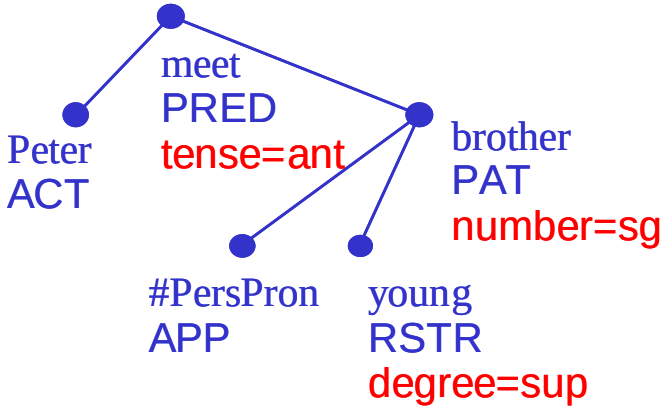
Grammatemes and Coreference in the PDT 2.0

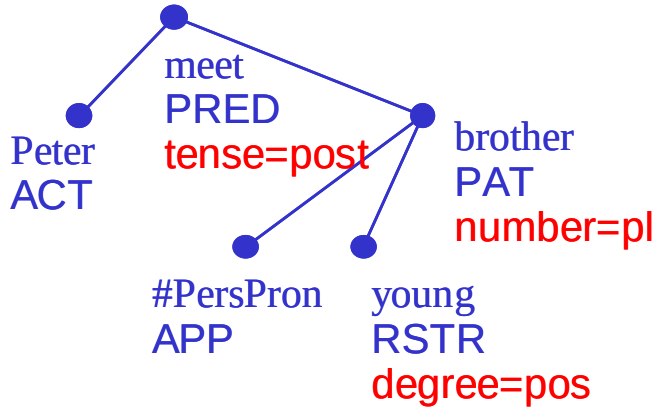
Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics
Charles University in Prague

What is a "grammateme"? (1)

Peter met her youngest brother.



Peter will meet her young brothers.



- the same t-lemmas, the same tree topology, the same functors, but the original sentences are obviously not synonymous and must be distinguished at the t-layer (must obtain different t-trees) !
- the difference is in grammatemes ~ t-node attribute-value pairs representing morphological meanings (semantically indispensable morphological categories)
- e.g. number for nouns, tense for verbs, degree for adjectives, deontic/verb/sentence modality ...

What is a "grammateme"? (2)

- grammatemes are not just straightforward counterparts of surface morphological categories (as stored in m-layer tags) !
- some morphological categories are only imposed by grammar and thus are not semantically relevant
  - gender, number or case of an adjective in a noun group come from agreement with the noun (e.g. in Czech or German), not from semantics
  - similarly, person is not a grammateme of verbs, as it is only induced by subject-verb agreement

What is a "grammateme"? (3)

- on the surface, grammatemes can be expressed both inflectionally and analytically
- info about grammatemes can be distributed over more than one m-layer token
  - comparative of adjectives in English (*more interesting*)
  - future tense of imperfectives in Czech (*budu chodit…/ I will go…*)

Complete list of grammateme attributes used in PDT 2.0

1. **gram/number** - number of semantic nouns
2. **gram/gender** - gender of semantic nouns
3. **gram/person** - person of pronominal semantic nouns
4. **gram/politeness** -basic vs. polite/esteemed form, relevant for pronominal semantic nouns
5. **gram/indeftype** (type of indefiniteness of pro-forms)
6. **gram/numertype** (type of numeric expression)
7. **gram/negation** - negation of semantic nouns, adjectives, and adverbs (not of verbs)
8. **gram/degcmp** - degree of comparison of semantic adjectives and adverbs
9. **gram/tense** - tense of verbs
10. **gram/aspect** - aspect of verbs
11. **gram/verbmod** - basic verb modality (indicative, imperative, conditional)
12. **gram/deontmod** - deontic modality expressed by modal verbs
13. **gram/dispmod** - dispositional modality (specific for Czech)
14. **gram/resultative** - resultativeness of verbs
15. **gram/iterativeness** - iterativeness of verbs
16. **sentmod** - sentence modality (enunciative, exclamative, desiderative, imperative, interrogative)

PDT 2.0

Grammateme number

- values:
  - sg - singular
  - pl - plural
  - nr - not recognized

- m-layer/t-layer asymmetry:
  - pluralia tantum: *jedny dveře/ dvoje dveře* (one door, two doors) - only the plural form exists at the m-layer, but sg/pl should be disambiguated at the t-layer
  - polite form: *"Viděl jste to, Petře?"* (Did you see it, Petr?) - complex verb form containing an auxiliary verb in plural at the m-layer, but at the t-layer the grammateme number (filled in the reconstructed #PersPron node) is equal to singular

