Annotation of the Topic-Focus Articulation in the Prague Dependency Treebank

Eva Hajičová

Overview

1. Linguistic motivation of TFA annotation:
   i. Basic notions
   ii. Why TFA should be annotated in the TGTS's: semantic relevance of TFA
2. TGTS attribute TFA and its values
3. Examples
4. Testing linguistic hypotheses on a deep layer of corpus annotation

Basic notions of TFA

- Information structure of the sentence
  - Topic-focus articulation
    - Topic, theme, …
    - Focus, rheme, …
  - based on *given* x *new*, but not identical to this cognitive dichotomy:
    - John and Mary entered the dining-room. They first went to the window …
    - Mary Called Jim a Republican. Then he insulted HER.
    - Mary called Jim a republican. Then he INSULTED her.

Semantic relevance of TFA

- Everybody in this room knows two languages.
  Two languages are known by everybody in this room.
- Many men read few books.
  Few books are read by many men
- Smoke in the hallway!
  In the hallway, you smoke.
- Staff behind the COUNTER.
  STAFF behind the counter.
- Carry DOGS.
  CARRY dogs. Dogs must be carried.

Topic-focus articulation in PDT

one attribute (TFA – topic-focus articulation)
with values concerning the *contextual boundness* of the nodes

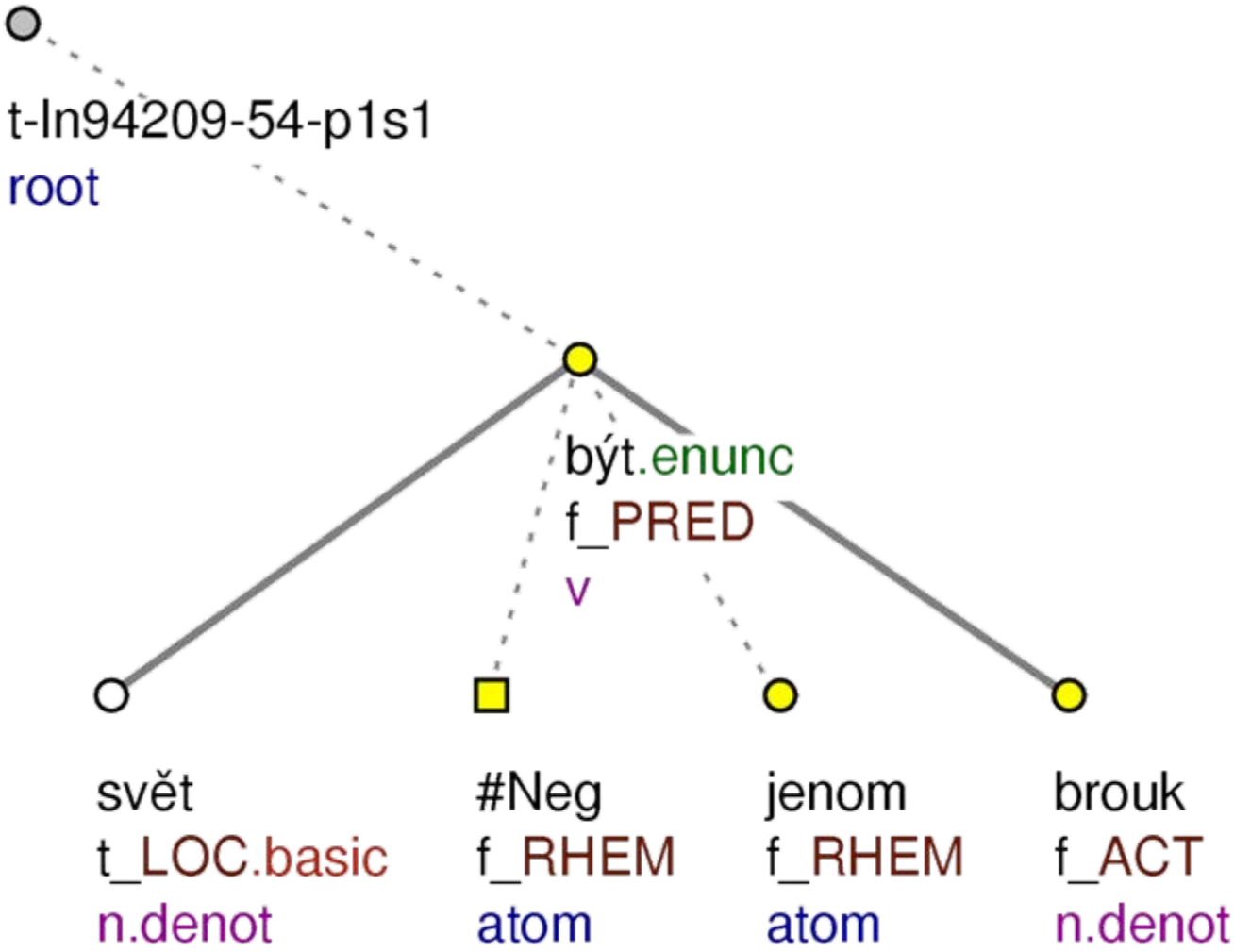
three values in the TFA attribute:

