Tectogrammatical Representation of Meaning vs. AMRs

Jan Hajič
Institute of Formal and Applied Linguistics & LINDAT-Clarin
School of Computer Science
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

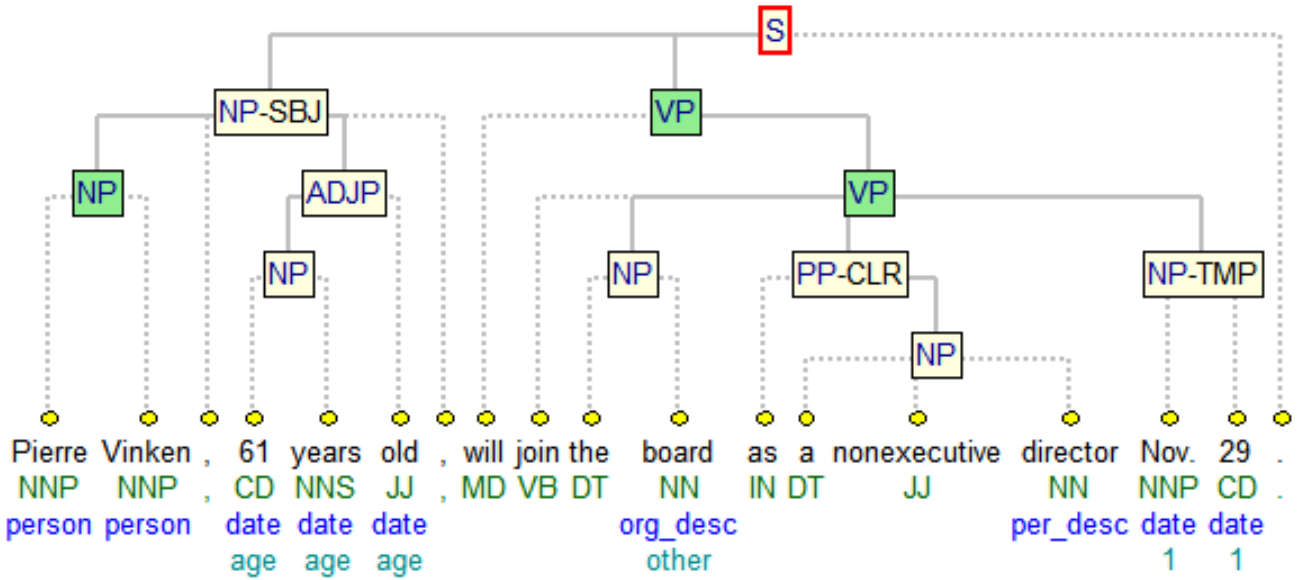
Outline

- Treebank styles, the Penn Treebank
- The Prague Dependency Treebank
  - Style of representation
    - Morphology
    - Syntax
    - Semantics aka Tectogrammatical Representation of meaning
  - Spoken language treebank: the challenges
- Relation to AMR
  - Tectogrammatical representation vs. AMR
  - Czech – English comparison (AMR)

Phrase- vs. Dependency-Based Treebanks

- The original: The Penn Treebank
  - Phrase-based style; good for parsing by CFG grammars
- Followers
  - Almost all Penn-based treebanks
    - Chinese, Arabic, Korean, …
  - Negra (German), many others
- Now: dependency parsing prevails; training data?
    - Conversion from phrase-based treebanks
      - Might lose information, heads added „ad hoc“
    - “native” dependency treebanks: annotated as such
      - Considered “better”
      - Hindi/Urdu, TIGER (sort of); both styles manually annotated
      - PDT (of course) and similar ones
        - » PDT style treebanks: Danish, Croatian, Slovene, Greek, Latin

The Penn Treebank



- Published (first) in 1993, now LDC99T42 (www.ldc.upenn.edu)
  - First the Wall Street Journal part (1 mil. words, 2312 documents)
- Added other text types
  - ATIS corpus (dialogs, travel reservations)
  - Brown corpus annotated for syntax
  - Switchboard (spoken language, tel. conversations)

The Penn Treebank(s)

- Extensions
  - Annotation of named entities, co-reference (BBN)
    - cf. also previous slide
  - Function labels (SBJ, OBJ, TMP, ...)
  - PropBank
    - Penn Treebank syntax + Predicate-argument relations, added “frame files” (predicate dictionary)

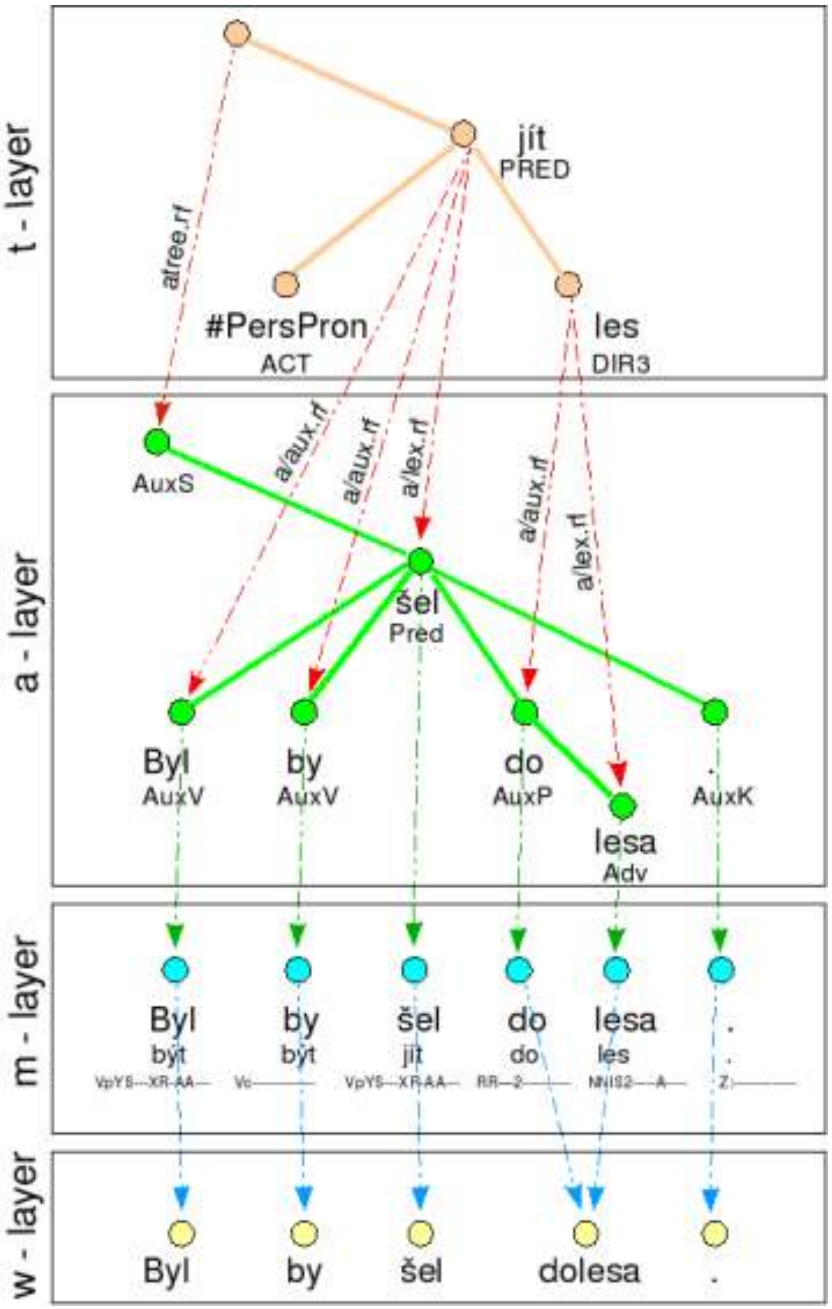
      (S (NP-SBJ (PRP I Arg0) VP (VBD gave Pred) (NP-DOBJ (PRP him Arg1) (NP-IOBJ (DET the) (NN book Arg2))) ...)
  - NomBank
    - Like PropBank, but for nouns and their “arguments”
- Other languages (Chinese, Arabic, ...)