PDT 2.0

Grammateme tense

- relative tense of verbs (with respect to the tense of the governing clause)
- values:
  - sim - simultaneous
  - ant - anterior
  - post - posterior
  - nil - absent (with infinitives)
  - nr - not recognized
- m-layer means for expressing tense=post in Czech:
  - inflection with perfectives (*uvařím* - I will cook)
  - auxiliary verb *být* with imperfectives (*budu zpívat* - I will sing)
  - prefix *po-*/*pů-* with a limited set of verbs (*pojedu* - I will go)

PDT 2.0

Grammateme indeftype (I)

- pro-form - a word used to replace or substitute other words, phrases, clauses...
- pronouns (pro-nouns), pro-adjectives, pro-numerals, pro-adverbs
- there are many semantically significant analogies present in the pro-forms systems, but usually not explicitly distinguished in the POS tag sets
- example of such parallelism:
  - nobody/never/nowhere... vs. everybody/always/everywhere...
- grammateme indeftype (type of indefiniteness) dedicated for all indefinite pro-forms
- to capture the parallelisms, each group of pro-forms is represented with t_lemma identical with the relative form: *někde->kde (nowhere->where), kdokoli->kdo (whoever->who), nikdy->kdy (never->when)*

Grammateme indeftype (II)

PDT 2.0

t-lemma:	*kdo*	*co*	*který*	*jaký*
value of the grammateme **indeftype**:				
relat	*kdo*	*co*	*který, jenž*	*jaký*
indef1	*někdo*	*něco*	*některý*	*nějaký*
indef2	*kdosi, kdos*	*cosi, cos*	*kterýsi*	*jakýsi*
indef3	*kdokoli(v)*	*cokoli(v)…*	*kterýkoli(v)*	*jakýkoli(v)*
indef4	*ledakdo, leckdo…*	*ledaco, lecco…*	*leckterý, ledakterý*	*lecjaký, ledajaký*
indef5	*kdekdo*	*kdeco*	*kdekterý*	*kdejaký*
indef6	*málokdo, kdovíkdo…*	*máloco…*	*málokterý…*	*všelijaký…*
inter	*kdo, kdopak…*	*co, copak…*	*který, kterýpak*	*jaký, jakýpak*
negat	*nikdo*	*nic*	*žádný*	*nijaký*
total1	*všechen*	*všechen, všechno, vše*	–	–
total2	–	–	*každý*	–

PDT 2.0

Grammateme indeftype (III)

- indefinite, negative, interrogative, and relative pronouns and other pro-forms are unproductive classes with (at least to a certain extent) transparent derivational relations also in other languages

- preliminary sketch of several English and German pronouns classified by indeftype

	English	English	German	German
Lemma	*who*	*what*	*wer*	*was*
indefype:				
relat	who	what	wer	was
indef1	someone	something	jemand	etwas
indef2	-	-	irgendjemand	irgendetwas
indef3	whoever	whatever	-	-
inter	who	what	wer	was
negat	nobody	nothing	niemand	nichts
total1	all	everything	alle	alles
total2	each	each	jeder	jedes

