Prague Dependency Treebank(s): Tectogrammatical annotation

Jan Hajič (et al.)
Institute of Formal and Applied Linguistics
LINDAT/CLARIN Research Infrastructure
School of Computer Science
Faculty of Mathematics and Physics
Charles University, Prague
Czech Republic

Tectogrammatical annotation layer in PDT

- Tectogrammatical representation
  - Place in the PDT scheme
- Sublayers
  - Structure, deep dependency relations
    - Valency lexicon
  - Topic / focus (information structure)
  - Coreference
  - Semantic features
- Discourse annotation
- Summary & references/pointers

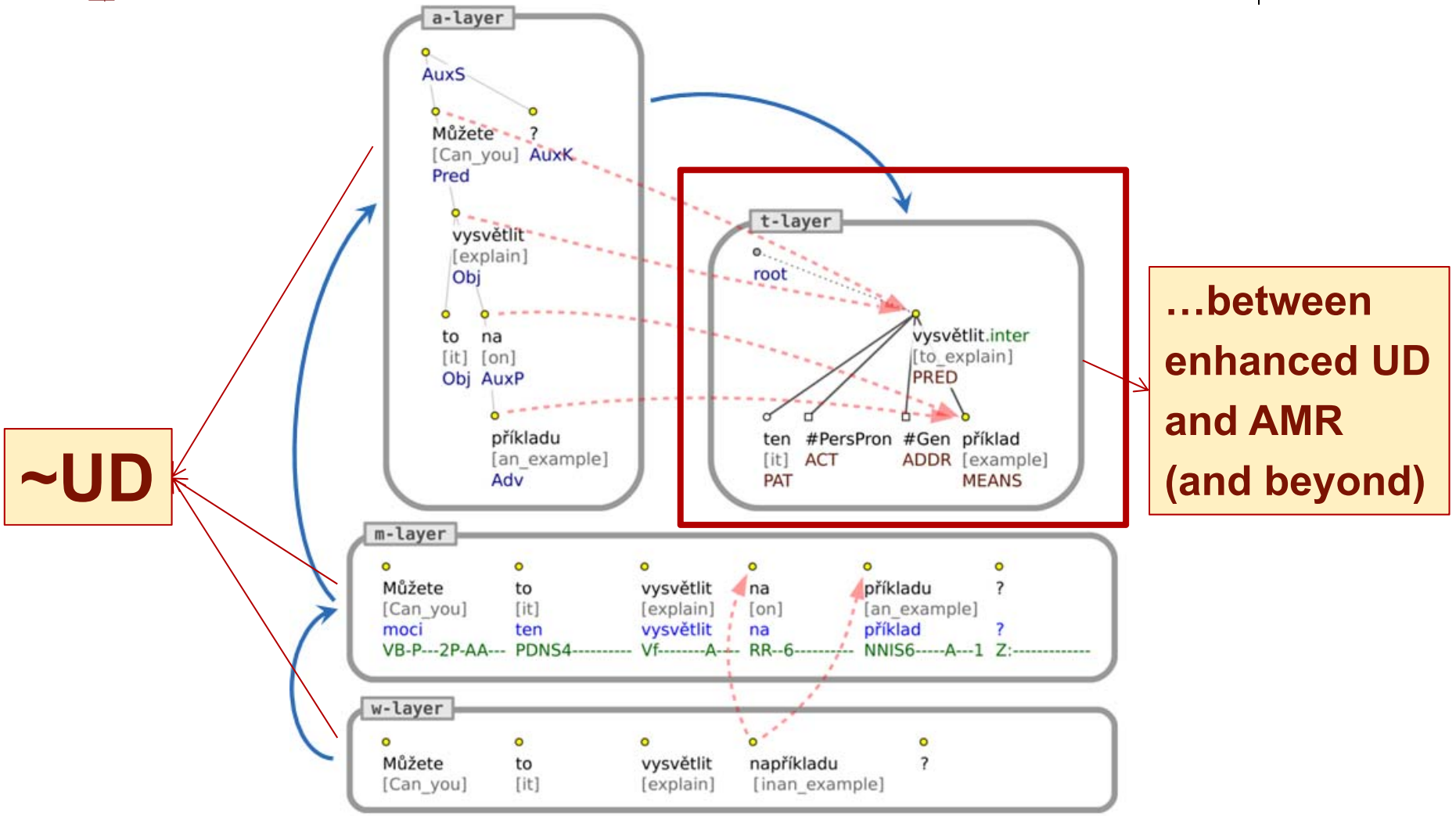
PDT Annotation Layers

- L0 (w) Words (tokens)
  - automatic segmentation and markup only
- L1 (m) Morphology
  - Tag (full morphology, 13 categories), lemma
- L2 (a) Analytical layer (surface syntax)
  - Dependency, dependency relation
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon

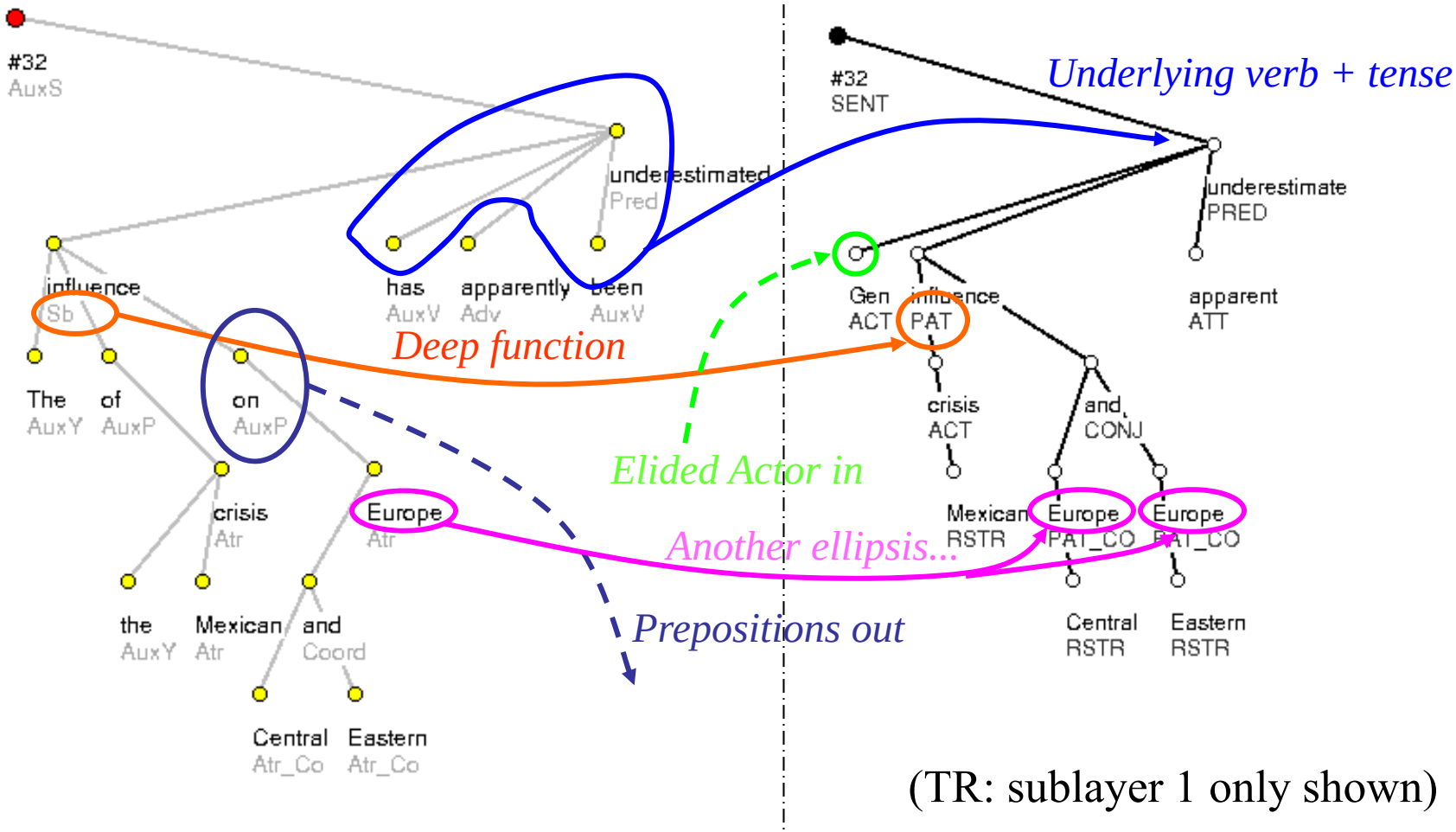
PDT Annotation Layers

- L0 (w) Words (tokens)
  - automatic segmentation and markup only
- L1 (m) Morphology
  - Tag (full morphology, 13 categories), lemma
- L2 (a) Analytical layer (surface syntax)
  - Dependency, dependency relation
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon

PDT Annotation Layers



Analytical vs. Tectogrammatical



(TR: sublayer 1 only shown)

Tectogrammatical layer (t-layer)

- 4 sublayers (integrated):
  - dependency structure, (detailed) functors
    - valency annotation
  - topic/focus and deep word order
  - coreference
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...
  - + discourse, MWE
- Total
  - 39 attributes (vs. 5 at m-layer, 2 at a-layer)

Tectogrammatical layer (t-layer)

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

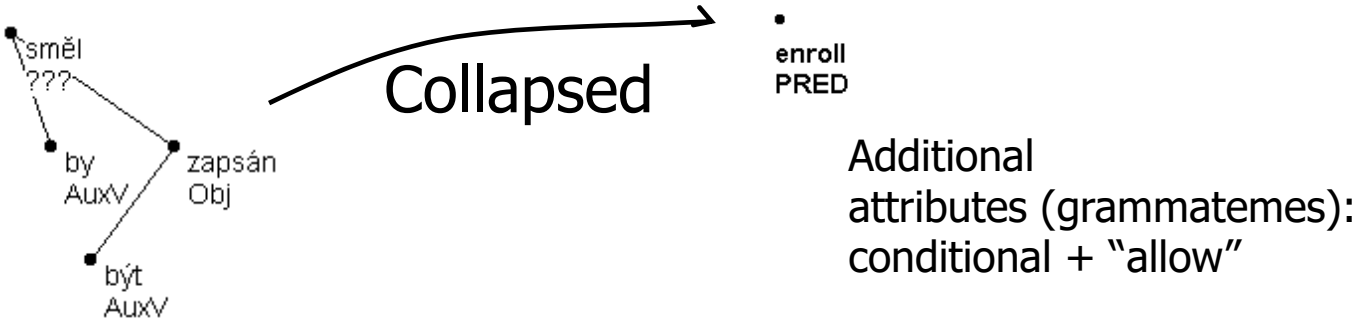
Tectogrammatical Functors
(deep syntactic/semantic relations)

"syntactic" "semi-"semantic

- "Actants": ACT, PAT, EFF, ADDR, ORIG
  - modify: verbs, nouns, adjectives
  - cannot repeat in a clause, usually obligatory
- Free modifications (~ 50), semantically defined
  - can repeat; optional, sometimes obligatory
  - Ex.: LOC, DIR1, ...; TWHEN, TTILL,...; RSTR; BEN, ATT, ACMP, INTT, MANN; MAT, APP; ID, DPHR, ...
- Special
  - Coordination, Rhematizers, Foreign phrases,...