The Prague Dependency Treebanks: the Basics

- 2001-, now at version 3.0
- Basic general features
  - Multilayered annotation, interlinked layers
  - Dependency-based syntax (both surface and deep)
  - Information structure of the sentence (topic/focus)
  - Grammatical and basic textual coreference
  - Multiword Entities
  - Discourse relations
- Languages: Czech, English (also parallel), Arabic
  - Student work on “samples”:
    - Indonesian, Urdu, Russian, …
  - Spoken: work started on Czech and English
    - non-parallel, dialogs

The Prague Dependency Treebank

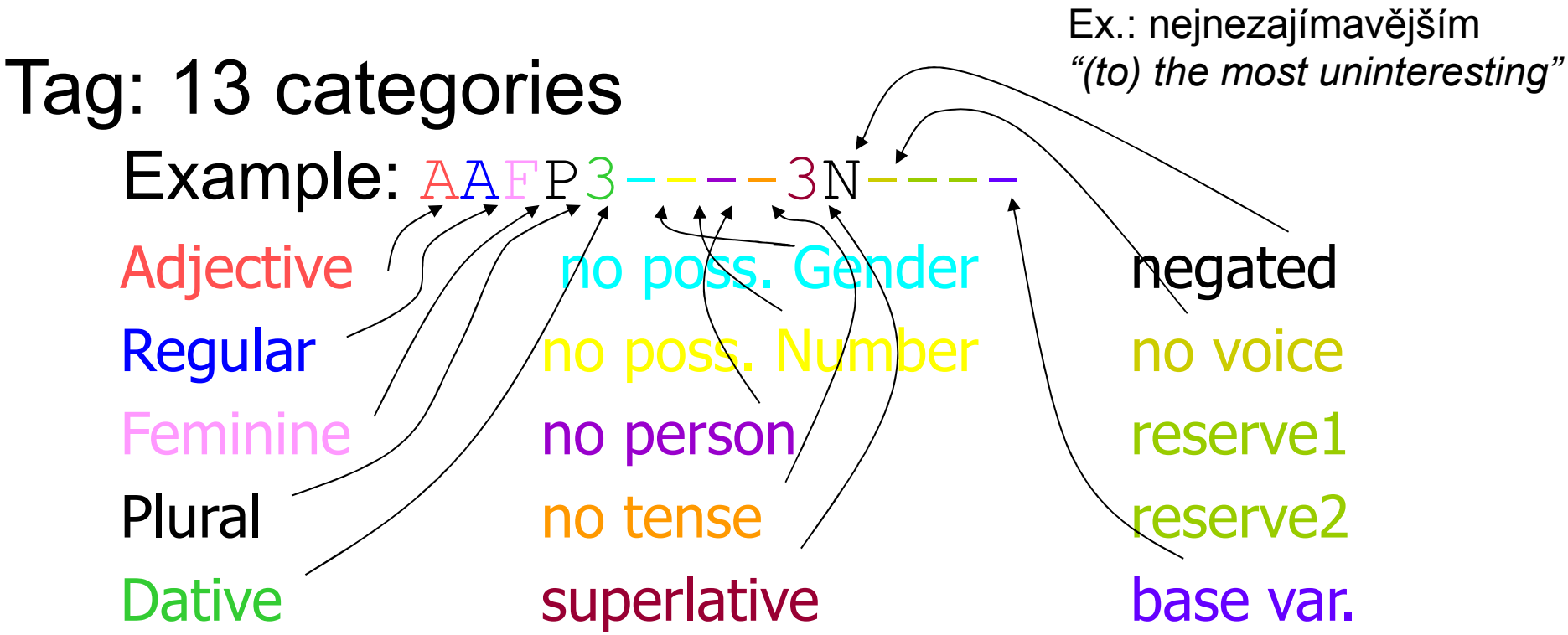
- Three basic layers of annotation
  - Morphological layer
  - Surface syntax ("analytical") layer
  - "Tectogrammatical" layer: underlying syntax, semantic roles (valency), inf. structure, coreference
- Size
  - 830,000 words (tokens)
    = 50000 sentences in 3165 full documents (texts)
- Format
  - Prague Markup Language (XML-based)
  - 830,000 words (tokens)
    = 50000 sentences in 3165 full documents (texts)



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, analytical dependency function
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon

Morphological Attributes



Lemma: POS-unique identifier

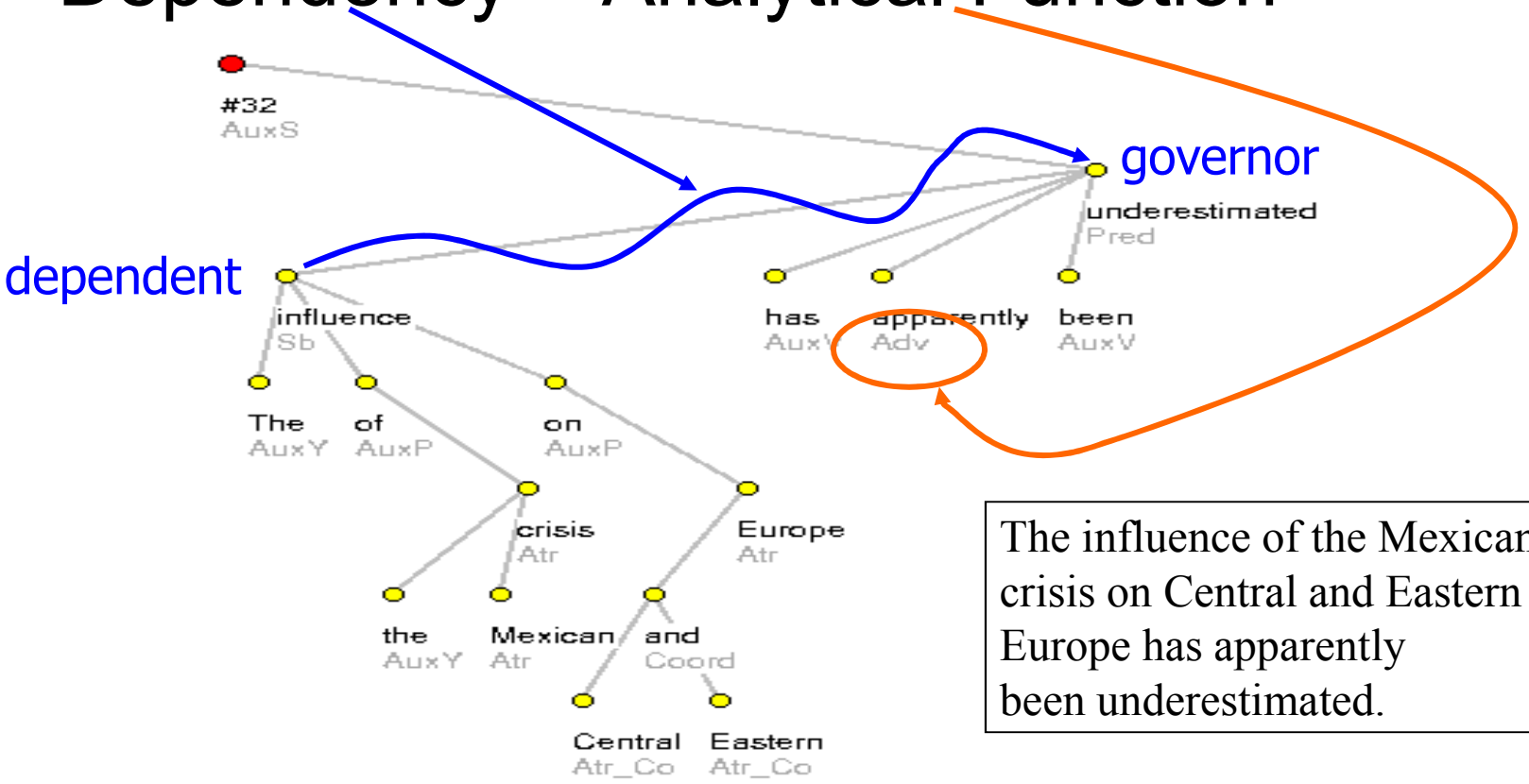
Books/verb -> book-1, went -> go, to/prep. -> to-1