PDT 2.0

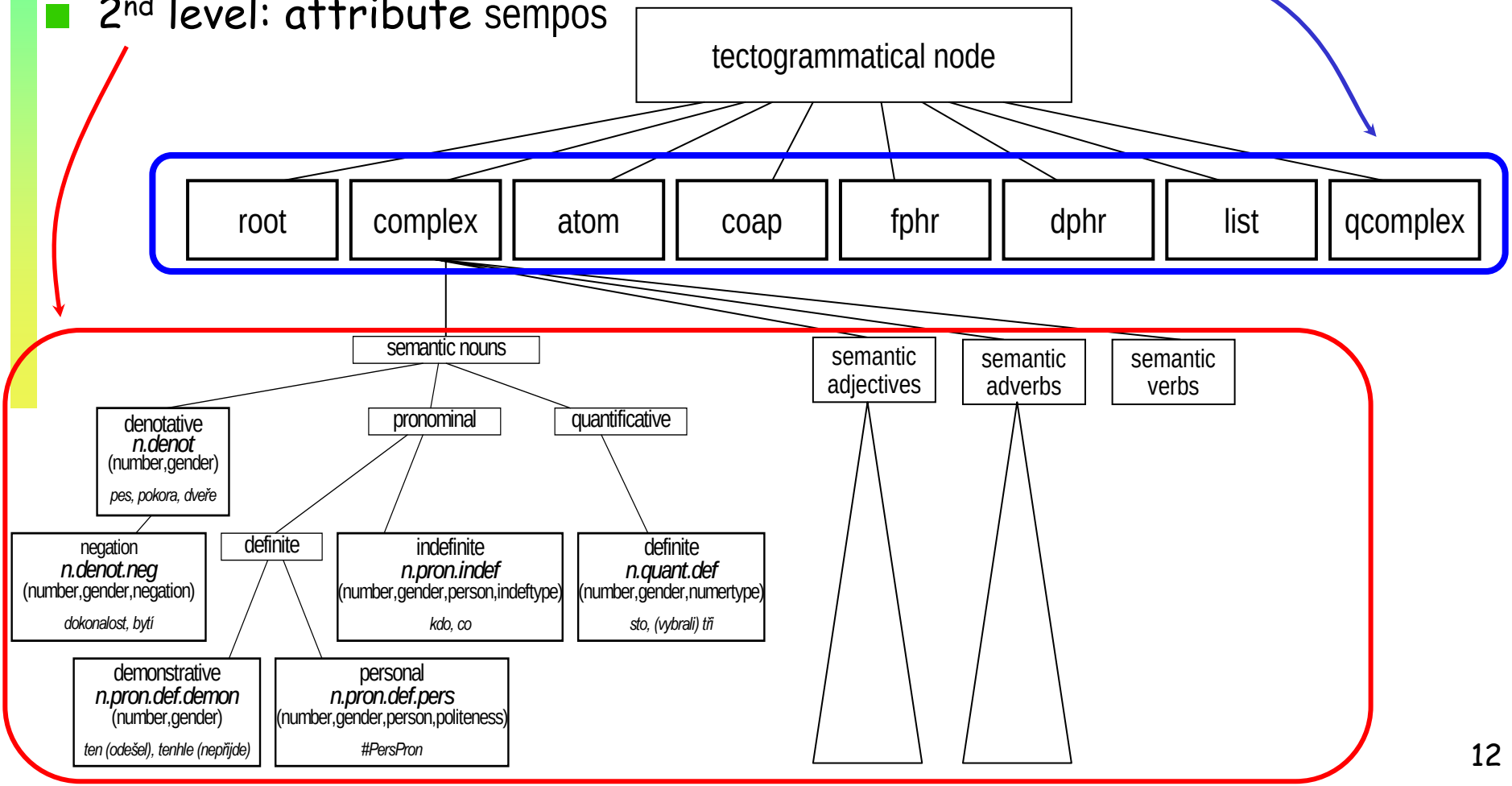
Typing of t-nodes

- unlike t_lemmas and functors, grammateme attributes are not relevant for all t-nodes
  - obviously, no tense for *dog*, no degree of comparison for *(he) waits*, etc.
- question: how to formally declare presence/absence of a certain grammateme in a certain t-node ? → the need for node typing
- our solution: two-level hierarchy of node types
  - 1st level: 8 coarse-grained types of nodes
  - 2nd level: 19 more specific subtypes, corresponding to detailed semantic parts of speech

PDT 2.0

Two-level hierarchy of t-node types

- 1st level: attribute nodetype
- 2nd level: attribute sempos



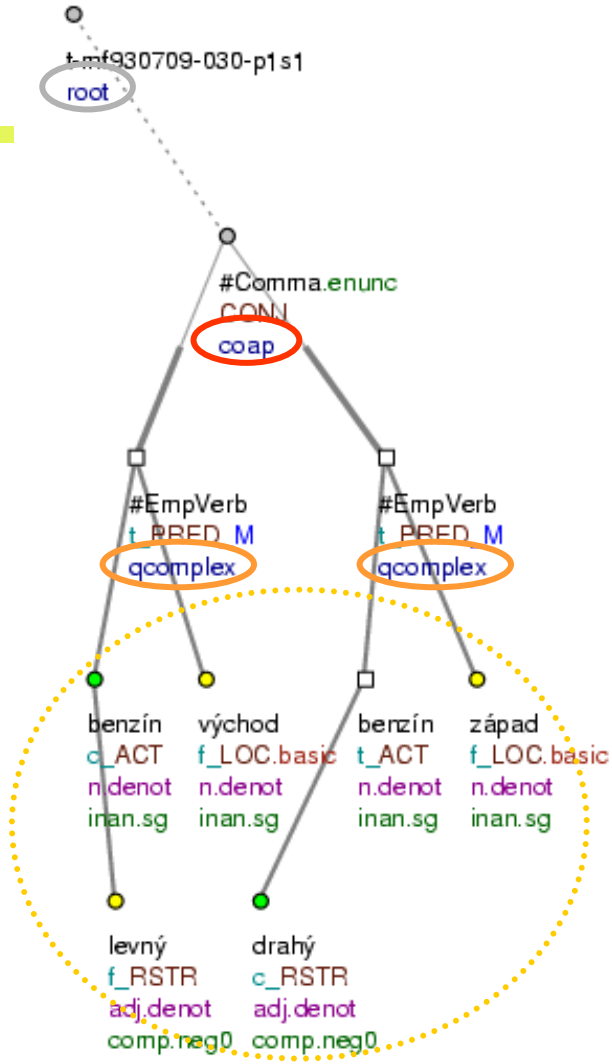
PDT 2.0

First level of the hierarchy: attribute nodetype

- 8 nodetype values:

  root | complex | qcomplex | list | atom | coap | dphr | fphr

- fully automatic annotation - use of
  - the tree structure → root
  - t-attributes
    - t-lemma → qcomplex | list
    - functor → atom | coap | dphr | fphr
  - otherwise → complex