Tectogrammatical Example

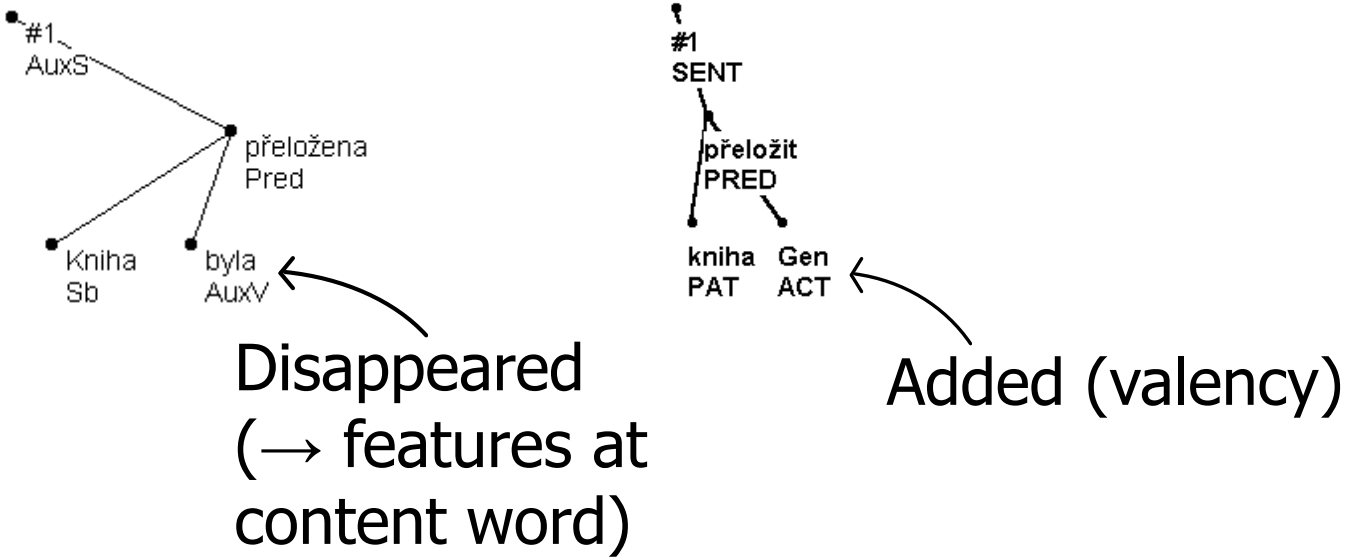
- Analytical verb form:
  - (he) allowed would-be to-be enrolled
  - směl by být zapsán



Additional attributes (grammatemes): conditional + "allow"

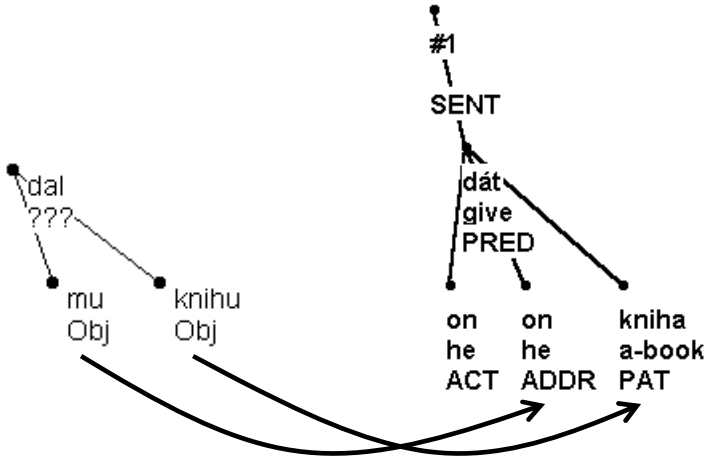
Tectogrammatical Example

- Passive (participle)
  - (The) book has-been translated [by Mr. X]
  - Kniha byla přeložena



Tectogrammatical Example

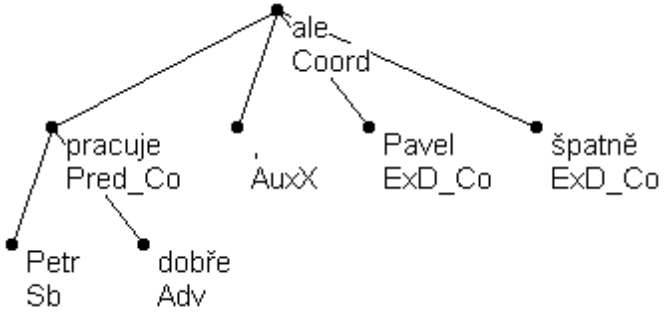
- Object
  - (he) gave him a-book
  - dal mu knihu

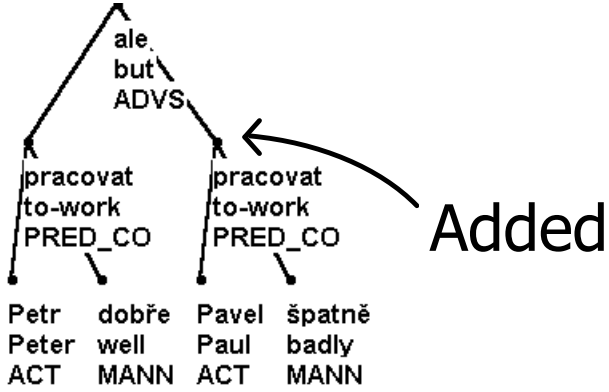


Obj goes into ACT, PAT, ADDR, EFF or ORIG based on governor's valency frame

Tectogrammatical Example

- Ellipsis (gap) (& coordination example)
  - Peter works well , but Paul badly
  - Petr pracuje dobře, ale Pavel špatně





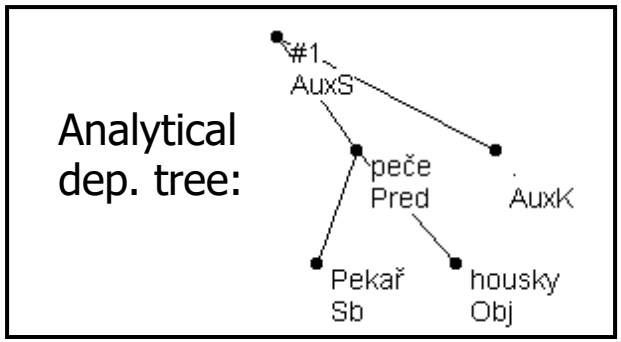
Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

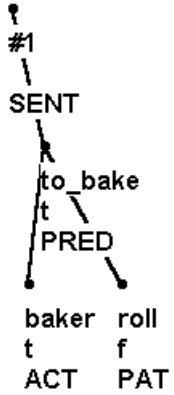
Deep Word Order Topic/Focus

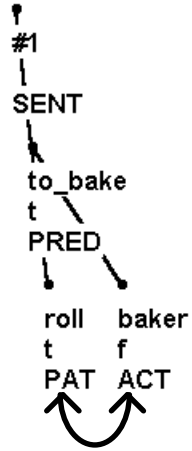
- Example:

Analytical dep. tree:



- Baker bakes rolls. vs. *Baker*[IC] bakes rolls.