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, analytical dependency function
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon

Layer 2 (a-layer): Analytical Syntax

- Dependency + Analytical Function



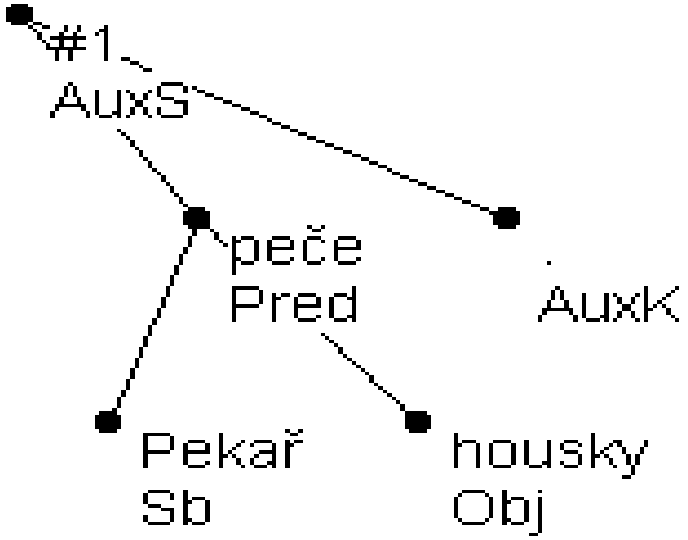
The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated.

Analytical Syntax: Functions

- Main (for [main] semantic lexemes):
  - Pred, Sb, Obj, Adv, Atr, Atv(V), AuxV, Pnom
  - "Double" dependency: AtrAdv, AtrObj, AtrAtr
- Special (function words, punctuation,...):
  - Reflefives, particles: AuxT, AuxR, AuxO, AuxZ, AuxY
  - Prepositions/Conjunctions: AuxP, AuxC
  - Punctuation, Graphics: AuxX, AuxS, AuxG, AuxK
- Structural
  - Elipsis: ExD, Coordination etc.: Coord, Apos

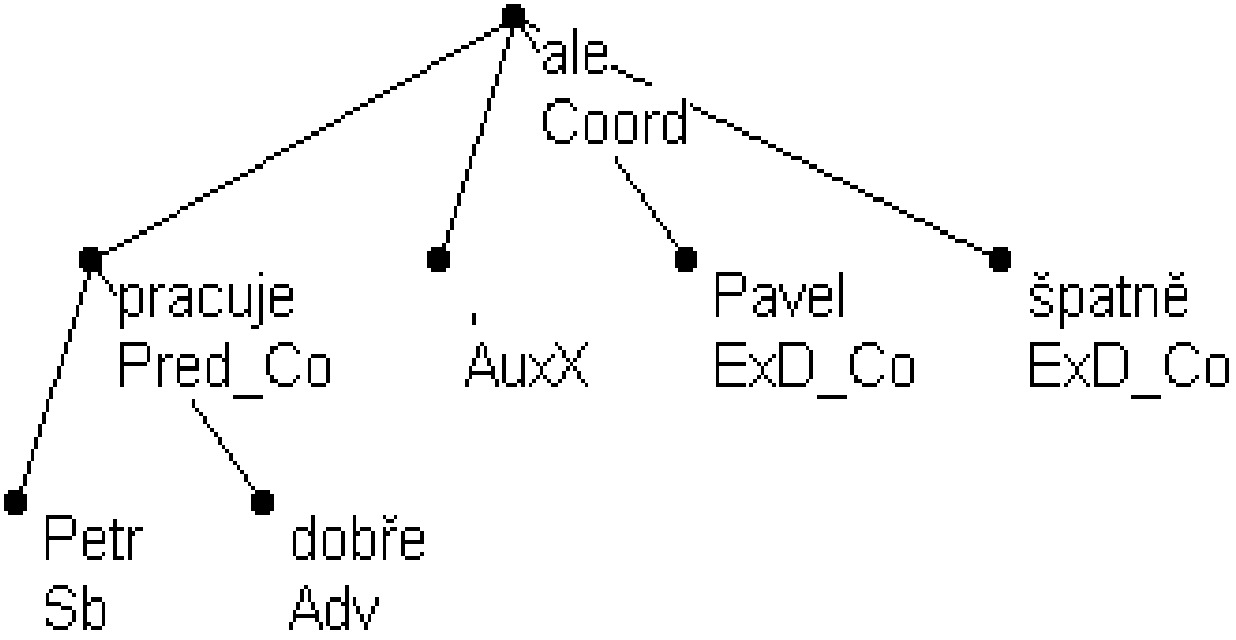
Surface Syntax Example

- Complete sentence: Sb, Pred, Obj
  - » The-baker bakes rolls.
  - » Pekař peče housky.