t – contextually bound non-contrastive

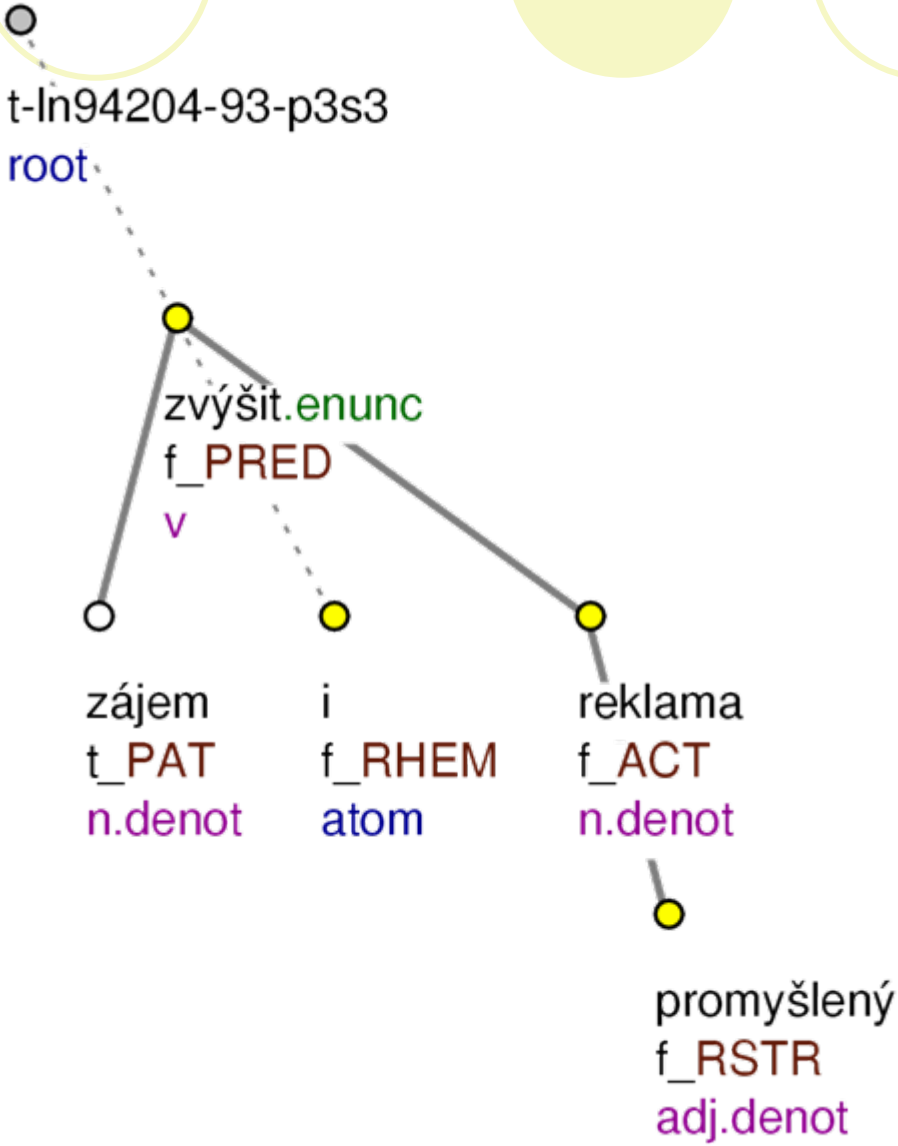
c – contextually bound contrastive

f – contextually non-bound