Deep Word Order Topic/Focus

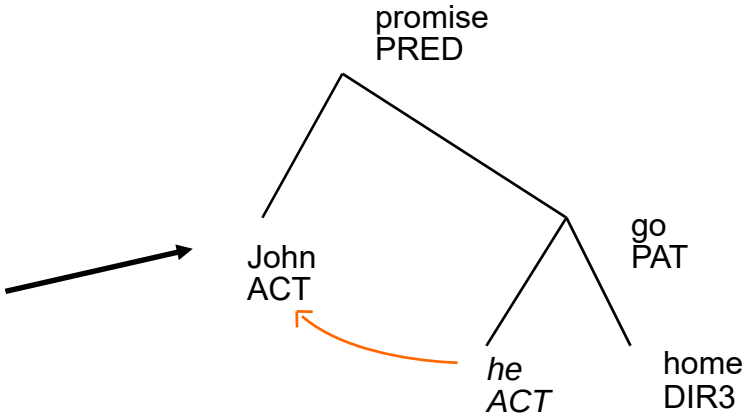
- Deep word order:
  - from “old” information to the “new” one (left-to-right) at every level (head included)
  - projectivity (almost) by definition
- Topic/focus/contrastive topic
  - attribute of every node (t, f, c)
  - restricted by d.w.o. and other constraints
- Every sentence: topic part (T) / focus part (F)
  - ~ scope

Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

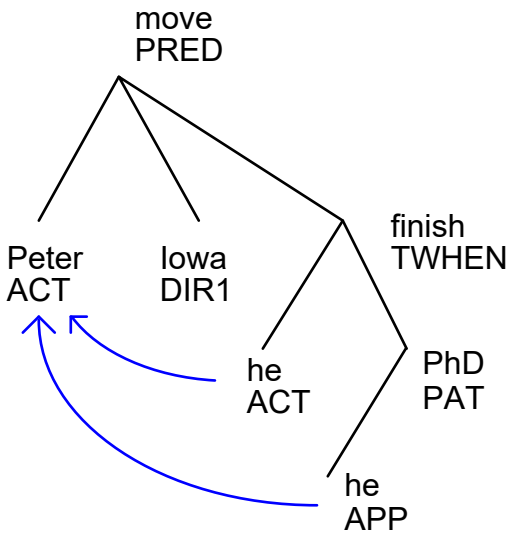
Coreference

- Grammatical
  - relative clauses
    - which, who
      - Peter and Paul, who ...
  - control
    - infinitival constructions
      - John promised to go ...
  - reflexive pronouns
    - {him,her,thme}self(-ves)
      - Mary saw herself in ...



Coreference

- Textual
  - Ex.: Peter moved to Iowa after he finished his PhD.



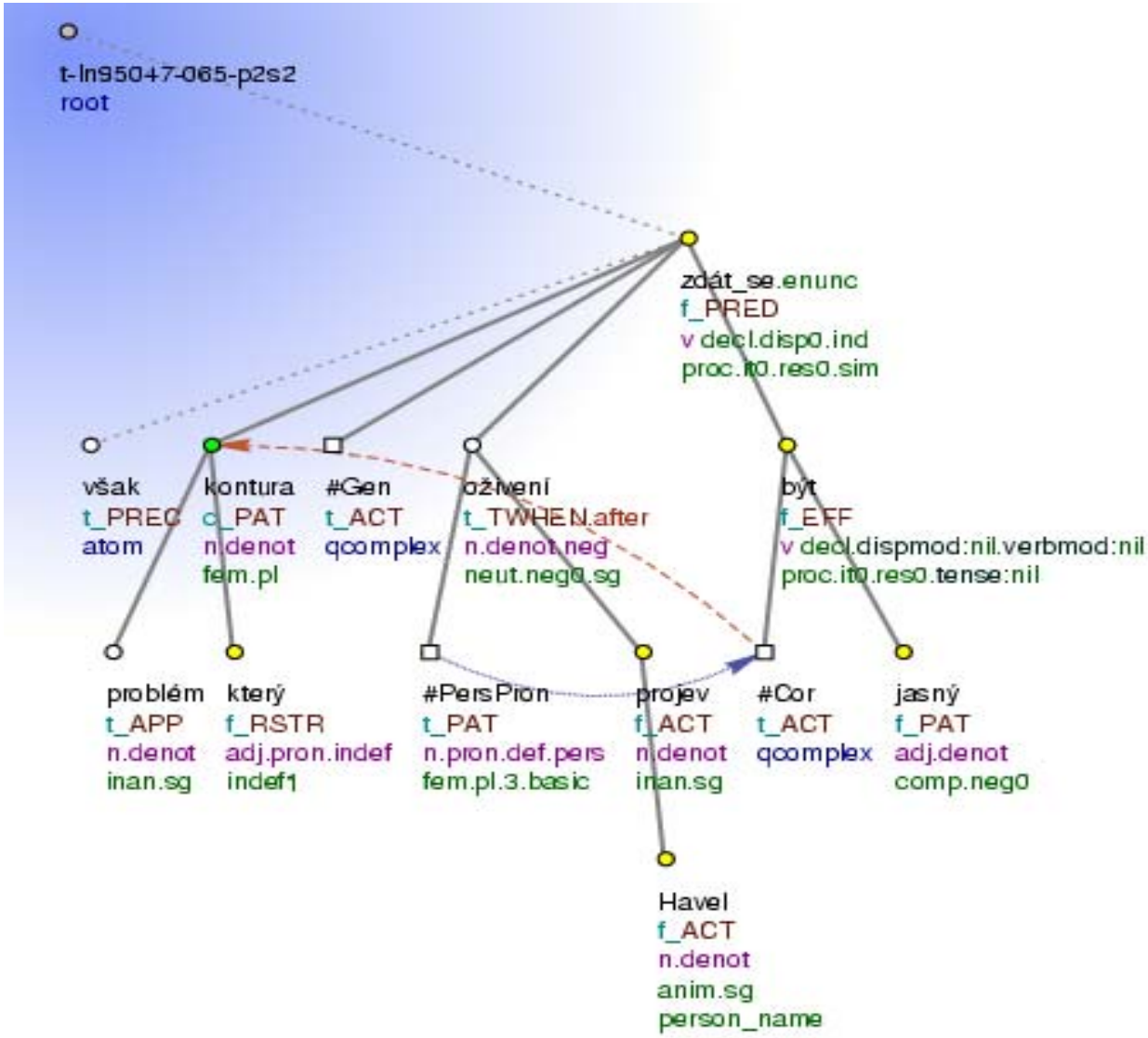
Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

