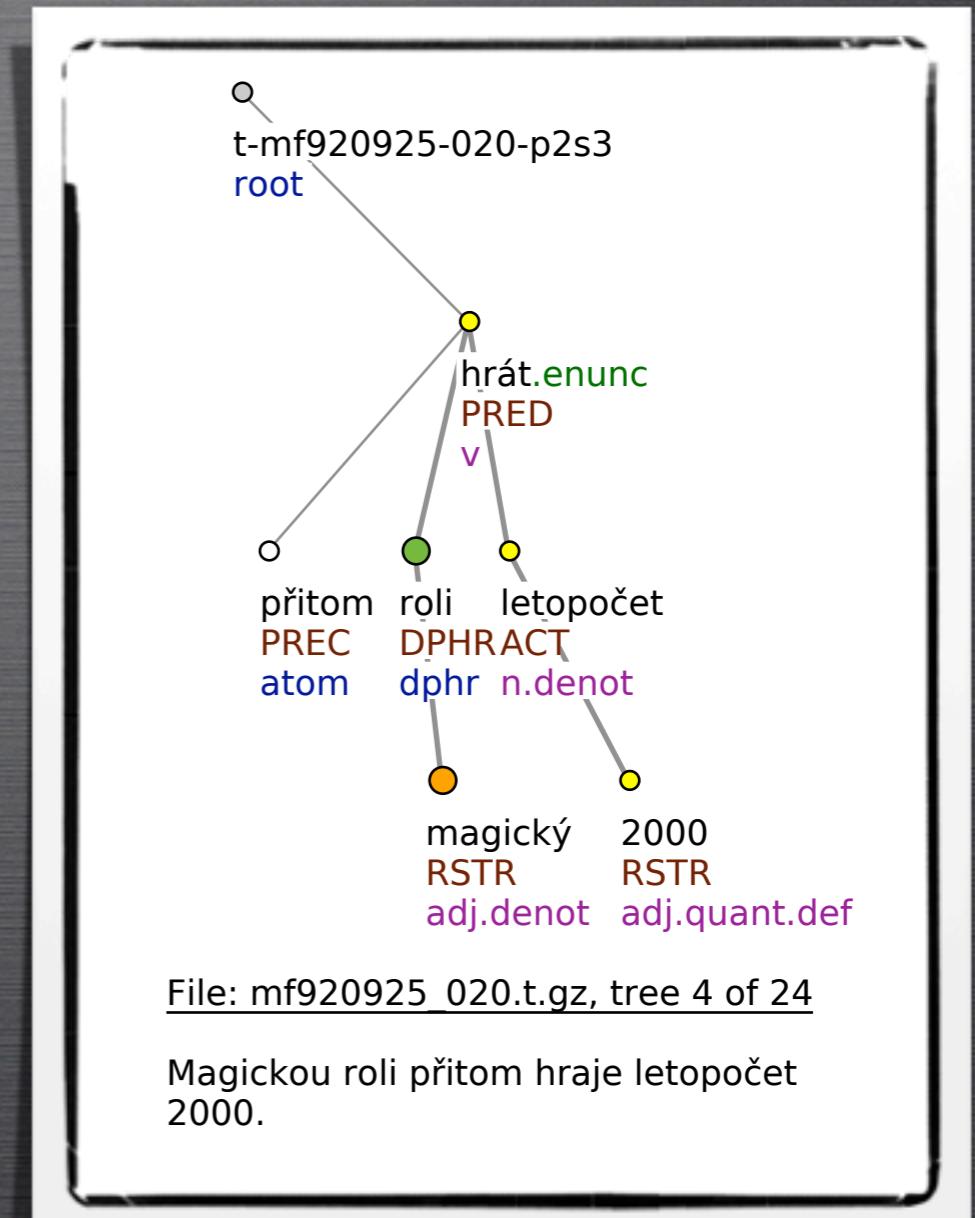
- Semantic Role Labels (Syntactic, actually), technical lemmas, attributes
- SRLs: ID, FPHR, CPHR, DPHR
- t-lemmas: #Idph, #Forn
- attribute of a t-node: ‘is-name-of-person’