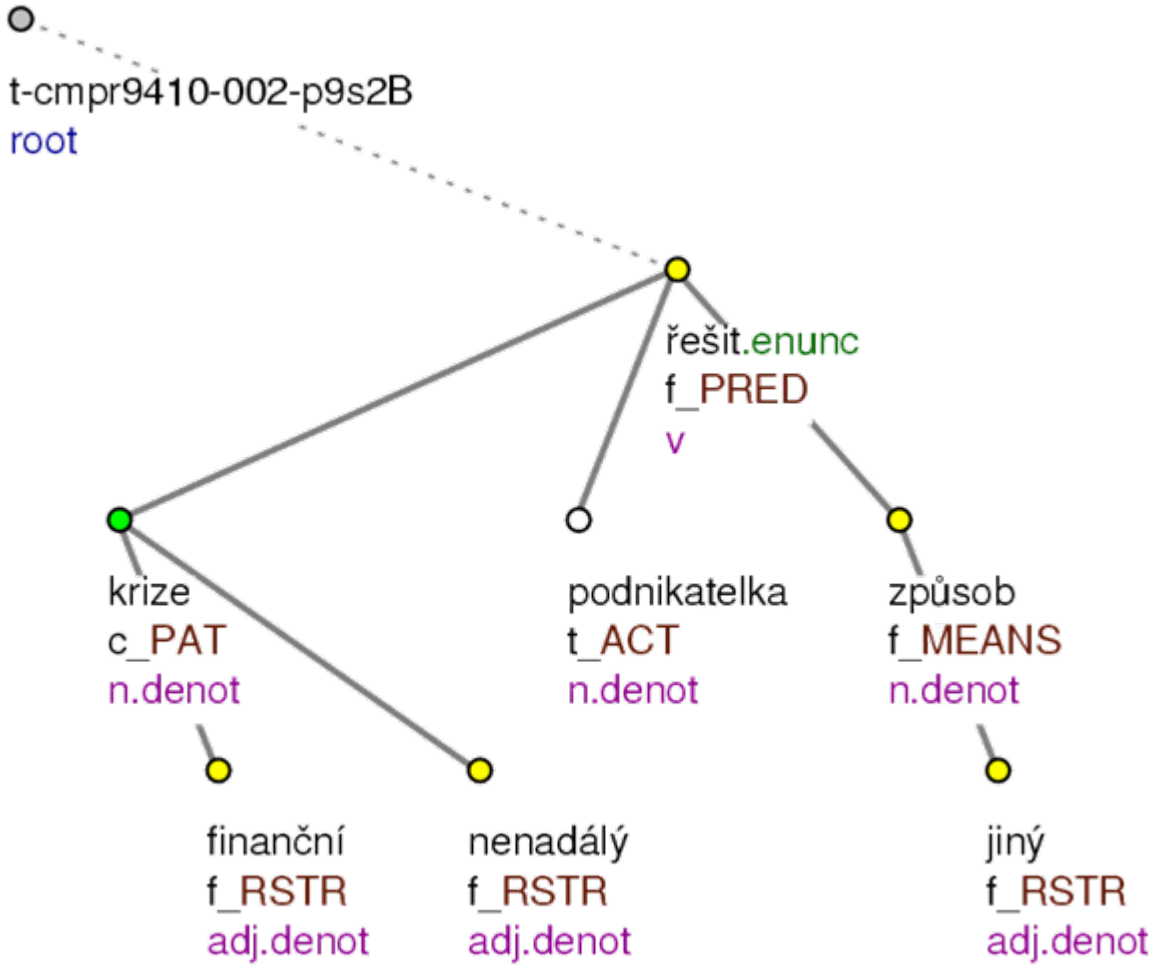
Example *Na světě nejsou jenom brouci. [In the-world there-are-not only beetles.]*