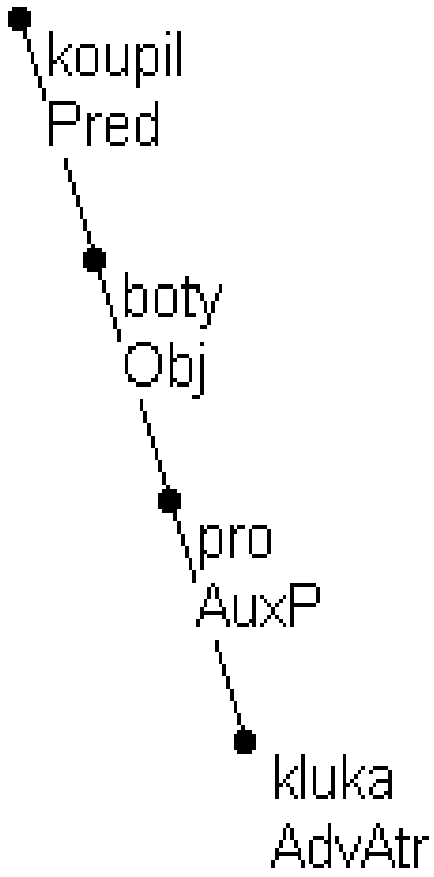
Surface Syntax Example

- Incomplete phrases
  - » Peter works well , but Paul badly
  - » Petr pracuje dobře, ale Pavel špatně



Surface Syntax Example

- Variants (equal meaning)
  - » (he) bought shoes for boy
  - » koupil boty pro kluka



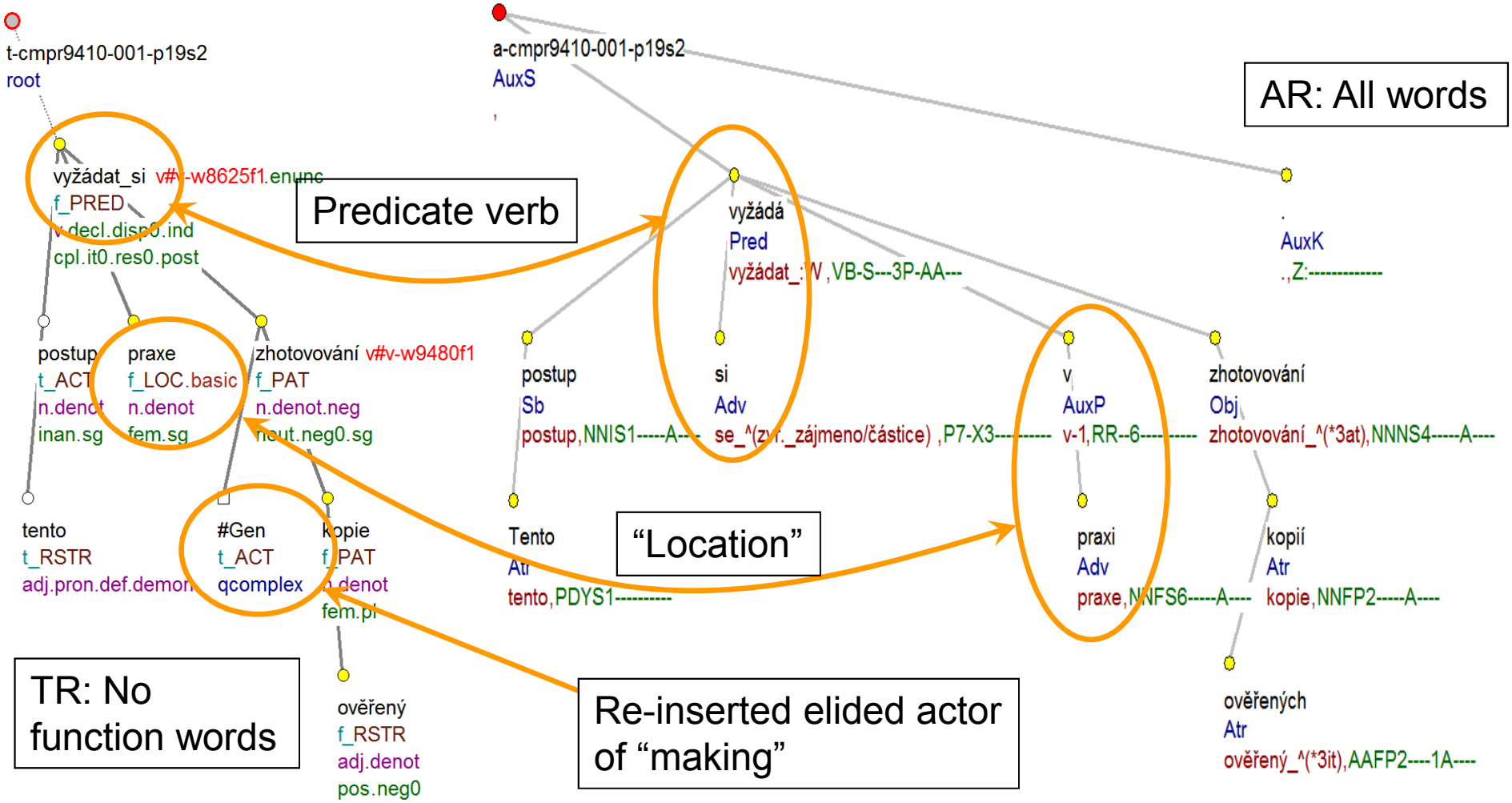
PDT Annotation Layers

- 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, analytical dependency function
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon