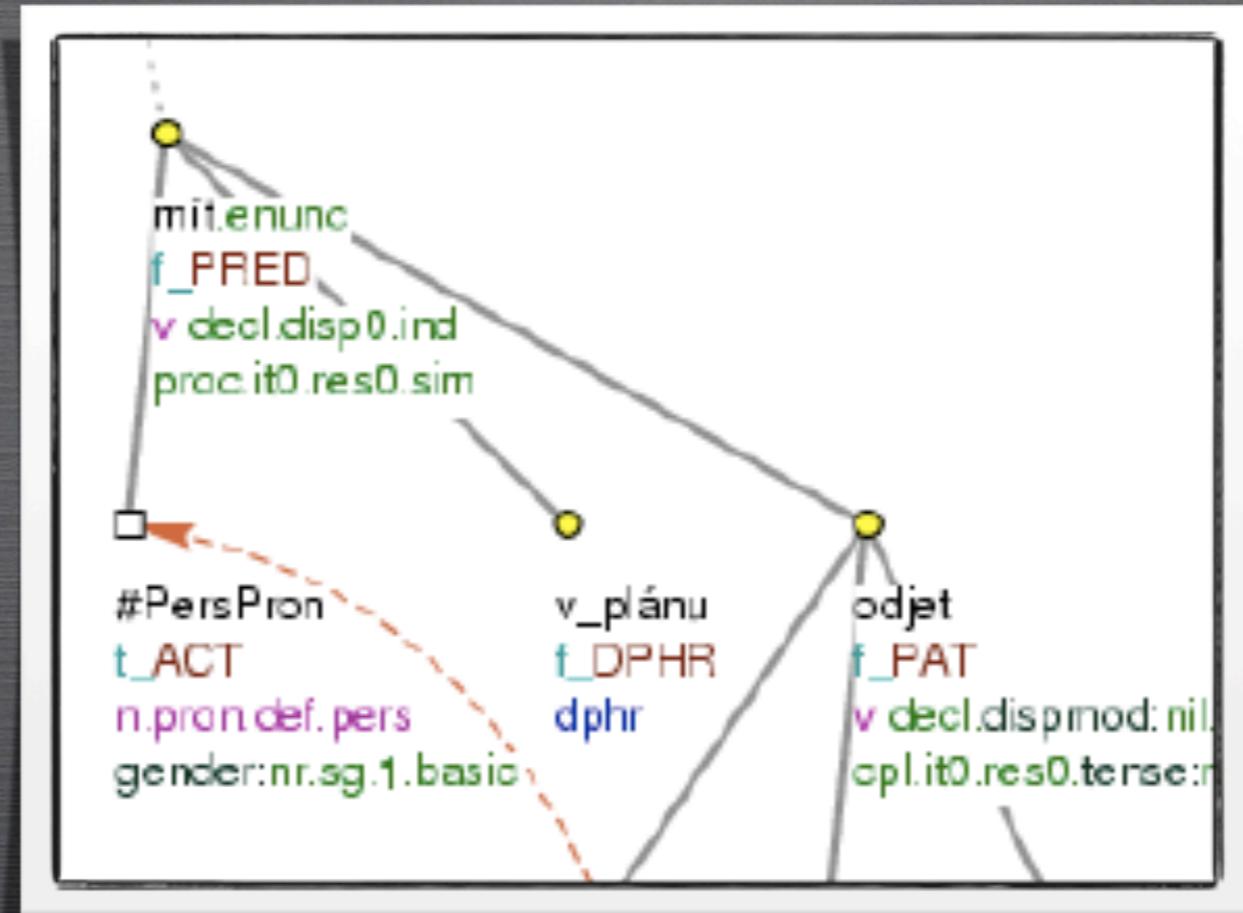
# DPHR

závislá část frazeologického spojení

978 výskytů v PDT 2.0

116 výskytů rozvíťých DPHR





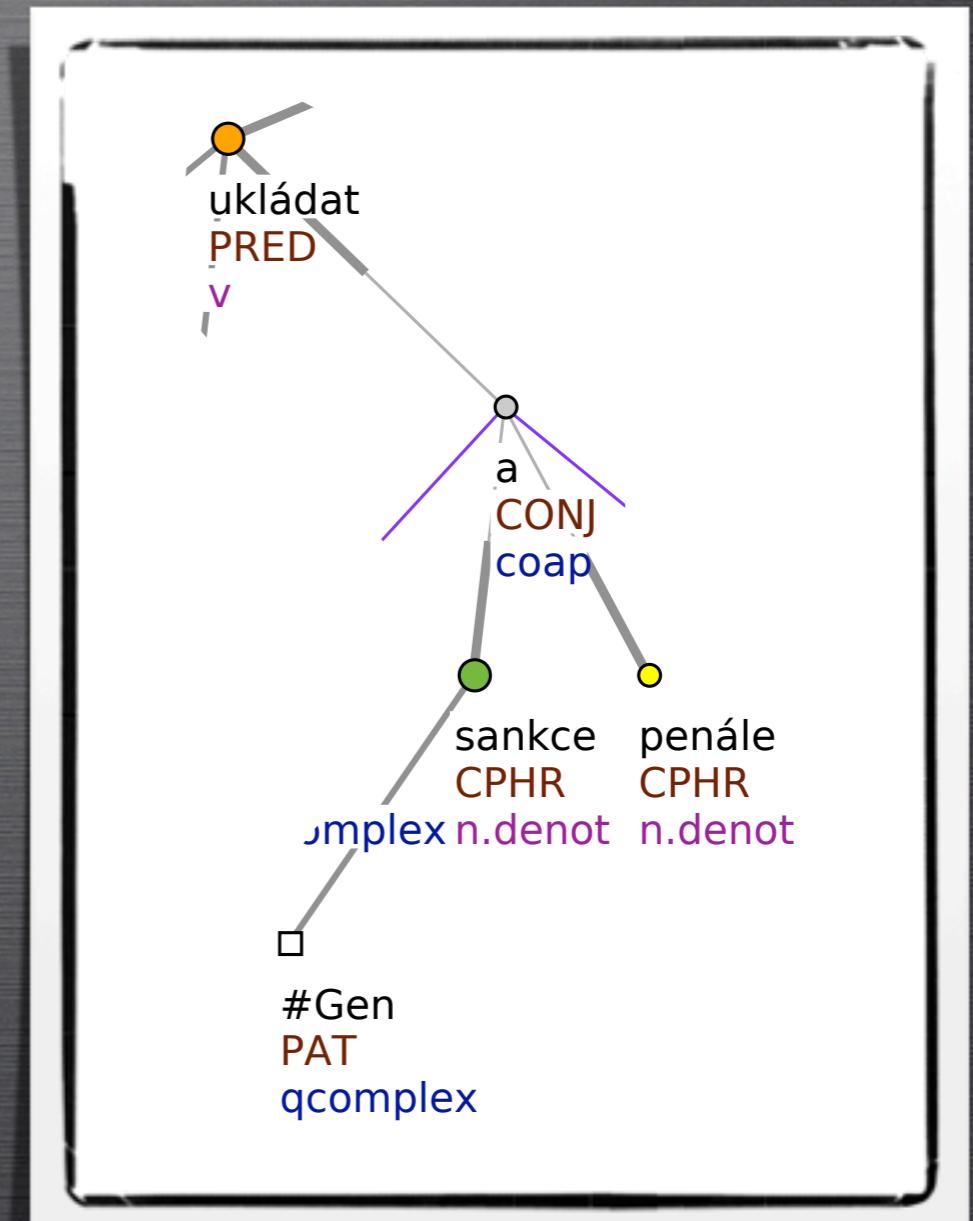