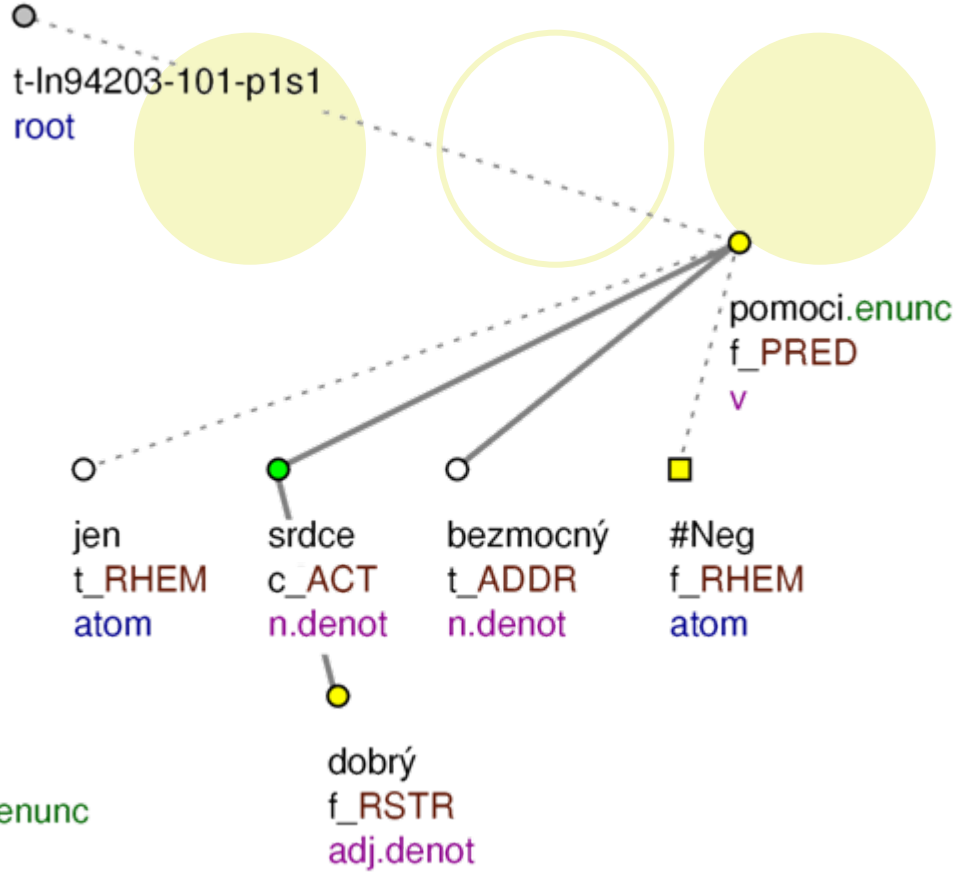
Example *Zájem zvýšila i promyšlená reklama. [The-interest_Acc raised also the-sophisticated campaign.]*

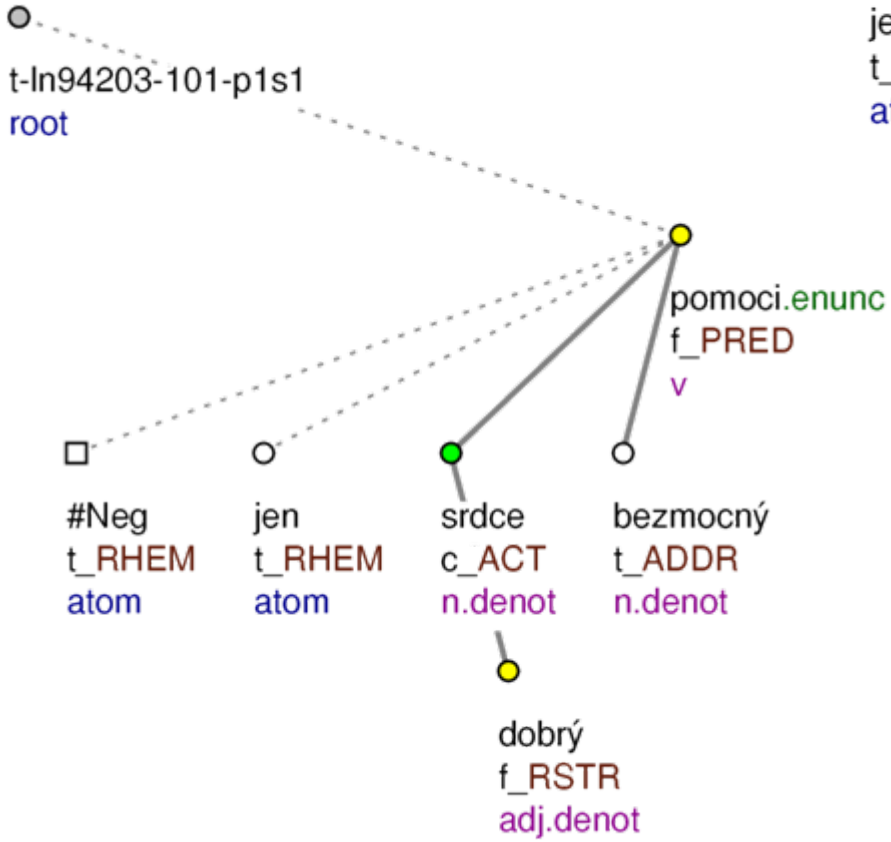


Example *Nenadálou finanční krizi podnikatelka řešila jiným způsobem. [The-sudden financial crisis_Acc the-entrepreneur_Nom solved by other means.]*



Example *Jen dobré srdce bezmocným nepomůže. [Only good heart the-helpless_Dat will-not-help.]*





Testing linguistic hypotheses

- Corpus annotation is not a self-contained task.
- A necessary condition for a usable annotated corpus: based on a sound linguistic theory.
- PDT: linguistic basis: Functional Generative Description.
- One of the important uses of corpus: test for linguistic theories.