Grammatemes (semantic features)

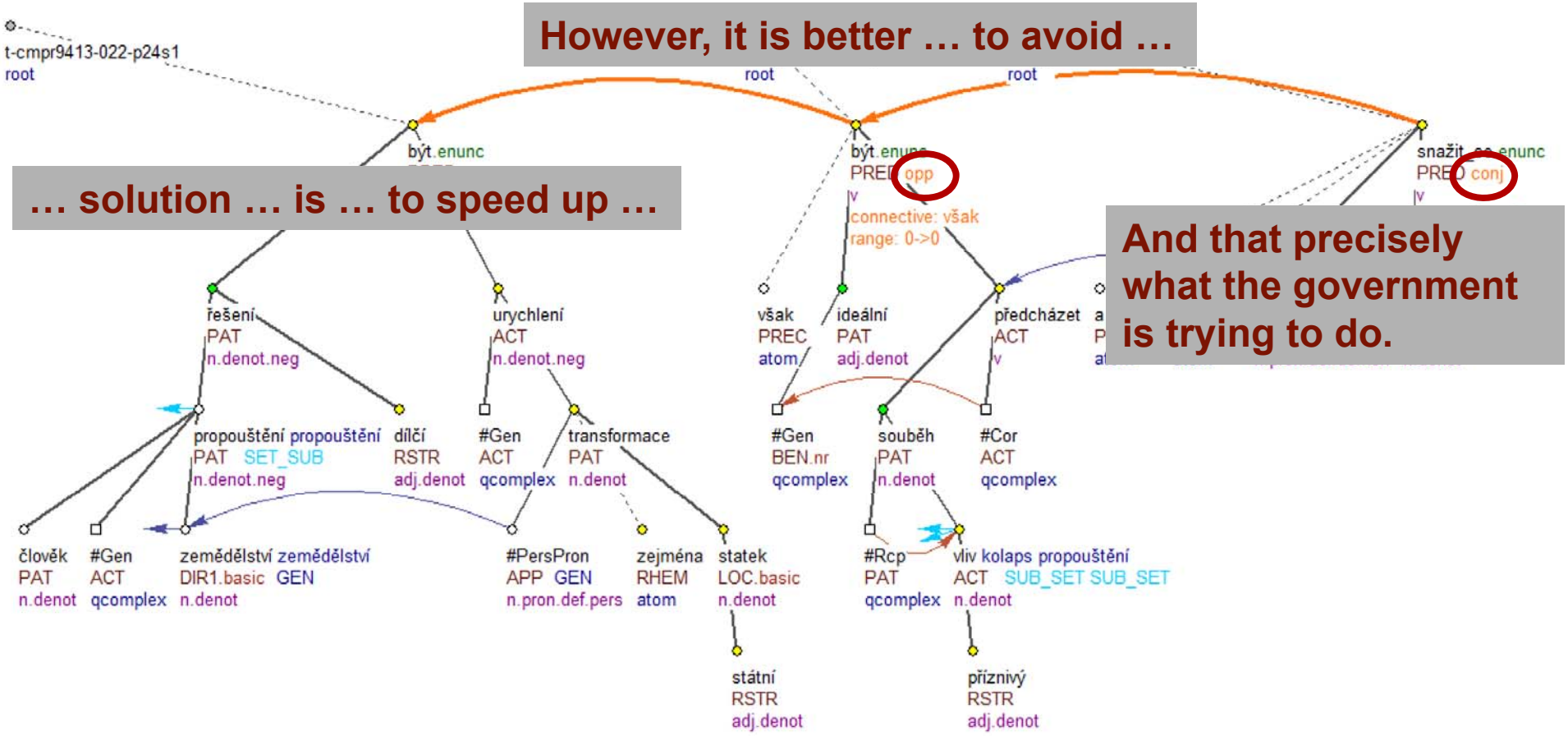
- Detailed functors (subfunctors)
  - only for some functors:
    - TWHEN: before/after
    - LOC: next-to, behind, in-front-of, ...
    - also: ACMP, BEN, CPR, DIR1, DIR2, DIR3, EXT
- Lexical (underlying)
  - number (SG/PL), tense, modality, degree of comparison, ...
  - only where necessary (agreement!)