DPHR “mít v plánu” + odjet

# CPHR

jmenná část složeného predikátu či  
neslovesná část kvazimodálního  
slovesa

2221 výskytů v PDT 2.0

76 v koordinacích

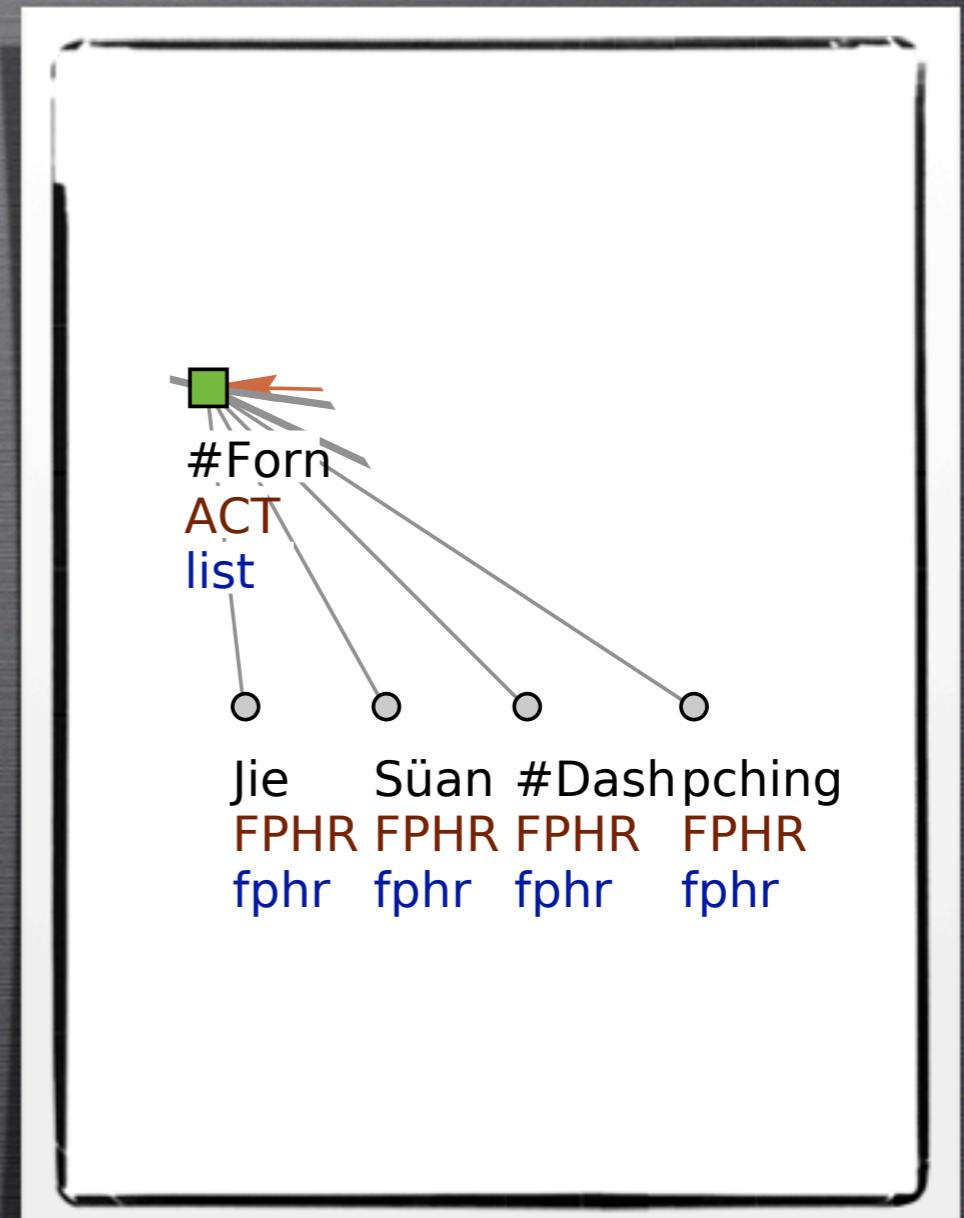