Layer 3 (t-layer): Tectogrammatical Annotation

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

Tectogrammatical vs. analytical syntax

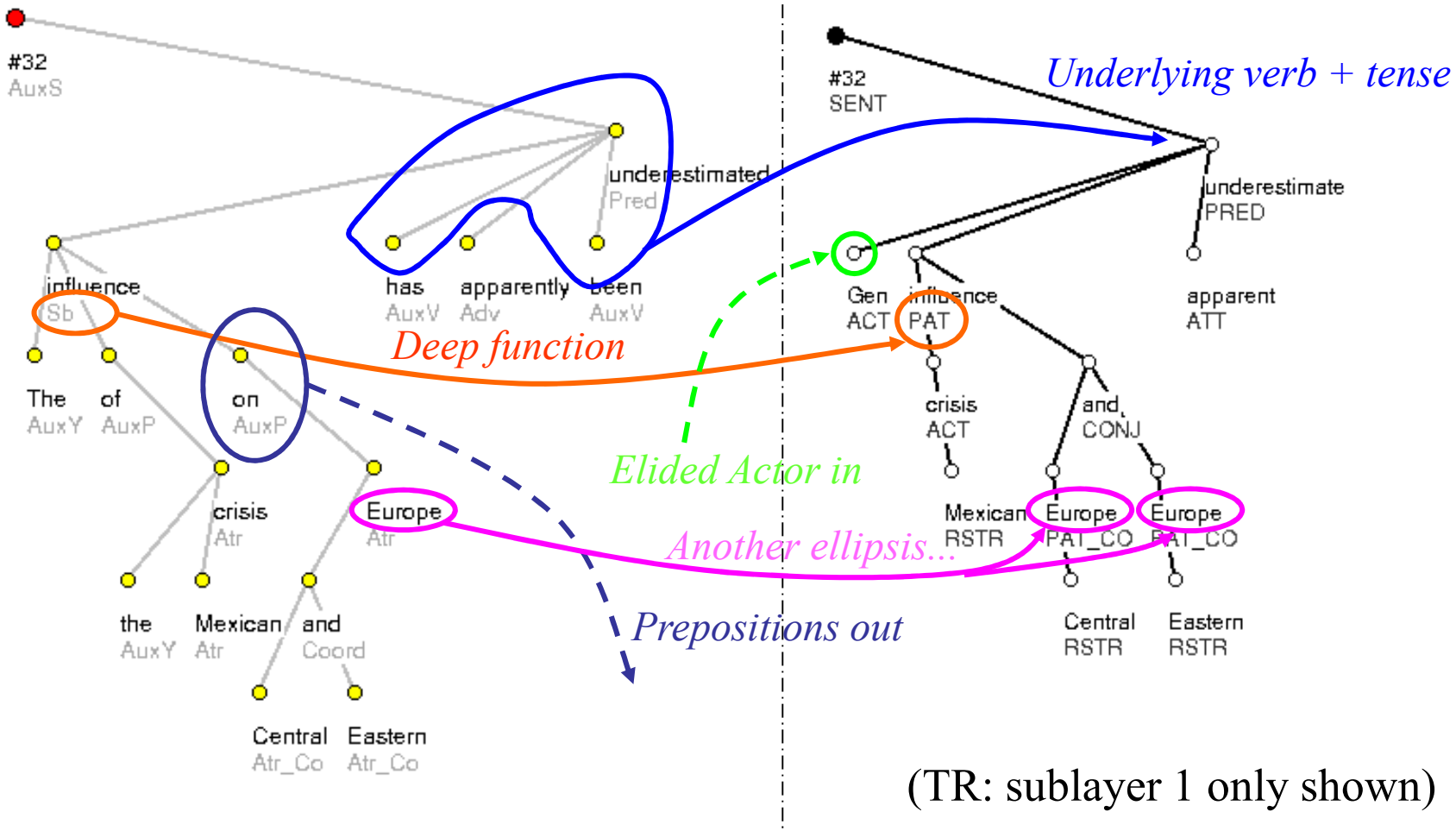


In practice, that procedure will require making of certified copies.

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, ...

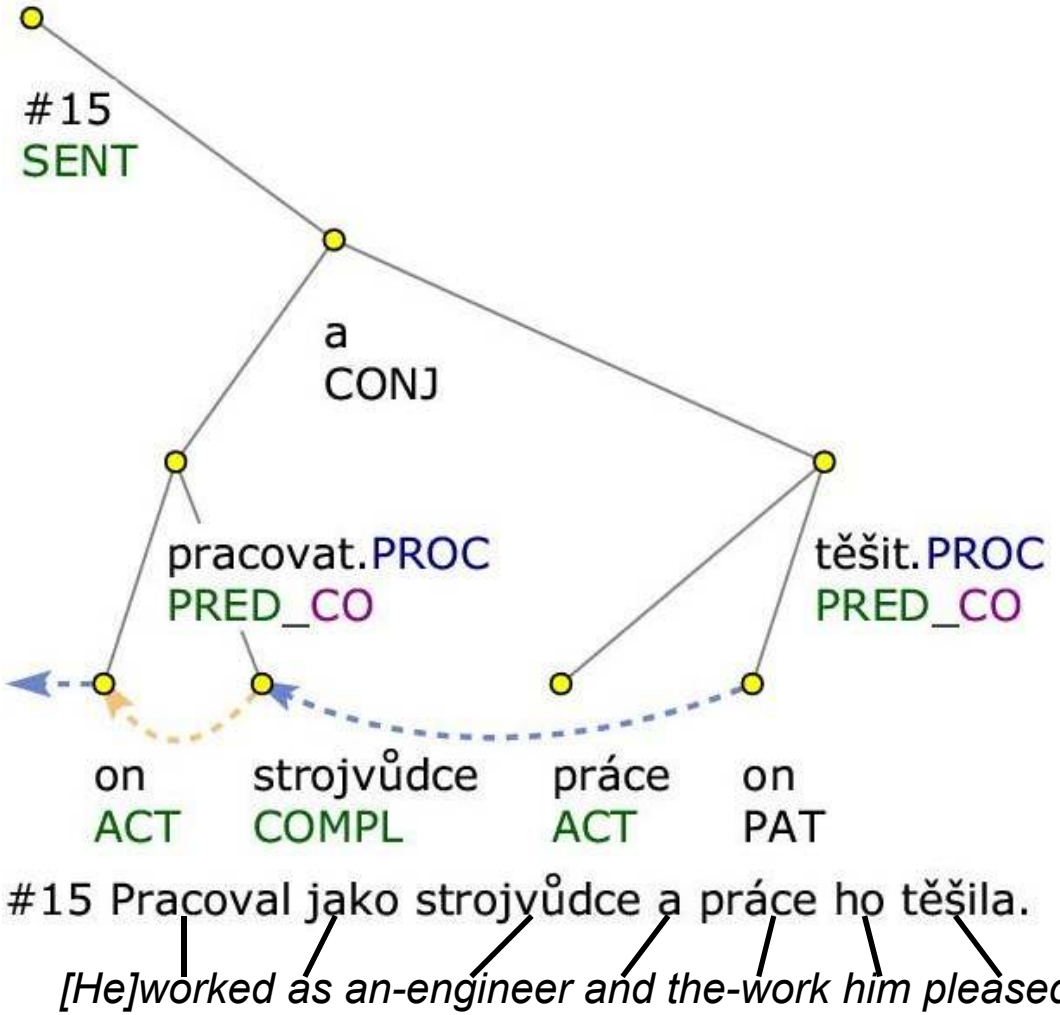
Analytical vs. Tectogrammatical



(TR: sublayer 1 only shown)

Example - TR

- Graphical visualization
- *He worked as an engineer and he liked the work*.



“Dependency” Structure

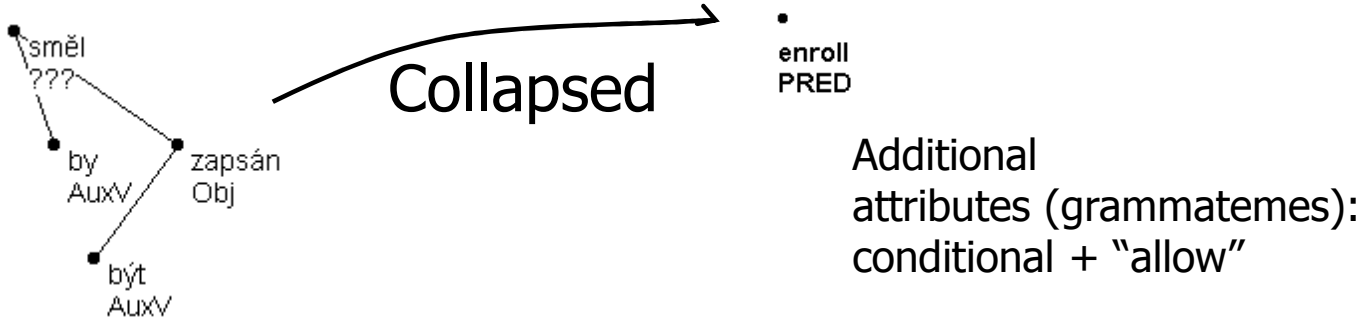
- Similar to the surface (Analytical) layer...
  ...but:
  - certain nodes deleted
    - auxiliaries, non-autosemantic words, punctuation
  - some nodes added
    - based on word (mostly verb, noun) valency
    - (some) ellipsis resolution
  - detailed “semantic” relation labels (***functors***)

Tectogrammatical Functors (Semantic Role Labels)

- "Actants": syntactic ACT, PAT, semantic 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

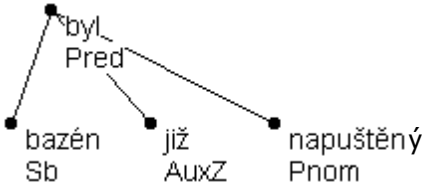
- Analytical verb form:
  - » (he) allowed would-be to-be enrolled
  - » směl by být zapsán