Hypothesis A1: Division into T and F based on boundness

Hypothesis A1:

- **the global division of the sentence into its TOPIC (what the sentence is about) and its FOCUS (what is said about the topic) can be made on the basis of boundness**

  Sgall (1979; see also Sgall et al. 1986 216f),

  original algorithm implemented and tested on the whole of PDT; the results reported in Hajičová, Havelka and Veselá (2005)

Hypothesis A1 (cont.)

a) main verb=f → belongs Focus (F); else, → T

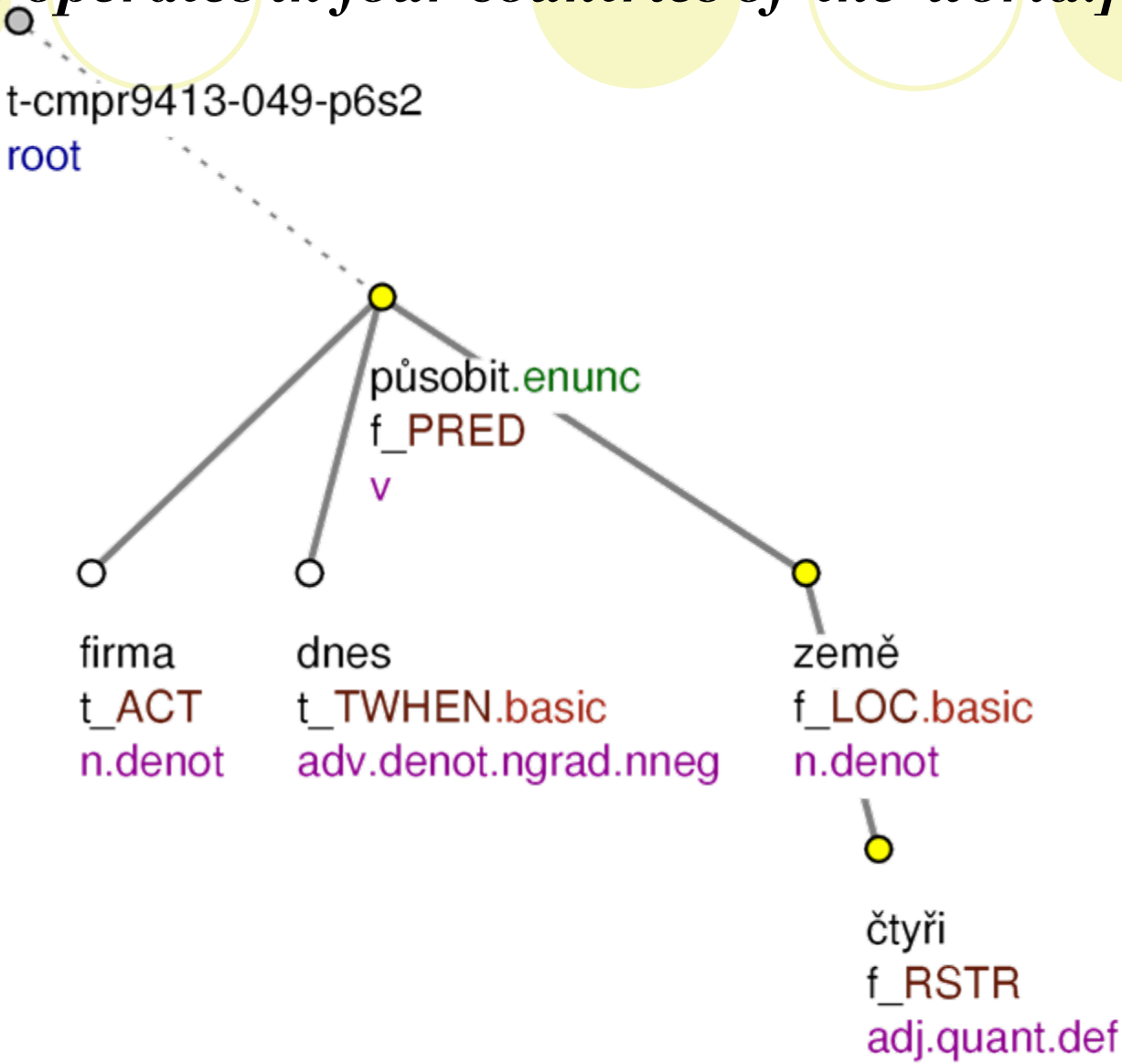
b) all the nodes directly dependent on the main verb and carrying t → T, together with all nodes depending on them

c) all the nodes directly dependent on the main verb and carrying f →F, together with all nodes depending on them