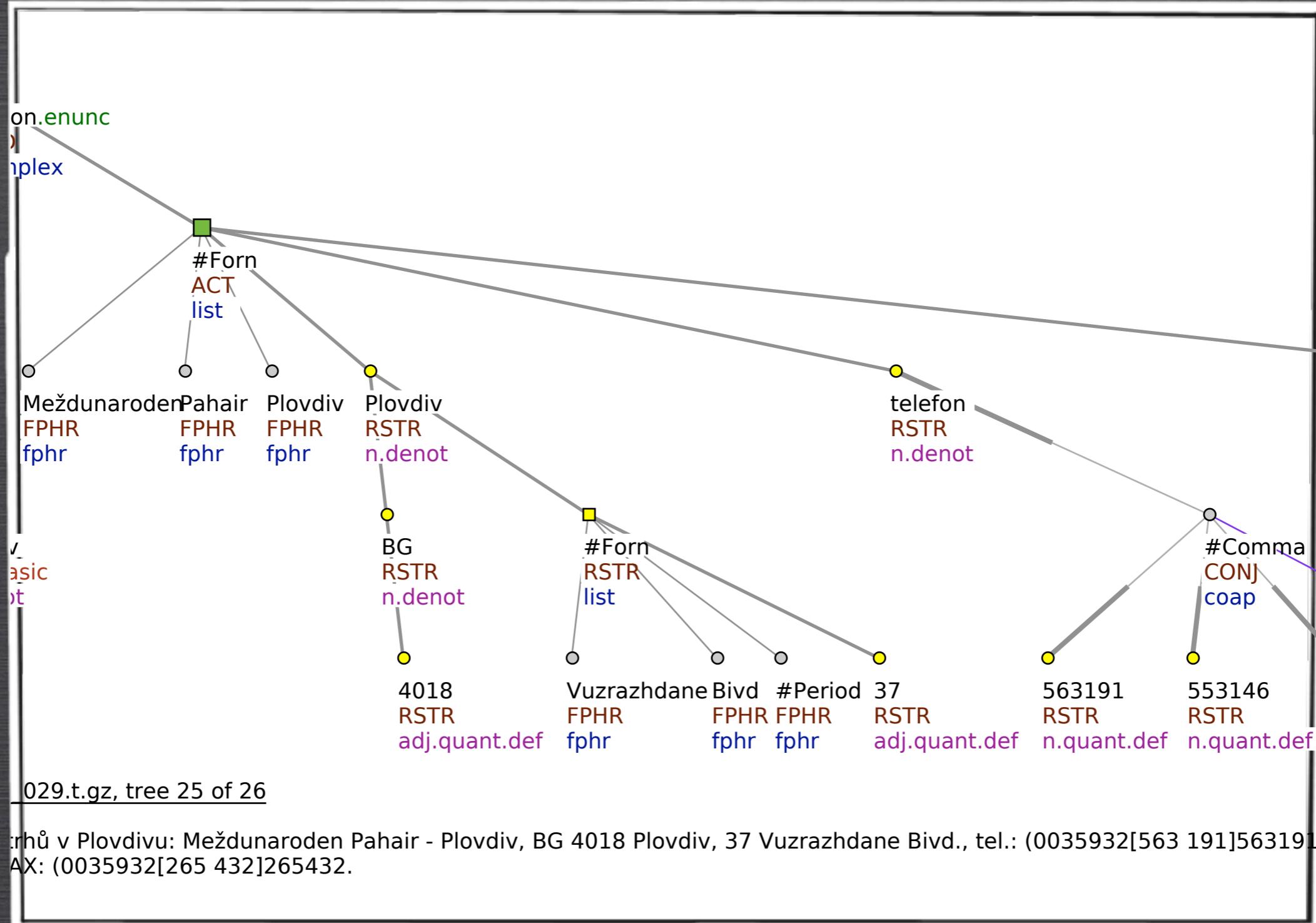
# FPHR

uzel reprezentující  
cizojazyčný výraz, který je součástí  
strukturně neanalyzovaného  
cizojazyčného textu