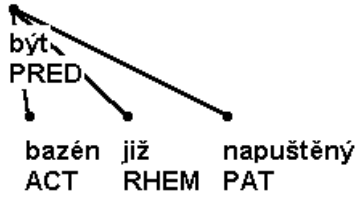


Additional attributes (grammatemes): conditional + "allow"

Tectogrammatical Example

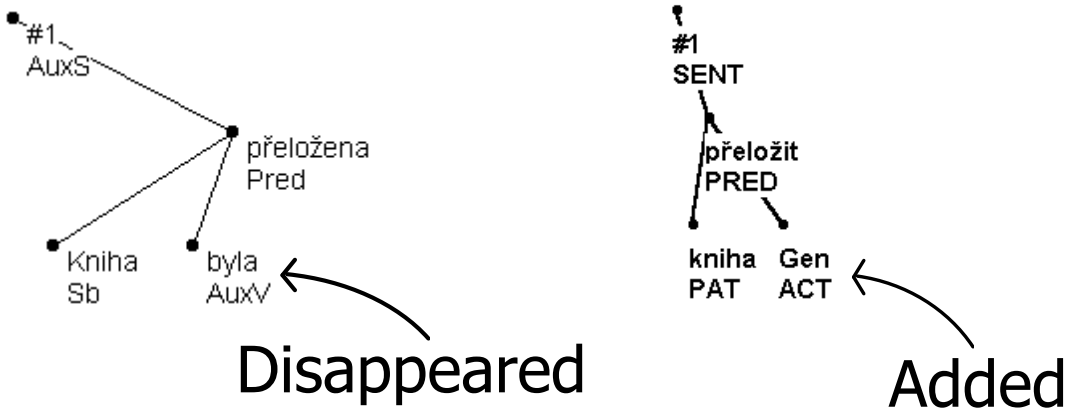
- Predicate with copula (state)
  - » (the) pool has-been already filled
  - » bazén byl již napuštěný





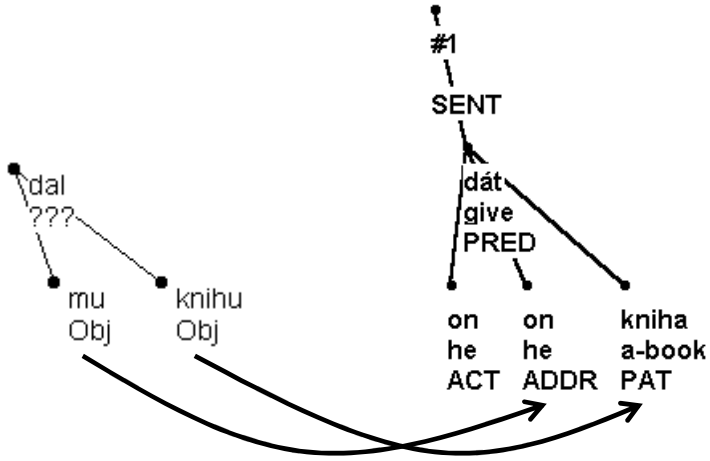
Tectogrammatical Example

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



Tectogrammatical Example

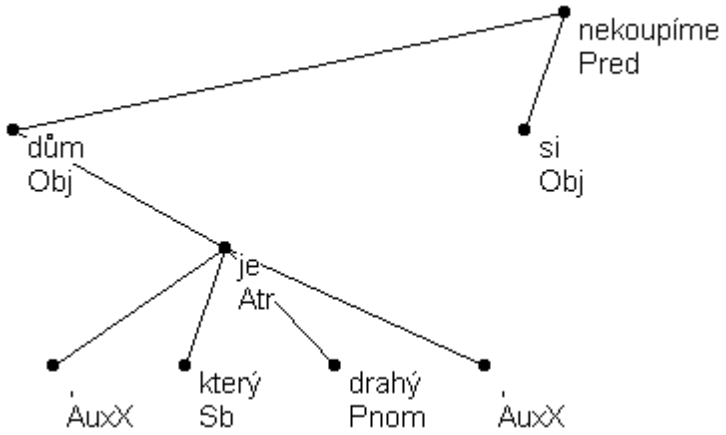
- Object
  - » (he) gave him a-book
  - » dal mu knihu

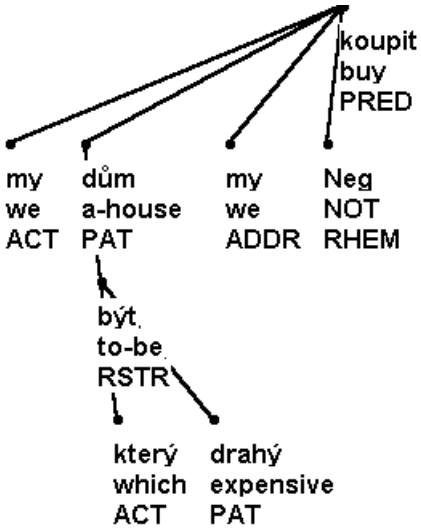


Obj goes into ACT, PAT, ADDR, EFF or ORIG based on governor's valency frame

Tectogrammatical Example

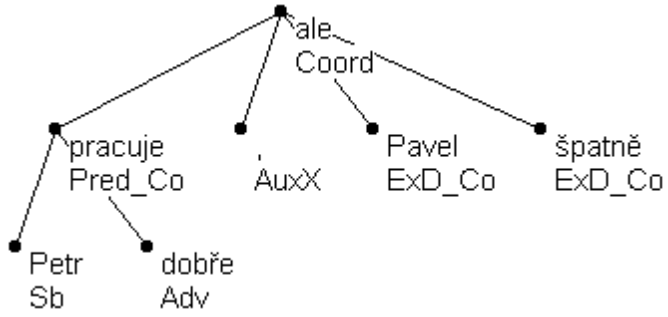
- Relative clause (embedded)
  - (a) house, which is expensive, (we) (to-ourselves) will-not-buy
  - dům , který je drahý , si nekoupíme

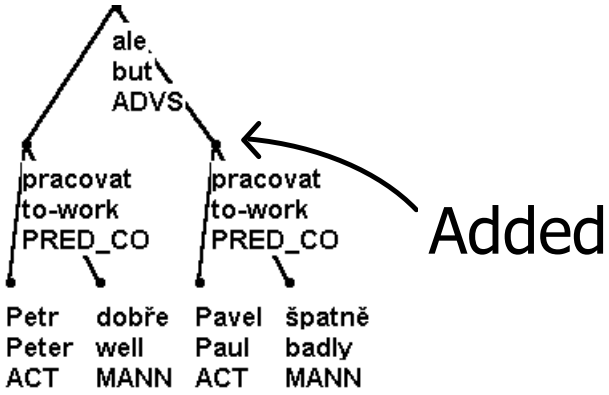




Tectogrammatical Example

- Incomplete phrases
  - » Peter works well , but Paul badly
  - » Petr pracuje dobře, ale Pavel špatně





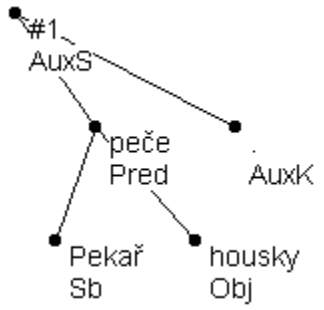
Layer 3: Tectogrammatical

- 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, ...