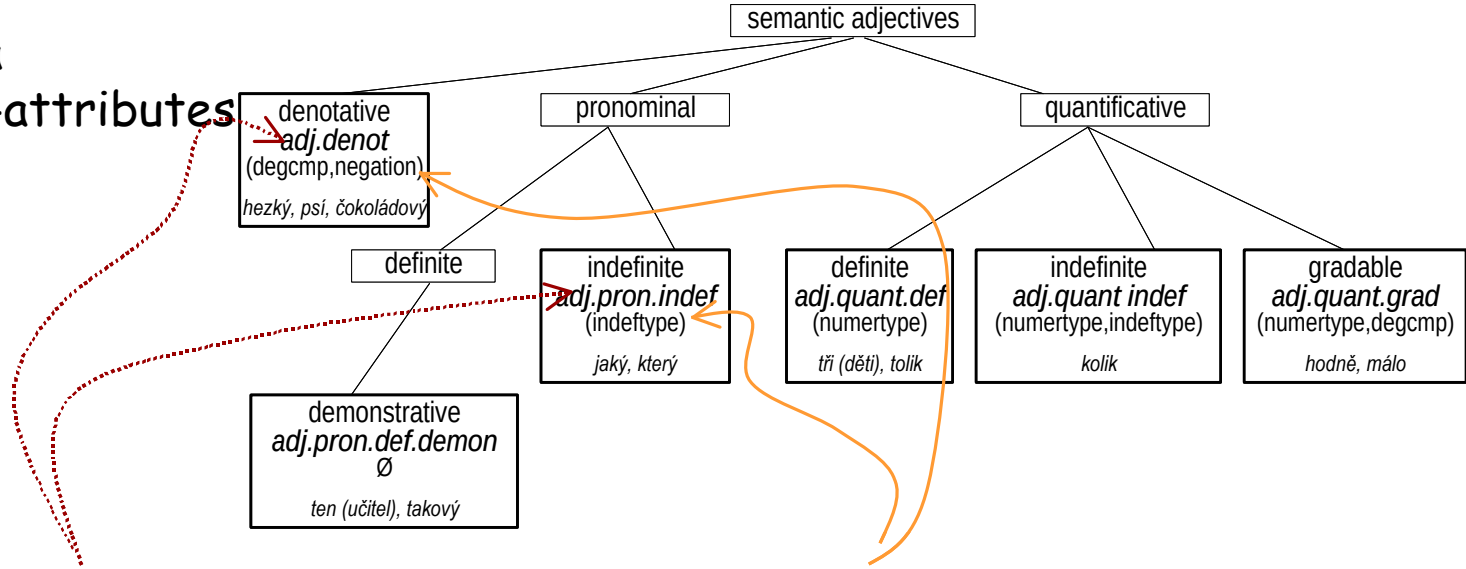


Levnější benzín na Východě, dražší na Západě
Cheaper gasoline in the East, more expensive one in the West

PDT 2.0

Second level of the hierarchy: attribute sempos

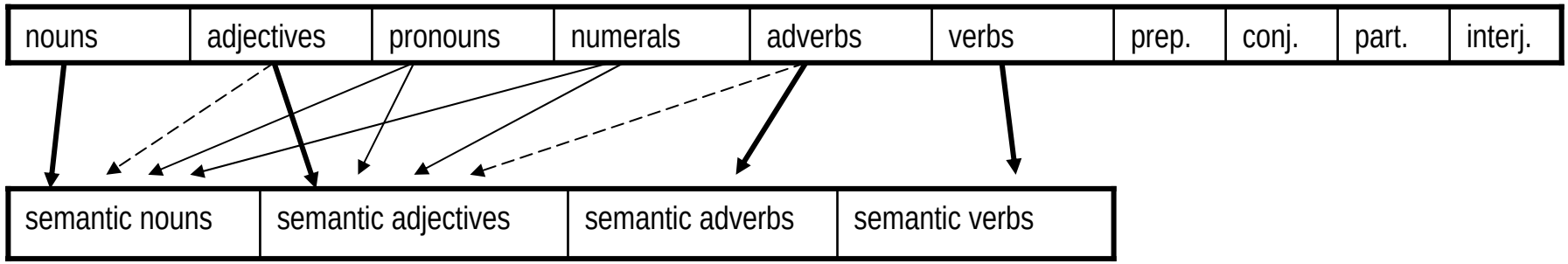
- sempos relevant only for nodetype=complex t-nodes
- 19 values of the attribute sempos:
  - n. ... | adj. ... | adv. ... | v. ...
- fully automatic annotation – use of
  - m-tag
  - t-lemma
  - other t-attributes