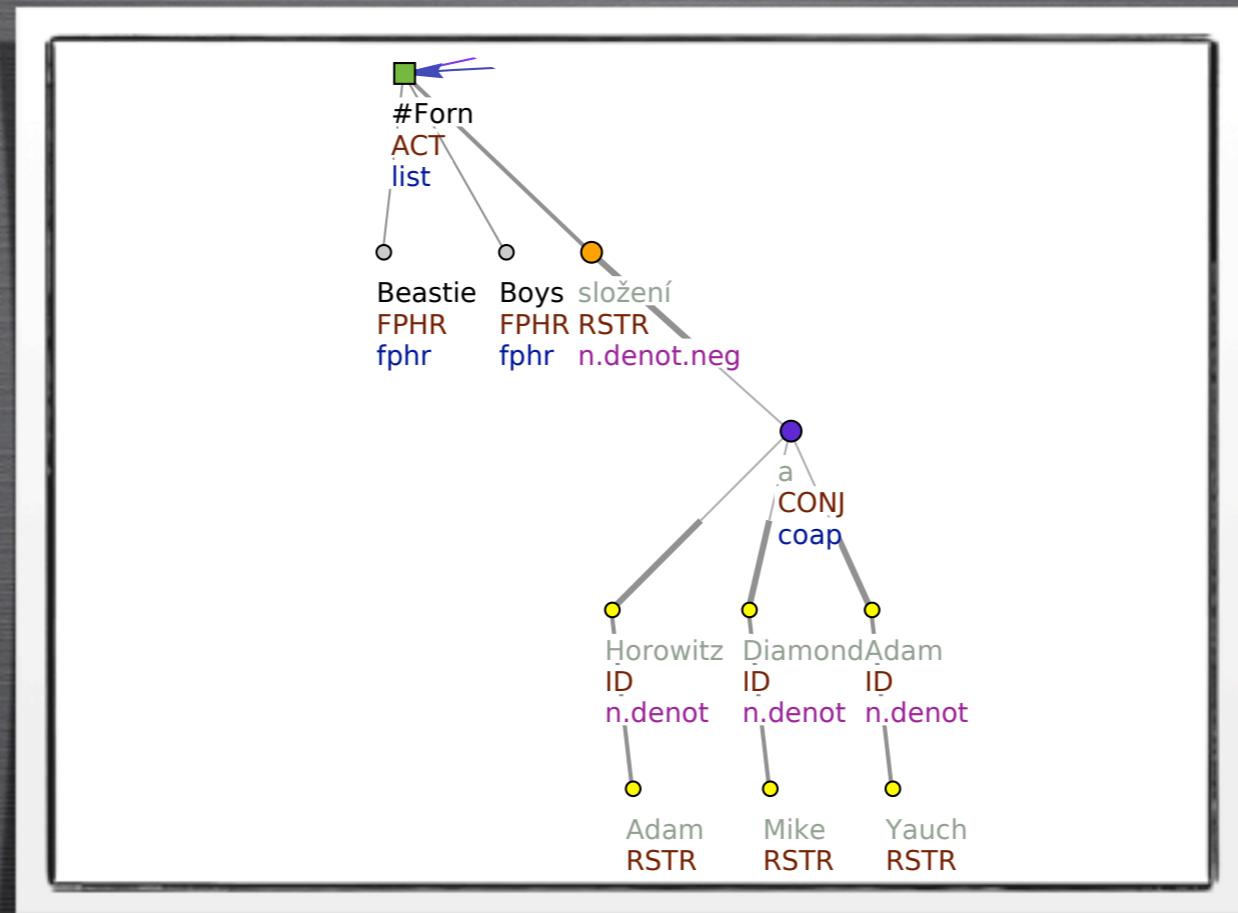
3620 výskytů v PDT 2.0

1310 konstrukcí  
(t-lemma = #Forn)