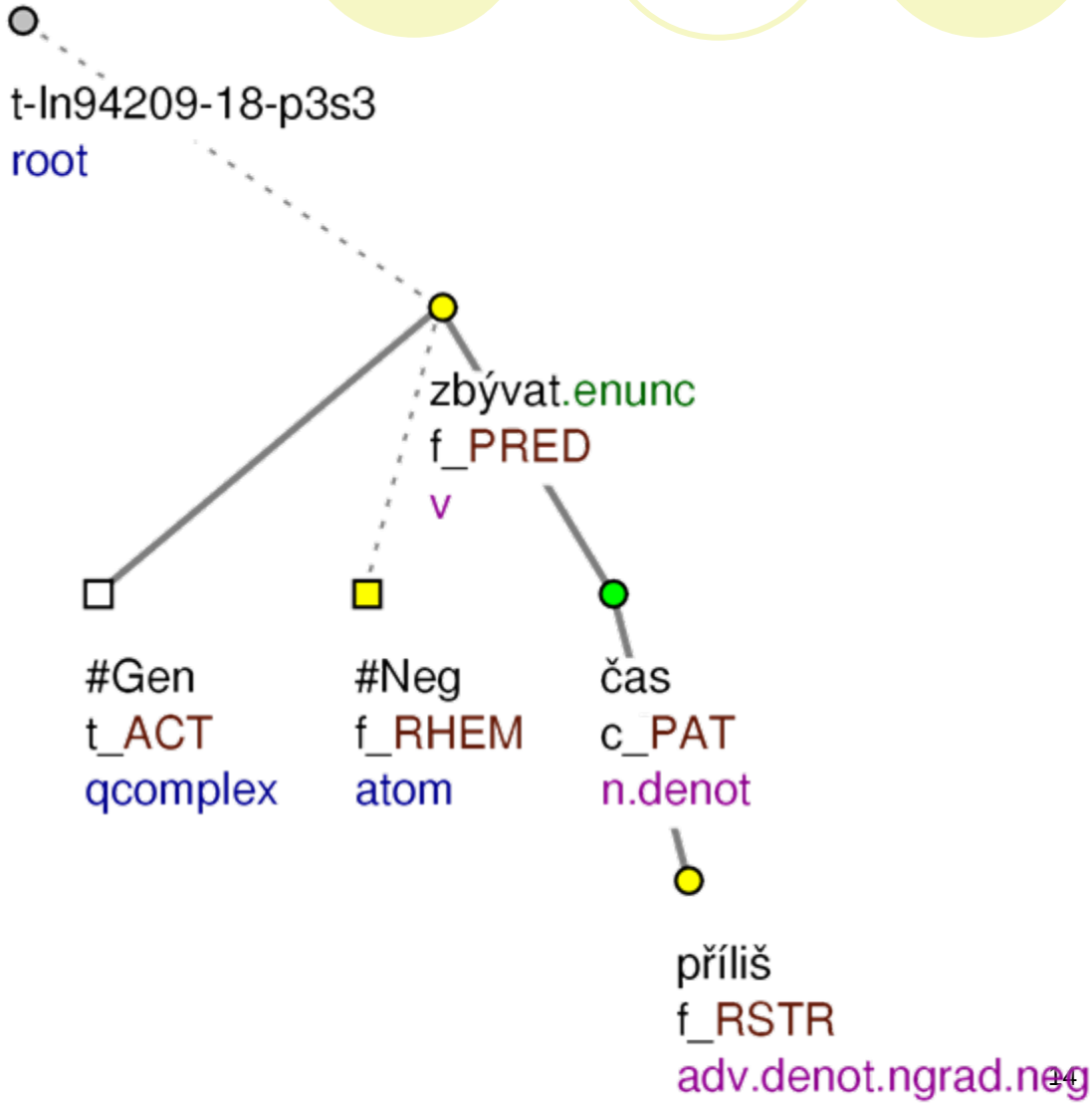
d) main verb = t & all nodes directly depending on the main verb = t: follow the rightmost edge leading from the main verb to the first node(s) on this path carrying the value f → this/these node(s) and all the nodes depending on it/them = F

Example *Firma dnes působí ve čtyřech zemích světa.*

[The-firm now operates in four countries of-the-world.]



Example *Času příliš nezbývá. [Time_Gen too-much does-not-remain.]*



Results of the implementation

F: V + subtrees	85,7%
F: right-attached subtrees of V.t	8,58%
Quasi-focus	4,41%
F interrupted by c-node	0,06%
Ambiguous partition	1,14%
No focus indentified	0,11%

Results

- **in Czech**: the **boundary** between Topic and Focus can be determined in principle on the basis of the consideration of the **status of the main predicate** and its direct dependents.
- TFA annotation leads to satisfactory results in cases of rather **complicated "real" sentences** in the corpus.