Fully Annotated Sentence



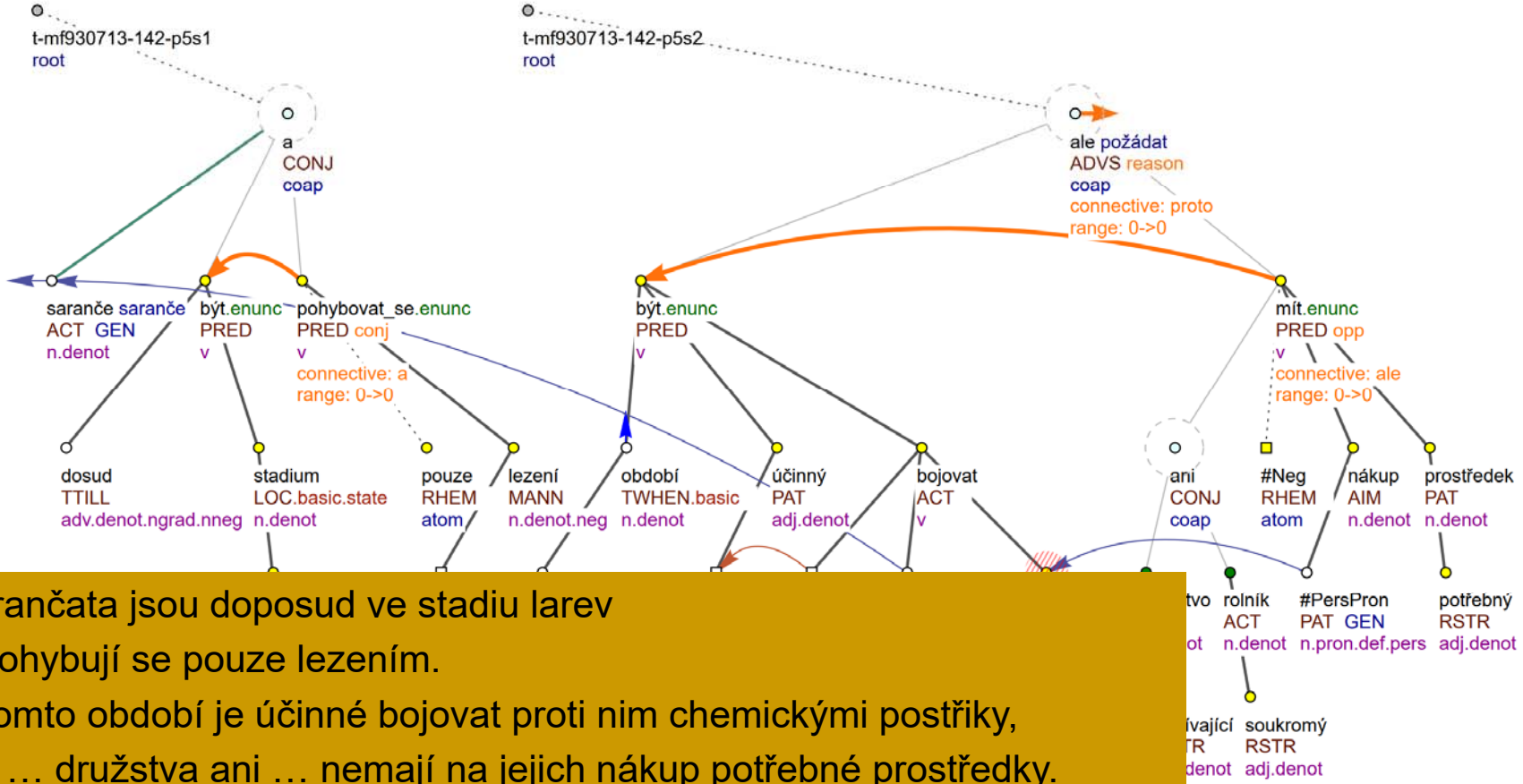
The boundaries of some problems seem to be clearer after they were revived by Havel's speech.

Discourse annotation
(~ Penn Discourse Treebank)



Discourse annotation

- Inspiration: the Penn Discourse Treebank

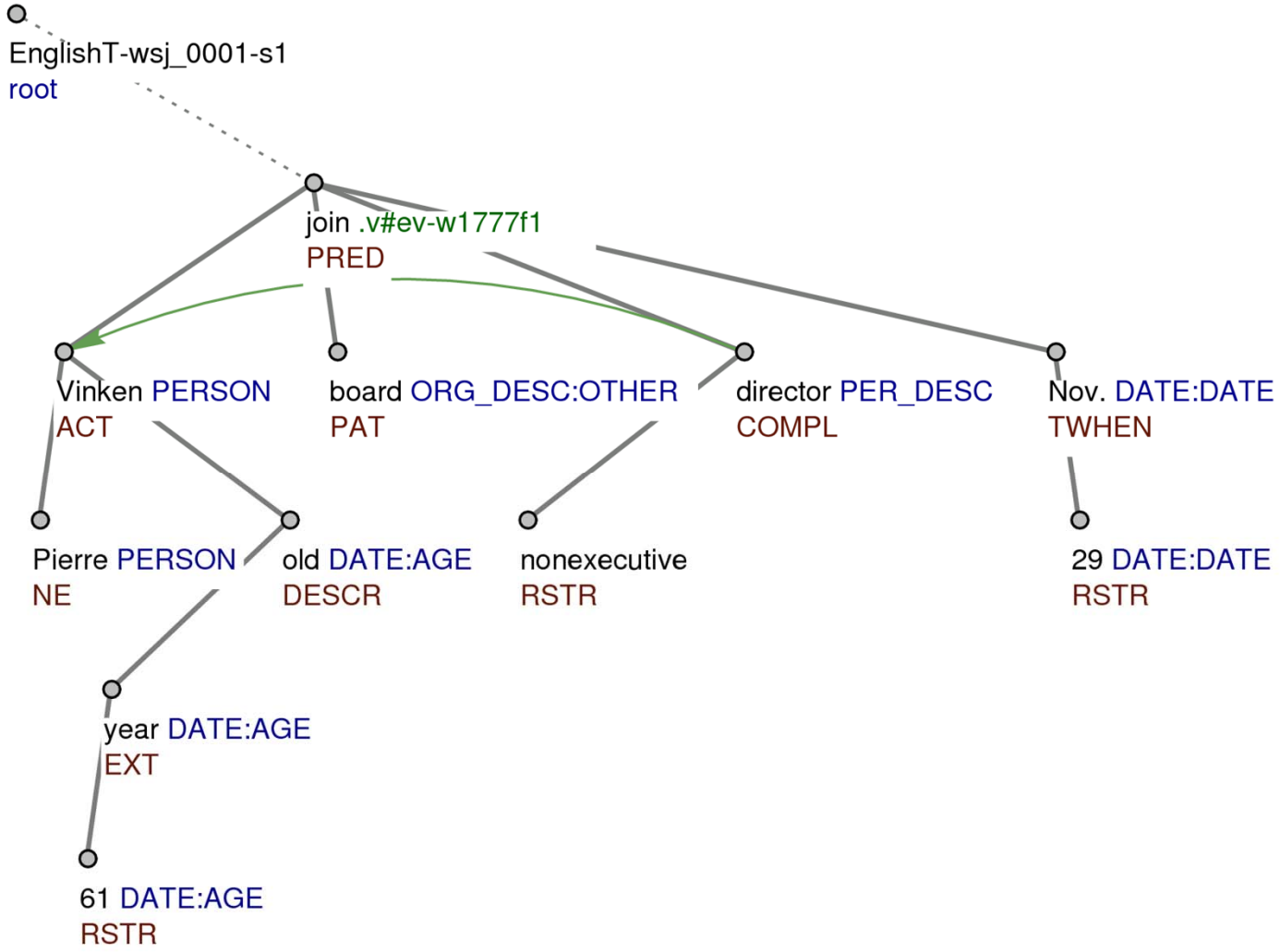


Sarančata jsou doposud ve stadiu larev
a pohybují se pouze lezením.
V tomto období je účinné bojovat proti nim chemickými postřiky,
ale … družstva ani … nemají na jejich nákup potřebné prostředky.
Proto požadují …