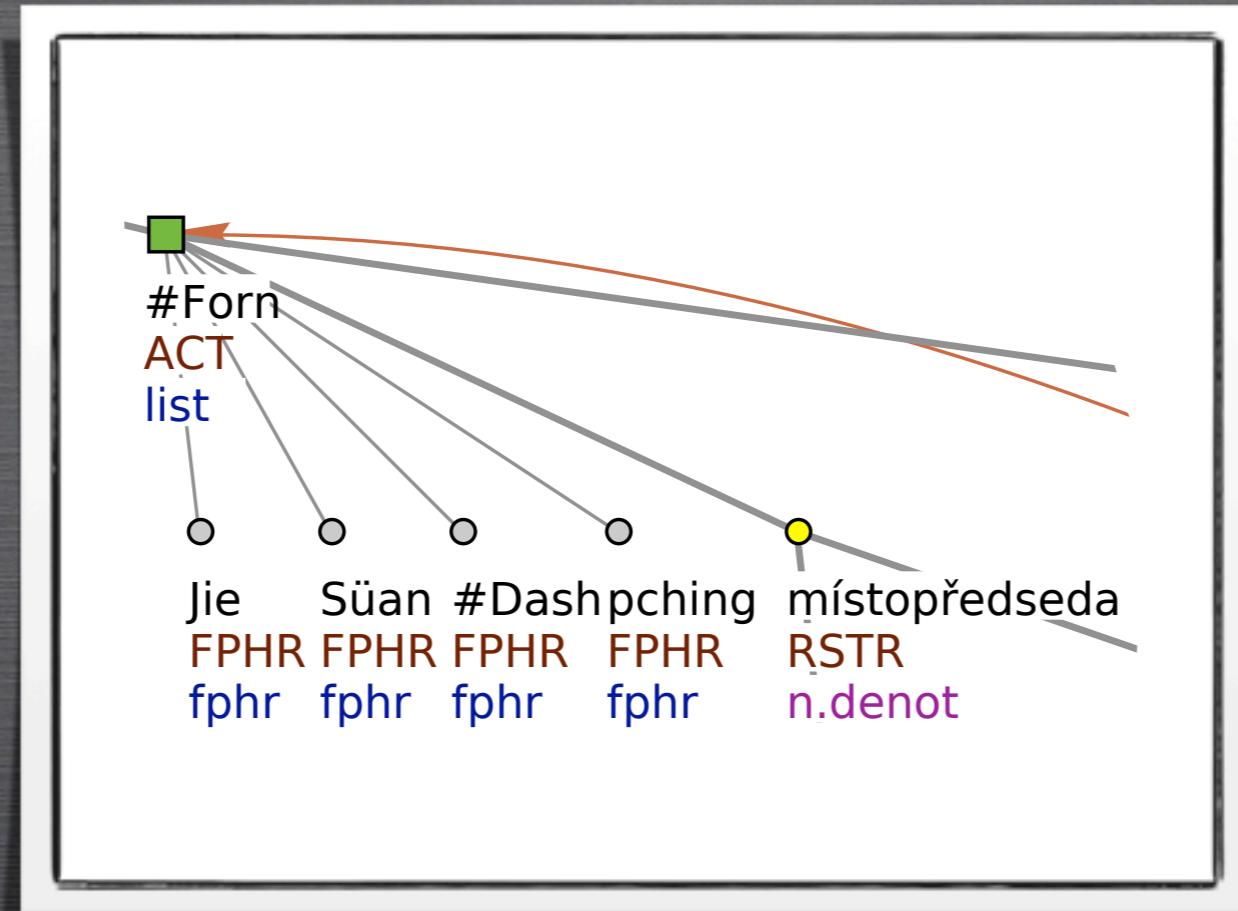




konflikt "adresy" a "cizí fráze"



Americká jména nejsou tolík cizí ...



... jako čínská

# LexemAnn

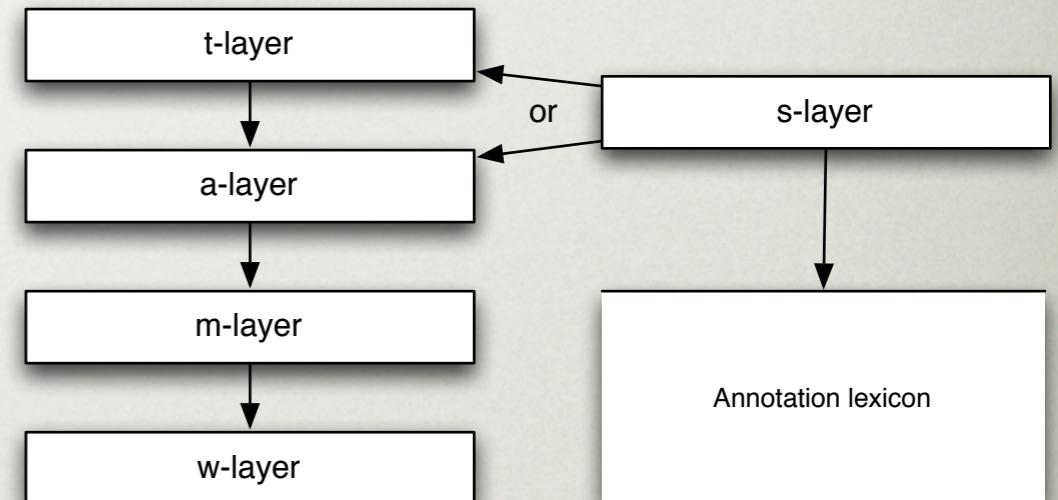
---

- = Lexico-semantic annotations
- = Annotations of lexias (or lexemes) ...
- Goal:
  - Identification of lexias
  - Over PDT 2.0:
    - t-node ≈ lexia

# Data files

---

- PDT 2.0 data; format: PML (XML based)
- PML:
  - stand-off annotation (4 layers)
  - addition of s-layer (“sense”)
    - not a deeper layer,  
refers to a-nodes or t-nodes
    - a list of pairs: *lexicon.ref => (t / a)-node.ref.list*