Certain modification of the annotation procedure necessary, but the material gathered and analyzed in this way may be further used for the study of several aspects of the **discourse patterning.**

Hypothesis A2: The so-called systemic ordering

Hypothesis A2:

In the focus part of the sentence the complementations of the verb (be they arguments or adjuncts) follow a certain canonical order (not necessarily the same for all languages).

tested with a series of psycholinguistic experiments (with speakers of Czech, German and English) but PDT offers a richer and more consistent material → work in progress (Lešnerová)

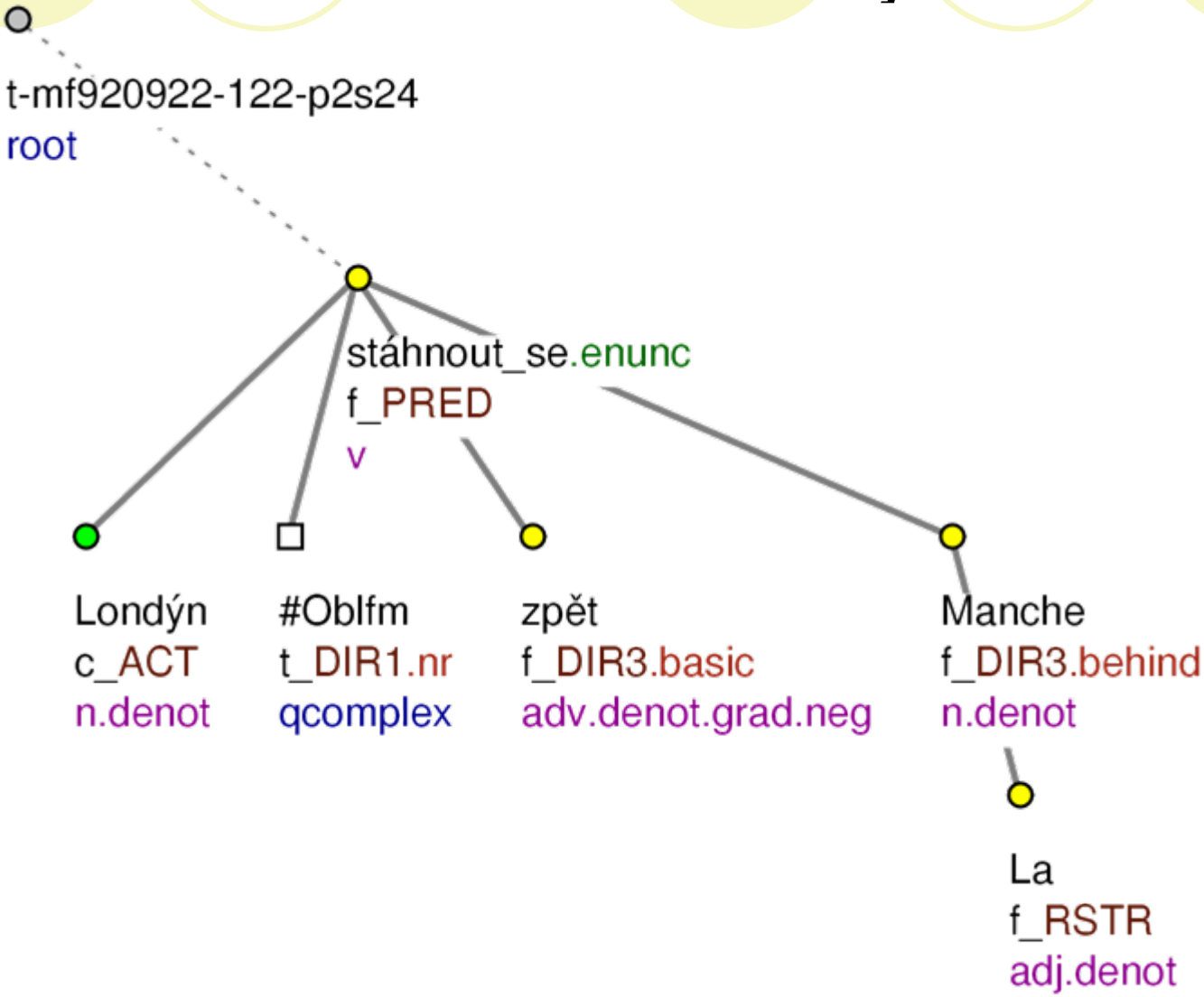
Hypothesis A2: (cont.)