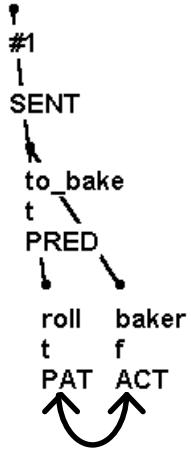
Deep Word Order Topic/Focus

- Example:

Analytical dep. tree:



- Baker bakes rolls. vs. *Baker*^{IC} bakes rolls.



Deep Word Order Topic/Focus

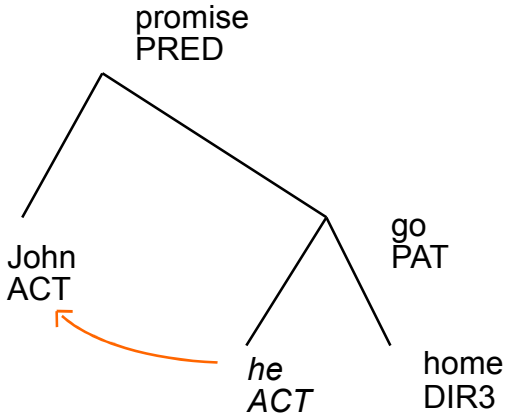
- Deep word order:
  - from “old” information to the “new” one (left-to-right) at every level (head included)
  - projectivity by definition (almost...)
    - i.e., partial level-based order -> total d.w.o.
- Topic/focus/contrastive topic
  - attribute of every node (t, f, c)
  - restricted by d.w.o. and other constraints

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, ...

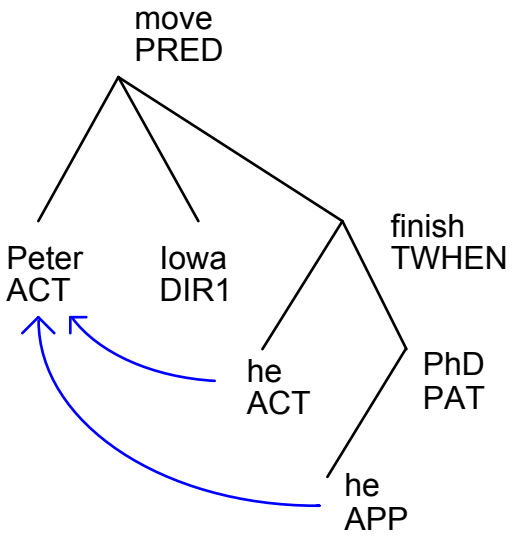
Coreference

- Grammatical (easy)
  - 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.



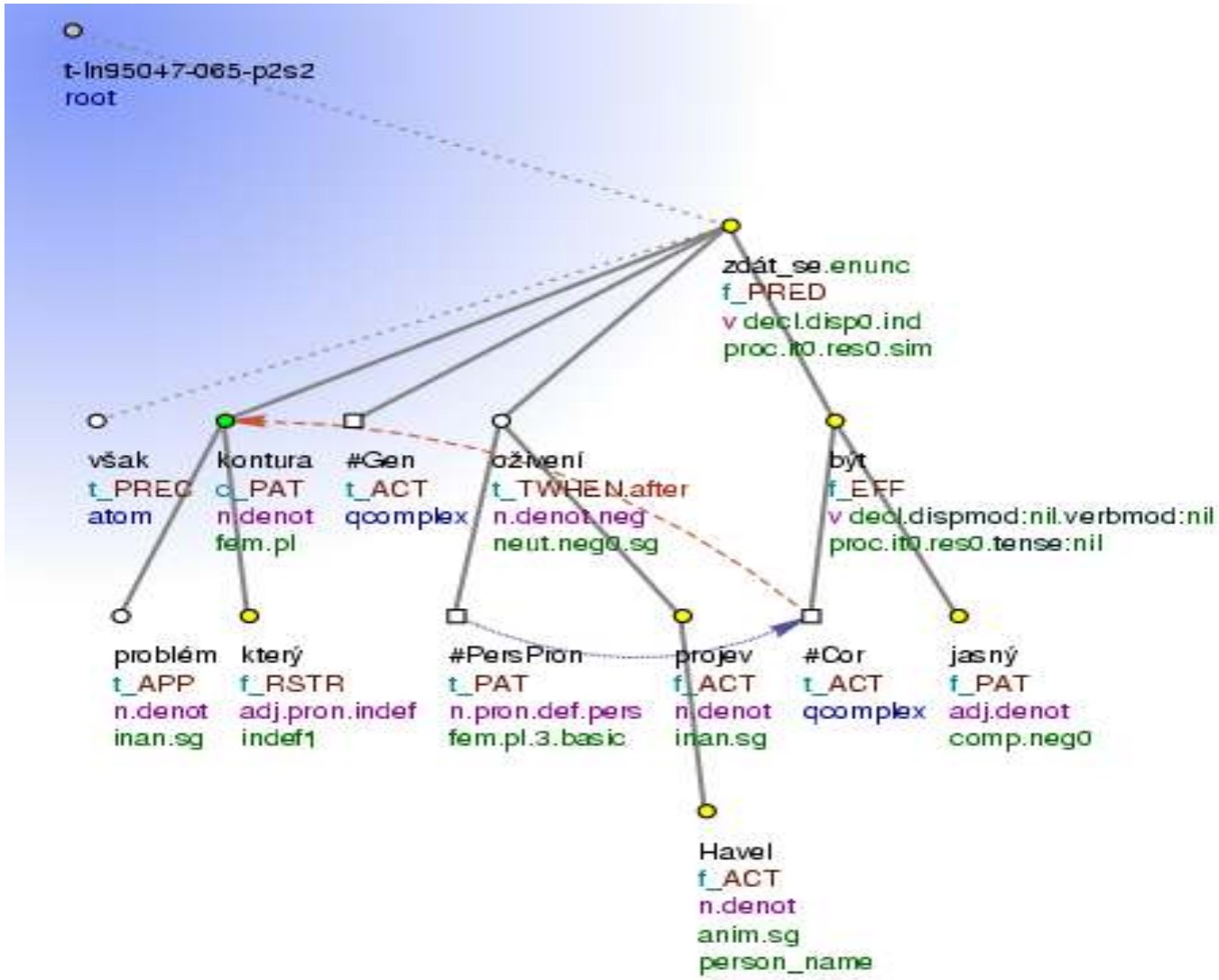
Layer 3: Tectogrammatical

- 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

- 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, ...