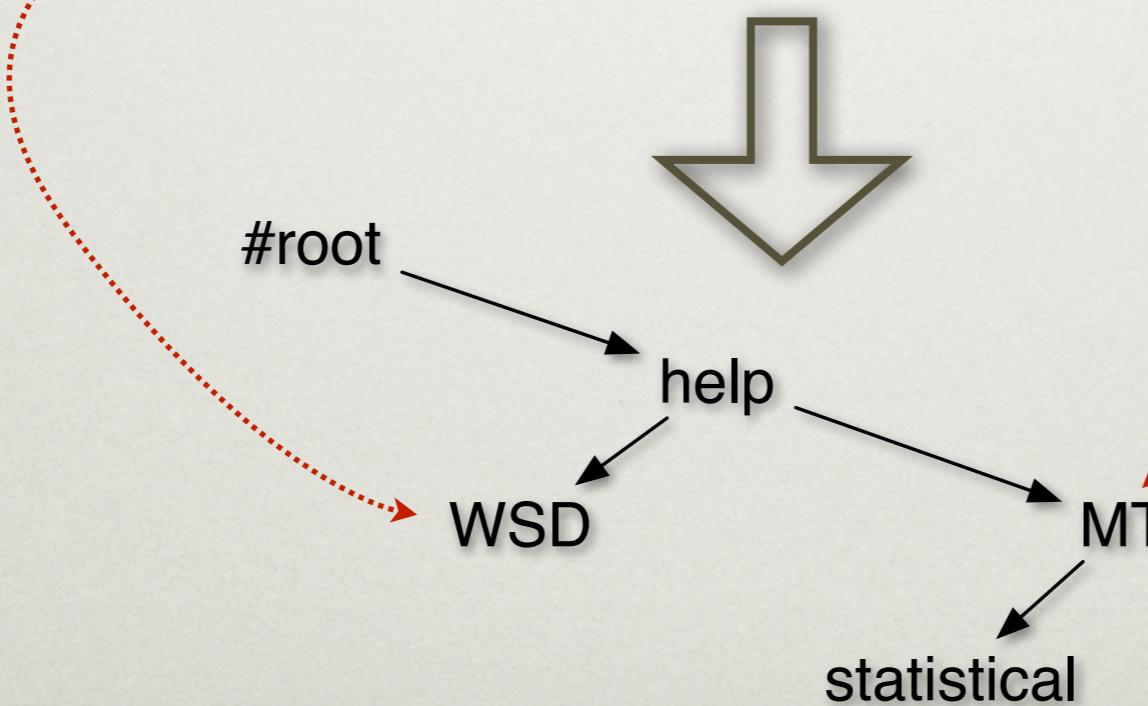
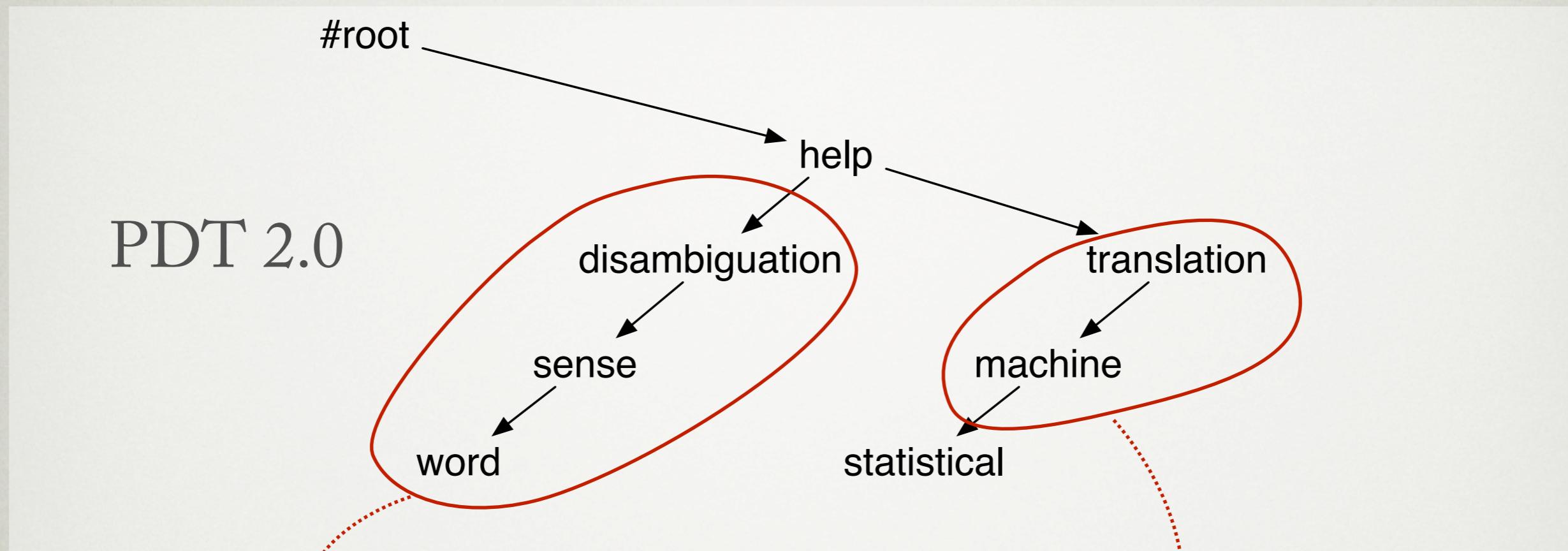