Tested on PDT

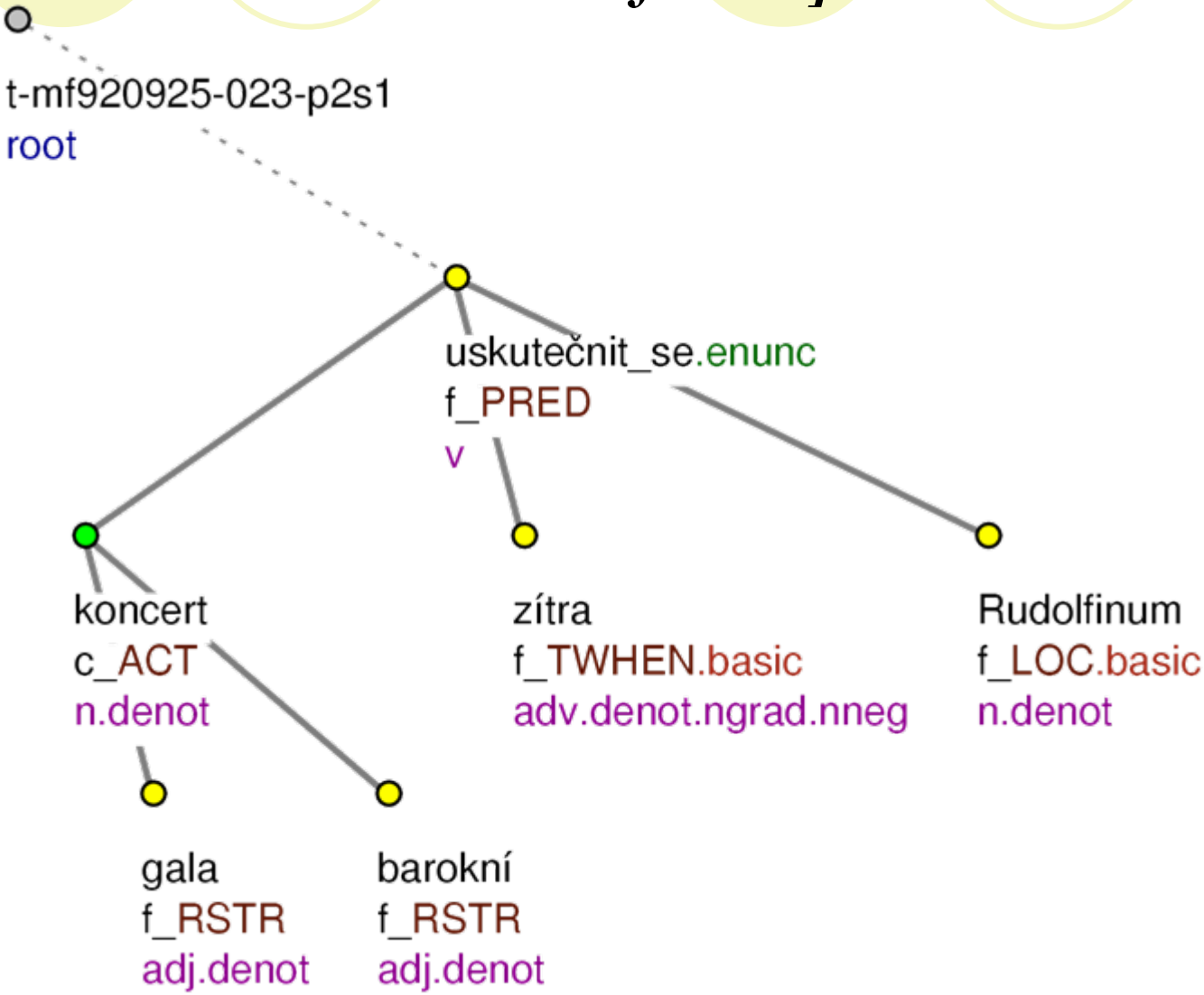
a) the **F** of the sentence identified (see A1)
b) in TGTS: the **surface order in F** preserved
c) **systemic ordering** hypothetically stated

→these pieces of information used to compare the order of the complementations in the actual sentence and the assumed order according to the scale of systemic ordering

Example *Londýn se stáhl zpět za La Manche. [London Refl. Withdrew back behind La Manche.]*



Example *Barokní gala koncert se uskuteční zítra v Rudolfinu. [A-baroque gala concert Refl. Will-take-place tomorrow at Rudolfinum.]*



Conclusions

- → importance of the deep-layer corpus annotation for the **study** of most various language phenomena
- → the tectogrammatical layer of annotation brings about an indispensable source of information for **testing** any linguistic **theory** and any **grammar build-up**