- sempos value delimits the set of relevant grammatemes

M-layer POS tags vs. sempos

PDT 2.0



"prototypical" relations between semantic and "traditional" parts of speech

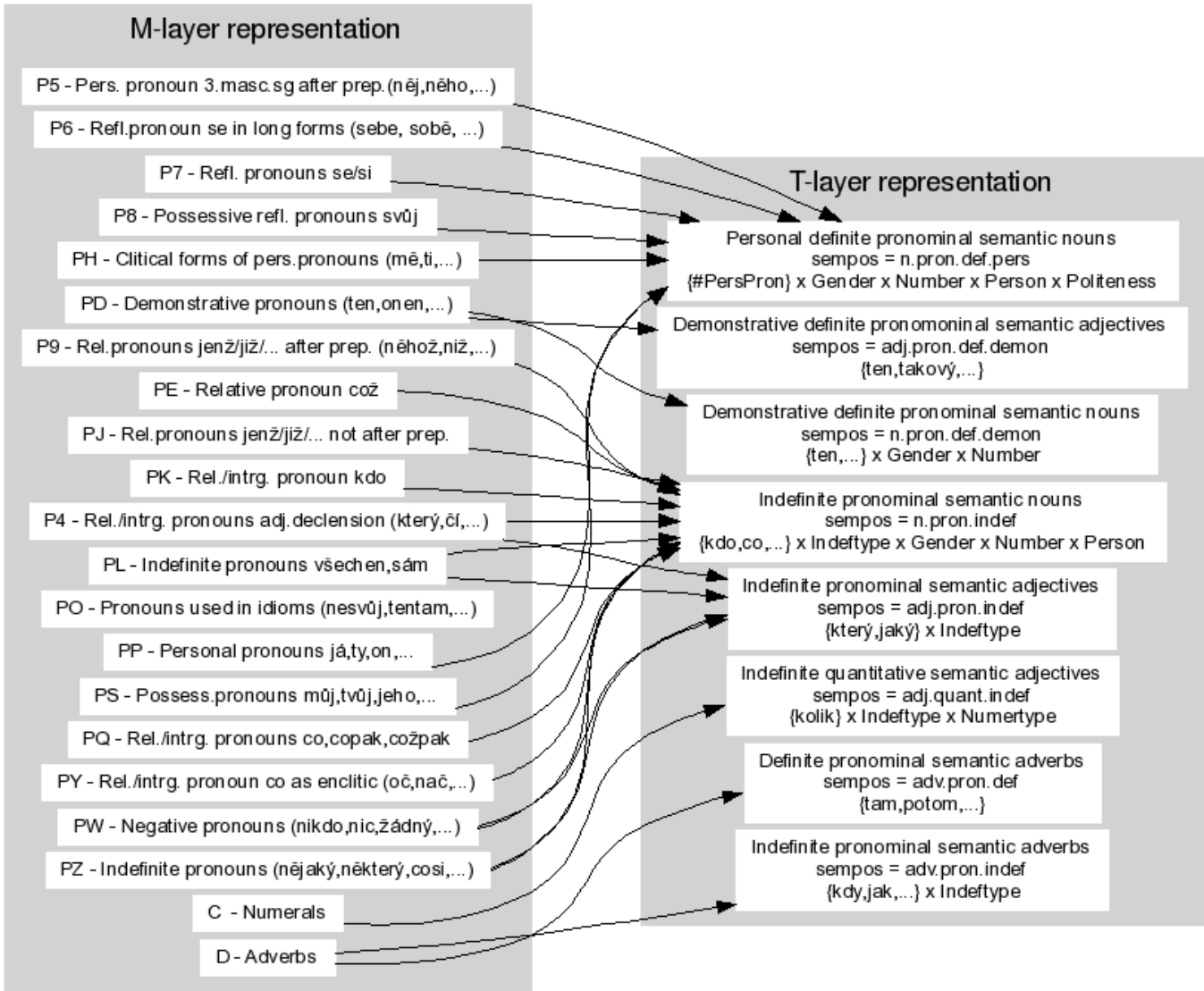
distribution of pronouns and numerals into semantic parts of speech

classification following the derivational information

- Examples of asymmetry:
  - m-layer possessive adjectives (e.g. *matčin*/mother's) converted to semantic nouns (*matka*/mother)
  - m-layer deadjectival adverbs (*pěkně*/nicely) converted to semantic adjectives (*pěkný*/nice)