English PDT-style Annotation

- Morphology and Syntax
  - By conversion
- Tectogrammatical annotation
  - Guidelines (English TR: by S. Cinková)
  - Pre-annotation
    - Transformation from Penn Treebank & Propbank (Palmer, Kingsbury) by Z. Žabokrtský et al.
  - Valency
    - From Propbank Frame Files (Cinková, Šindlerová, Nedolužko, Semecký)

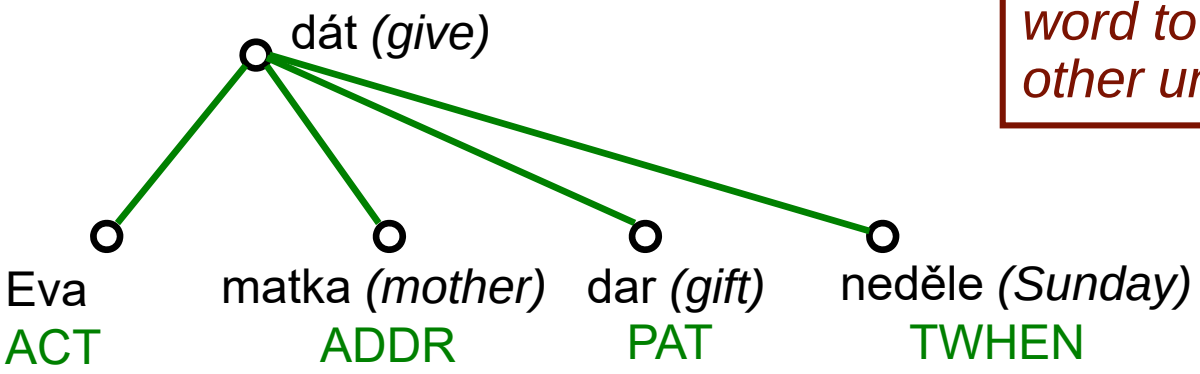
Example - English TR



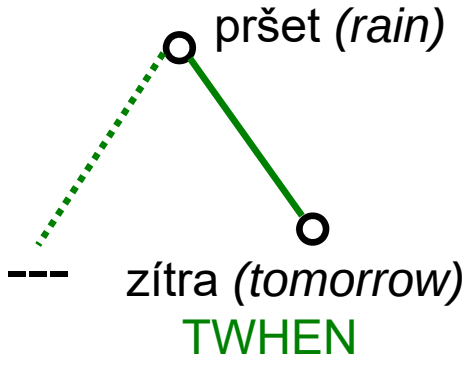
- Words
- Dependencies
- Sem. function
- Valency (predicates)
- Coref (BBN)
- Named Entities (BBN)

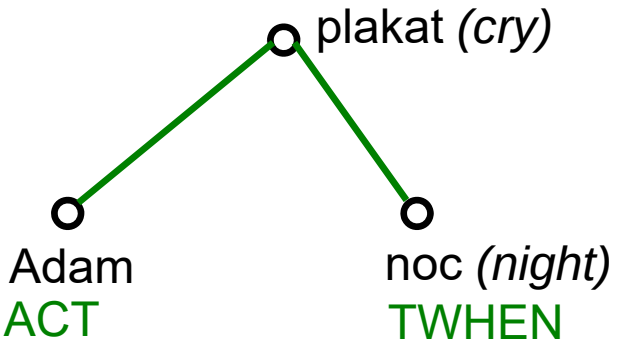
Valency in the PDT

Valency: *specific ability of a word to combine itself with other units of meaning*



Specifies nothing





Valency - Basic Principles

inner participants vs. free modifications
(arguments vs. adjuncts)

obligatory vs. optional modifications
(the dialogue test)

Inner Participant …

… Free Modification

ACT(or), PAT(ient)
ADDR(essee), EFF(ect),
ORIG(in) (5)

- each occurs just with particular verbs
- each modifies the verb only once (in a clause)

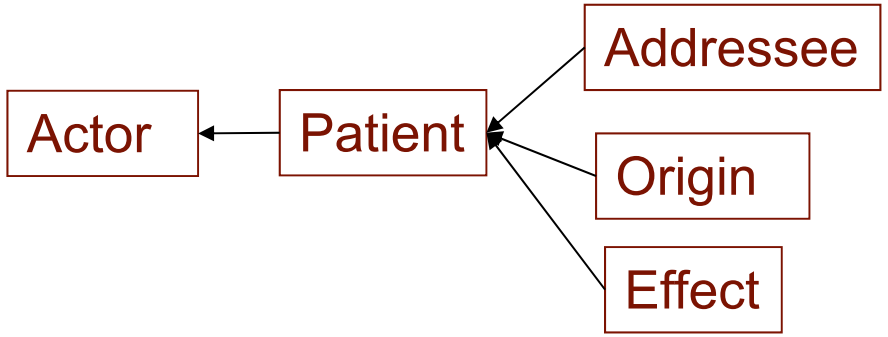
Location (LOC, DIR1,…)
Time (TWHEN, TTILL, …),
Manner, Intention,… (70)

- can modify in principle any verb
- can be repeated (within the same clause)

Inner Participants

syntactic criteria - Actor and Patient
semantic criteria for other inner participants (if a verb has more than two arguments)

Argument shifting



Petr has dug a hole.

➔ Semantic Effect (as a cognitive role) shifted to the position of Patient.

The teacher asked a pupil.

➔ Semantic Addresse shifted to the position of Patient.

Obligatory … Optional

The Dialogue Test

Answering a question about a semantically obligatory modification, the speaker cannot say: *I don't know*.