# Methodology

---

- Two rounds:
  1. multi-word lexemes and named entities
    - there is no point in assigning single-word senses to tokens inside these units (in data)
  2. remaining single-word lexemes

# Can word sense disambiguation help statistical machine translation?



# Annotation

---

- Goal:
  - Identification of multiword lexias
- Requirements:
  - A semantic lexicon — set of all possible meanings (tags) for each unit.
  - A method / procedure that assigns a semantic tag to each occurrence of a word.

# Multiword expressions

---

- “idiosyncratic interpretations that cross word boundaries” (Sag et al., 2002)
- a type of *lexias*
  - (a)idioms, phrasemes, LVC ...
  - (b)named entities
    - Are they really lexias (or semantemes)?
    - What is a *sense*? Is there a *concept* of UFAL?

# SemLex

---

- annotation lexicon
- MW lexias:
  - Czech WordNet (CWN),
  - Eurovoc,
  - Dictionary of Czech Phraseology and Idiomatics (SČFI)
- New lexias added by annotators - trees

# Czech WordNet

---

- Developed at The Masaryk University, Brno
- Originally in EuroWordNet 2, continuing development within the Balkanet project
- Mapped directly to the Princeton WordNet 2.0
- XML format
- 17,000 nouns; 2,000 verbs; 4,000 adjectives and adverbs

# Eurovoc

---

- multilingual, polythematic thesaurus focusing on the law and legislation of the European Union (EU)
- [eurovoc.europa.eu](http://eurovoc.europa.eu)
- 19,563 Czech lexemes (15,176 multi-word)
  - compare CWN: 23,000 lexemes (2,896 multi-word)
- microthesauri (politics, law, science, trade, ...)
- 17 official languages of EU

# SemAnn – an annotation tool

---

- plain text generated from t-layer
  - words remain linked to the t-nodes
    - tagging words → tagging t-nodes
- integrated browser and editor of SemLex
- efficient tagging - modal interface (compare Vi)
- Automatic pre-annotations of once annotated MWEs (they have a tree structure in SemLex)

Soubor Úpravy Debug

Festival Starý zákon v umění zahájí v září Job Petra Ebena

Varhanní kompozice Petra Ebena Job zahájí 3. září v Lichtenštejnském paláci v Praze mezinárodní festival nazvaný Starý zákon v umění. Rozsáhlá akce soustředí výstavy, divadelní a baletní představení i koncerty, které se uskuteční především v pražském Rudolfinu, u sv. Jakuba a v kostele sv. Šimona a Judy. Kromě České republiky se zatím k účasti přihlásily Norsko, Německo, Itálie, Rakousko, Švédsko, Izrael a Rusko.

Z hudebních kolektivů se v Praze představí například Norští sólisté, Jeruzalémský symfonický orchestr a Komorní orchestr z Halle. Pravděpodobně vystoupí i Česká filharmonie, Symfonický orchestr Českého rozhlasu, brněnská Státní filharmonie, Talichův komorní orchestr a soubory Archi Boemi a Virtuosi di Praga. Do programu festivalu přispěje Státní opera Praha, Divadlo na Vinohradech a balet Národního divadla z Brna. V rámci festivalu se bude konat reprezentativní izraelská archeologická výstava nazvaná City of David (Město Davidovo), výstavy výtvarného umění i expozice studentských prací pražské Akademie výtvarných umění. Další akce připraví pražské židovské muzeum, Památník národního písemnictví, Národní galerie a Národní muzeum.

Heslo zobrazeno.

Značky

Obecné Pojmenované entity  Automatická anotace

Ukázat Odstranit Jméno Instituce Místo Objekt Adresa Čas Biblio Foreign X

SemLex

Označkovat Nové heslo Ulož heslo  A=a  Hledat P N

ID: 0000003772 Source: Eurovoc POS: N Základní tvar: výtvarná umění  Lematizovaný tvar: výtvarný umění

Příklad: Synonyma:

Glosa:  Změněno:

# Inter-annotator Agreement

---

Table I. Annotated instances  
of significant types of MWEs

type of MWE	A	B
SemLex entries	8,447	8,312
- different items	3,844	4,089
Named Entities	8,435	8,903
- person/animal	2,797	2,811
- institution	1,702	2,047
- number	1,343	1,053
- object	1,129	888

# Kappa

---

$$\kappa = \frac{A_o - A_e}{1 - A_e} = 0.68$$

$$\kappa_w^U = \frac{A_o - A_e}{\widehat{U} - A_e} = \frac{0.154 - 0.046}{0.213 - 0.046} \doteq 0.644.$$