PDT 2.0

Pro-forms: m-layer tags vs. t-layer sempos



Grammatemes: Annotation process

- implementation: 2000 Perl LOCs in the ntred environment
- + 2000 lines of linguistic rules
- extensive usage of m-layer and a-layer manual annotation => mostly automatic annotation possible
- only 5 man-months of human annotation needed
- grammatemes available in all tectogrammatical trees of PDT 2.0

PDT 2.0

Grammatemes - summary

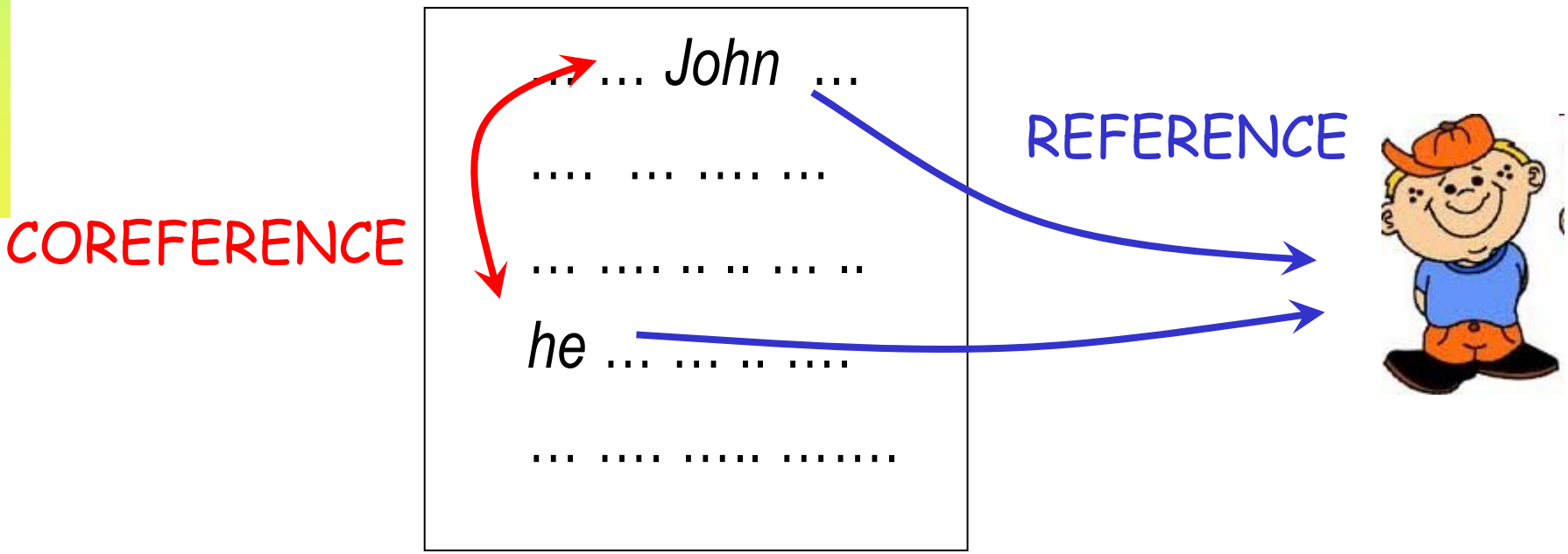
- grammateme attributes
  - component of the tectogrammatical layer
  - semantically indispensable morphological categories
  - i.e., not those imposed by agreement or other grammatical rules
  - e.g. number with nouns, tense with verbs, but not number with verbs
  - t-nodes types determine which grammatemes must be present

Part II
Coreference

PDT 2.0

What is coreference?

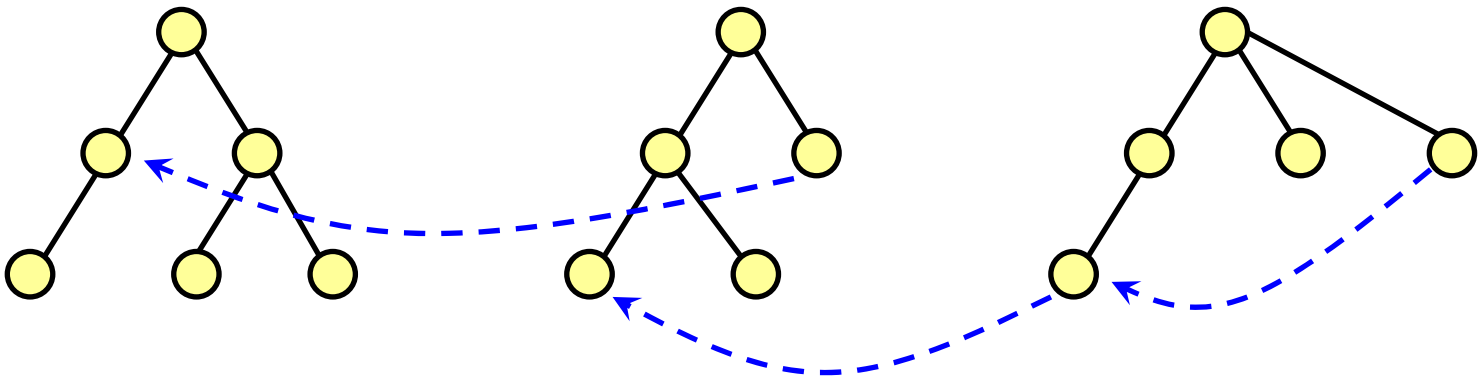
- multiple expressions in a sentence or document can refer to the same thing