A: *John left.*
B: *From where?*
A: **I don't know.*

A: *John left.*
B: *To where?*
A: *I don't know.*

„from where“
➔ obligatory modification

„to where“
➔ optional modification

Valency frame

Structure:

	obligatory	optional
argument		
adjunct		

Contents:

- functor
- obligatoriness
- surface form

one meaning of the word ➔ one valency frame

word: *leave*
meaning 1: *sb left sth* → frame1: ACT PAT
meaning 2: *sb left from somewhere* → frame2: ACT DIR1

Valency lexicon: PDT-VALLEX

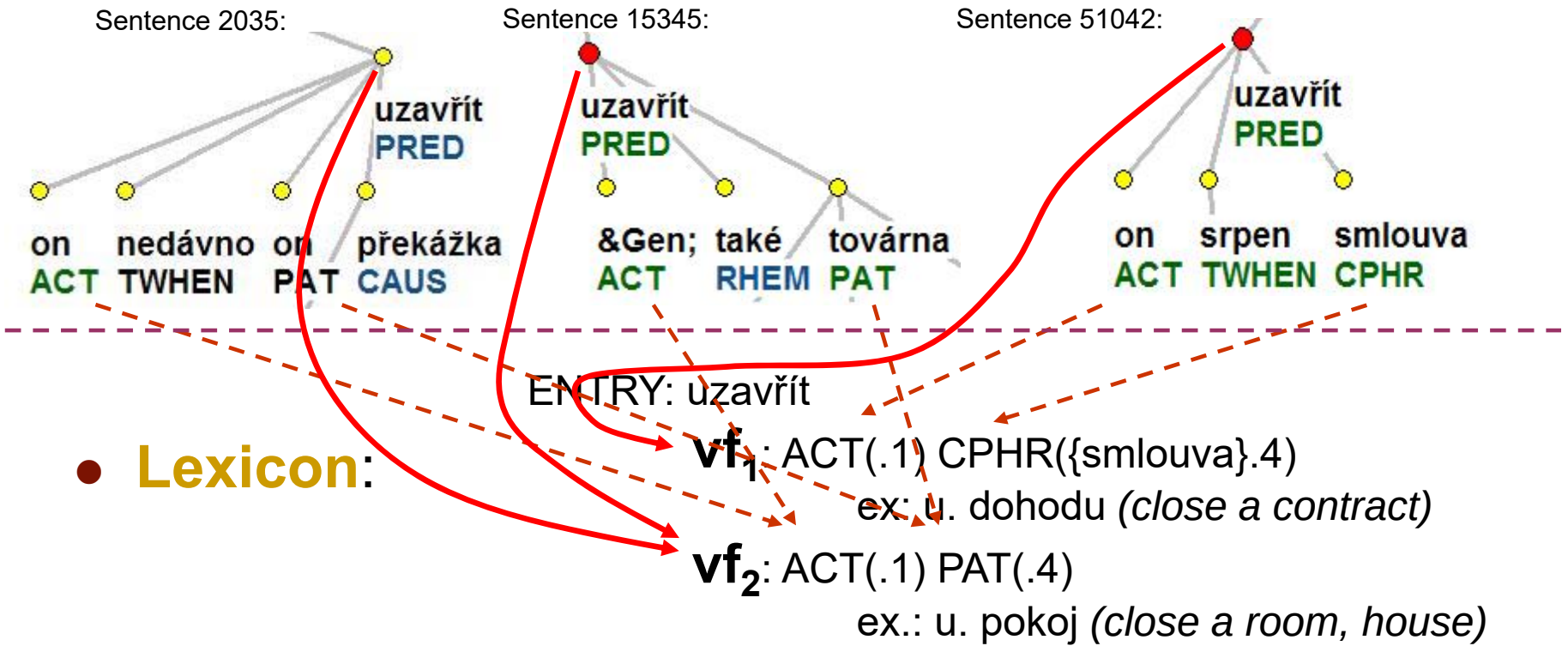
- 11500 verb senses / valency frames
- 9000 noun sense / valency frames
- some adjectives and adverbs

PDT-VALLEX Entry
verb: *dosáhnout*
meaning 1: *to reach sth*
meaning 2: *to get sb to do sth*
meaning 3: …
meaning 4: …

*** dosáhnout**
ACT(.1) PAT(.2,.4) v-w714f1 **Used:** 272x
dosáhnout určité úrovně
mzda d. v tomto oboru 80 tisíc
d. pokročilého věku
ACT(.1) PAT(.2,*aby*[.v]) ?ORIG(*na-1*[.6],*od-1*[.2]) v-w714f2 **Used:** 7x
dosáhl na něm slibu
dosáhli na sobě slibu
ACT(.1) DPHR(*svůj-1*.2) v-w714f3 **Used:** 2x
dosáhl svého
ACT(.1) DIR3(*) v-w714f4 **Used:** 2x
dosáhl na strop
rukou.MEANS

Corpus ↔ Valency Lexicon

- **Corpus** – occurrences of „uzavřít“ (*to close*) :



- **Lexicon**:

Tectogrammatical Parsing

- 4 phases
- Transformation -based learning
- FnTBL
- Largely language independent
- Coreference: >90%

(V. Klimes' thesis)

CoNLL 2009 ST: Czech 83.27

Attribute	m- and a-layer: manual	auto
structure	89,3 %	76,4 %
functor	85,5 %	77,4 %
val_frame.rf	92,3 %	90,9 %
t_lemma	93,5 %	90,9 %
nodetype	94,5 %	92,6 %
gram/sempos	93,8 %	91,5 %
a/lex.rf	96,5 %	95,1 %
a/aux.rf	94,3 %	90,3 %
is_member	94,3 %	89,5 %
is_generated	96,6 %	95,2 %
deepord	68,0 %	66,7 %

To summarize…

- PDT is/has (a)…
  - Dependency-based treebanking project
    - Czech (other languages: – Eng, Ar)
    - Ongoing projects (other inst.): Italian, Old Greek, Latin, …
  - ~ 1mil. words
    - sufficient size for ML experiments
  - 4 layers of annotation
    - token, morphology, syntax, **deep syntax/semantics**)
    - interlinked (for the development of parsers/generators)
  - Valency dictionary integrated (links from data)
  - Multiword expressions, discourse

Some pointers

- Current version of PDT: v3.5
  - all three levels, 1.9/1.5/0.8 Mwords
  - http://ufal.mff.cuni.cz/pdt3.5
  - LINDAT/CLARIN – search for “prague dependency”
- http://ufal.mff.cuni.cz
  - Research -> Corpora (Treebank(s))
- http://www.ldc.upenn.edu
  - LDC2004T23 (PADT 1.0), LDC2012T08 (PCEDT 2.0), LDC2006T01 (PDT 2.0)
- http://ufal.mff.cuni.cz/pcedt2.0
  - Parallel Czech-English Dependency Treebank
- Valency lexicons
  - https://lindat.mff.cuni.cz/services/CzEngVallex/
  - https://lindat.mff.cuni.cz/services/EngVallex