PDT 2.0

Coreference in PDT

- links between tectogrammatical nodes
- technically: pointer from an anaphor t-node to its antecedent t-node
- links can form chains



PDT 2.0

Two types of coreference

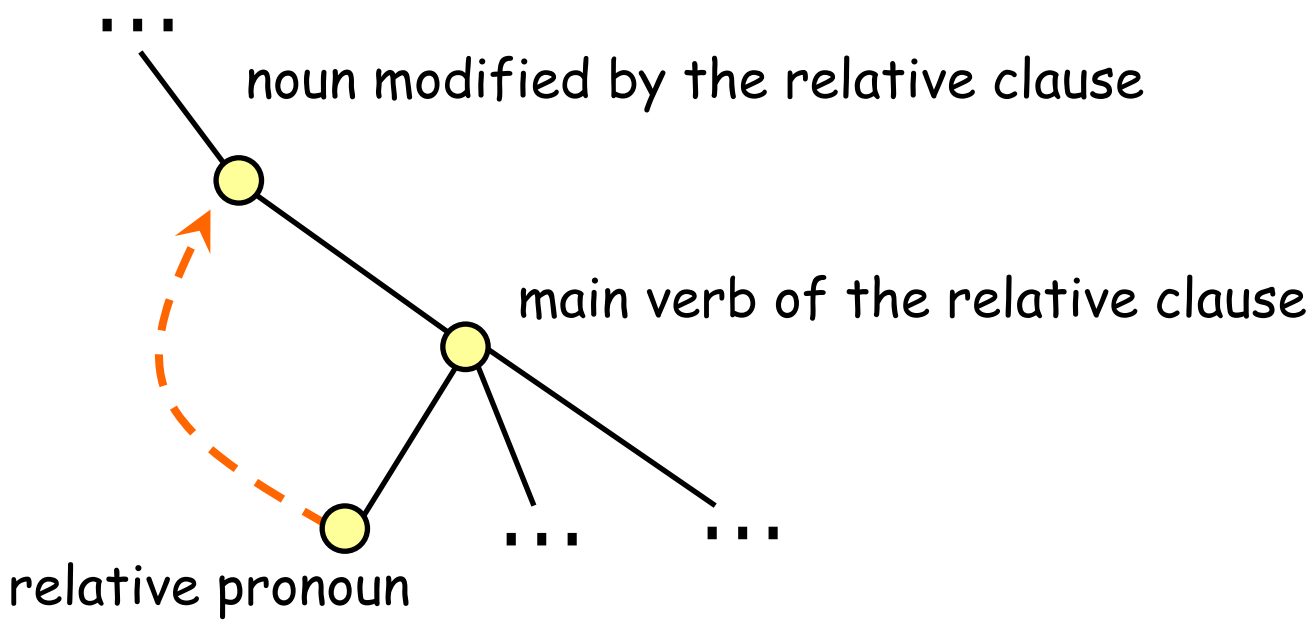
- according to Functional Generative Description, two types of coreference distinguished:

- **grammatical** coreference
  - (partially) determined by grammar rules

- **textual** coreference
  - determined only by text meaning

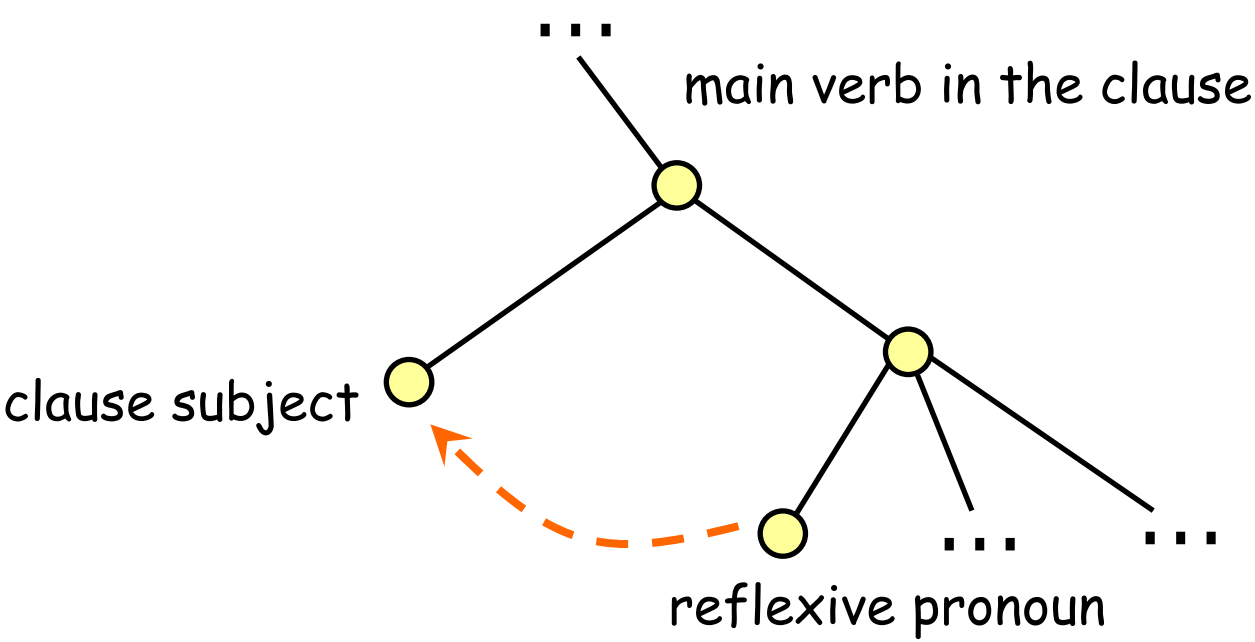
Grammatical coreference (1)

- relative pronouns
- *"The man, who…", "The man, whose …"*
- typical local configuration:



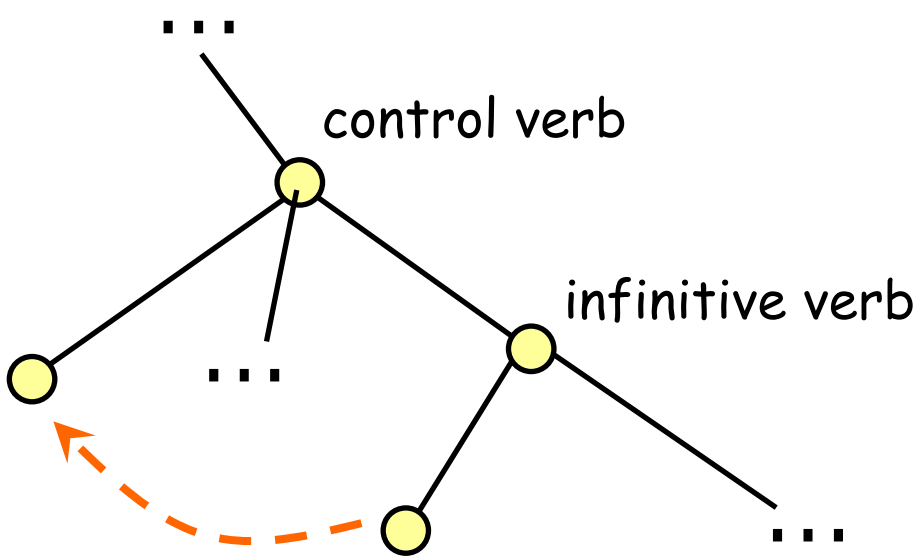
Grammatical coreference (2)

- reflexive pronouns
- in Czech, pronouns referring to clause subject have reflexive form
- typical local configuration:



Grammatical coreference (3)

- reconstructed (surface-unexpressed) actor of infinitive verbs
- *"He started to sing." "They asked him to come."*
- typical local configuration:



#Cor.ACT - reconstructed coreferential actor

PDT 2.0

Textual coreference

- anaphors:
  - personal pronouns
  - possessive pronouns
  - reconstructed pronouns (pro-drop)

PDT 2.0

Special cases

- multiple antecedent:
  - two or more parallel links from a plural anaphor (*Peter and Paul … they…*)
- cataphora
  - left-to-right links
- segm – vague reference to the previous context
- exoph - exophora

PDT 2.0

Annotated data

- manually annotated coreference in 50,000 sentences
- around 45,000 coreference links

Coreference - summary

- coreference in PDT 2.0
  - t-layer component
  - one of the largest manually annotated coreference resources
  - two types of coreference links
    - grammatical coreference
    - textual coreference
  - anaphors:
    - pronouns (personal, possessive, relative, reflexive)
    - reconstructed nodes (pro-drops,actants of infinitive verbs,...)