Fully Annotated Sentence

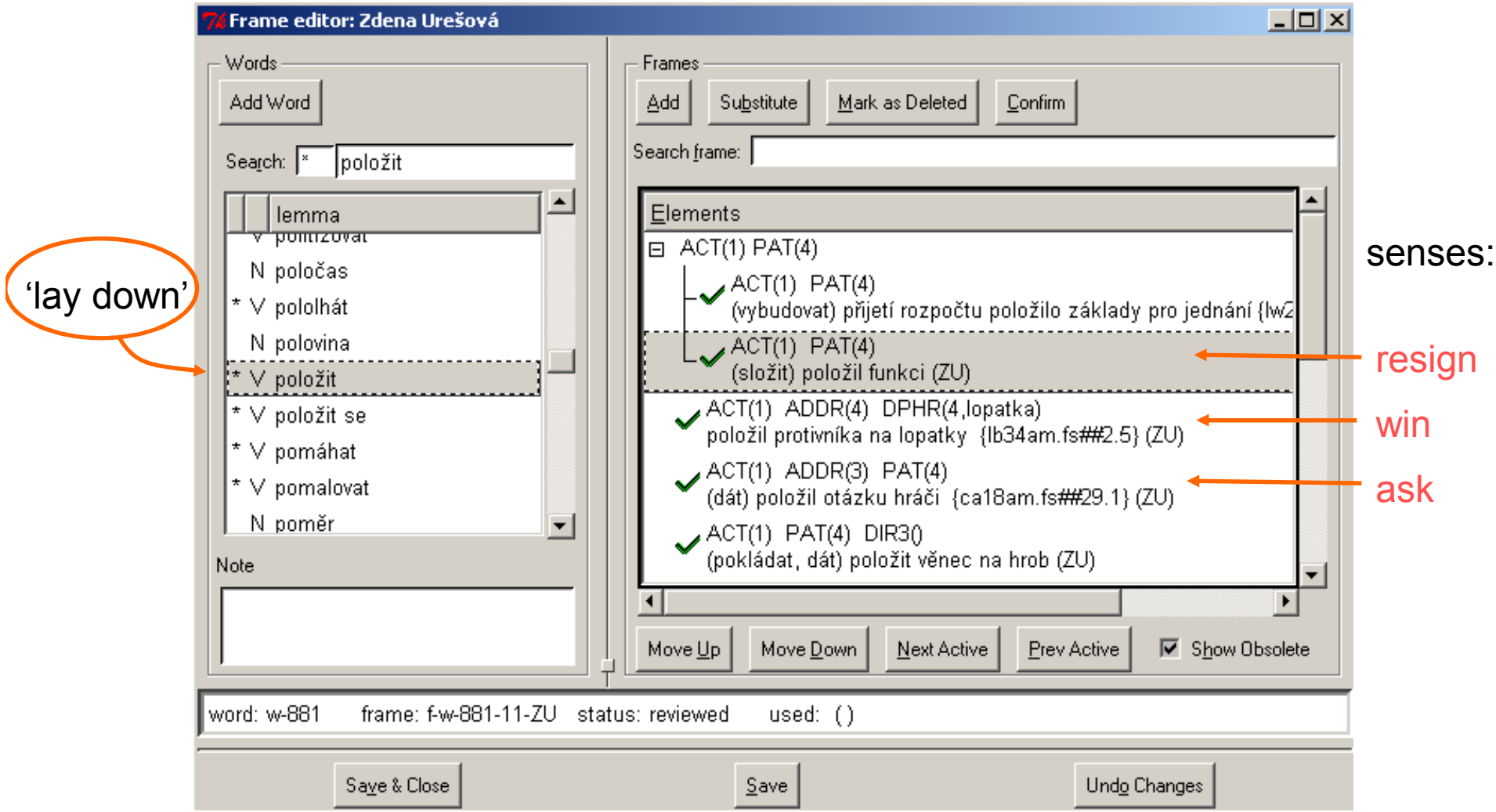


The boundaries of some problems seem to be clearer after they were revived by Havel's speech.

Prague Dependency Treebank & Valency

- Valency in the PDT (~ PropBank in/over PTB)
  - Valency lexicon for PDT
- Properties of valency:
  - Specific for every word meaning (in general)
    - leave: *sb left sth for sb* vs. *sb left from somewhere*
    - same as in PropBank *leave.02* vs. *leave.01*
  - Typically strongly correlates with surface form
    - morphological case (~ ending), preposition+case, ...
- Every verb occurrence annotated
  - incl. verb sense

Example PDT-Vallex entry



Spoken Language Annotation

- Create gold-standard data for
  - (Statistical) training
  - Testing
- Use in machine learning of automatic
  - *speech reconstruction*
  - *(eventually) language understanding*
- Go beyond state-of-the-art
  - ASR Post-Correction / disfluency removal
  - (cf. e.g. Fitzgerald, 2008, or Lopez/Cozar & Callejas, 2008)

What is Speech Reconstruction

... apart from disfluency removal?

and we ‘re sitting on the step of I think it ‘s my aunt Molly ‘s house

Disfluency removal

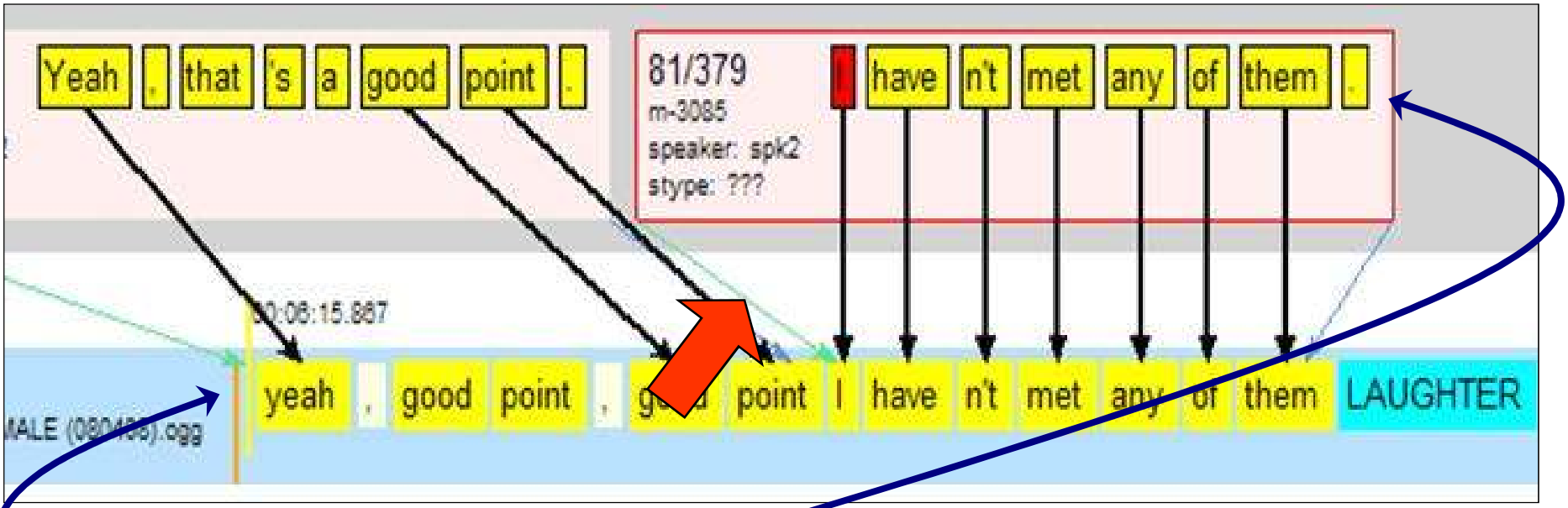
?? I think

We’re sitting on the step of my aunt Molly’s house, .

Speech reconstruction

- original transcript → edited transcript
  - resembling an interview editing for print
  - “standard”, grammatical text -> ***<u>can be treebanked</u>***

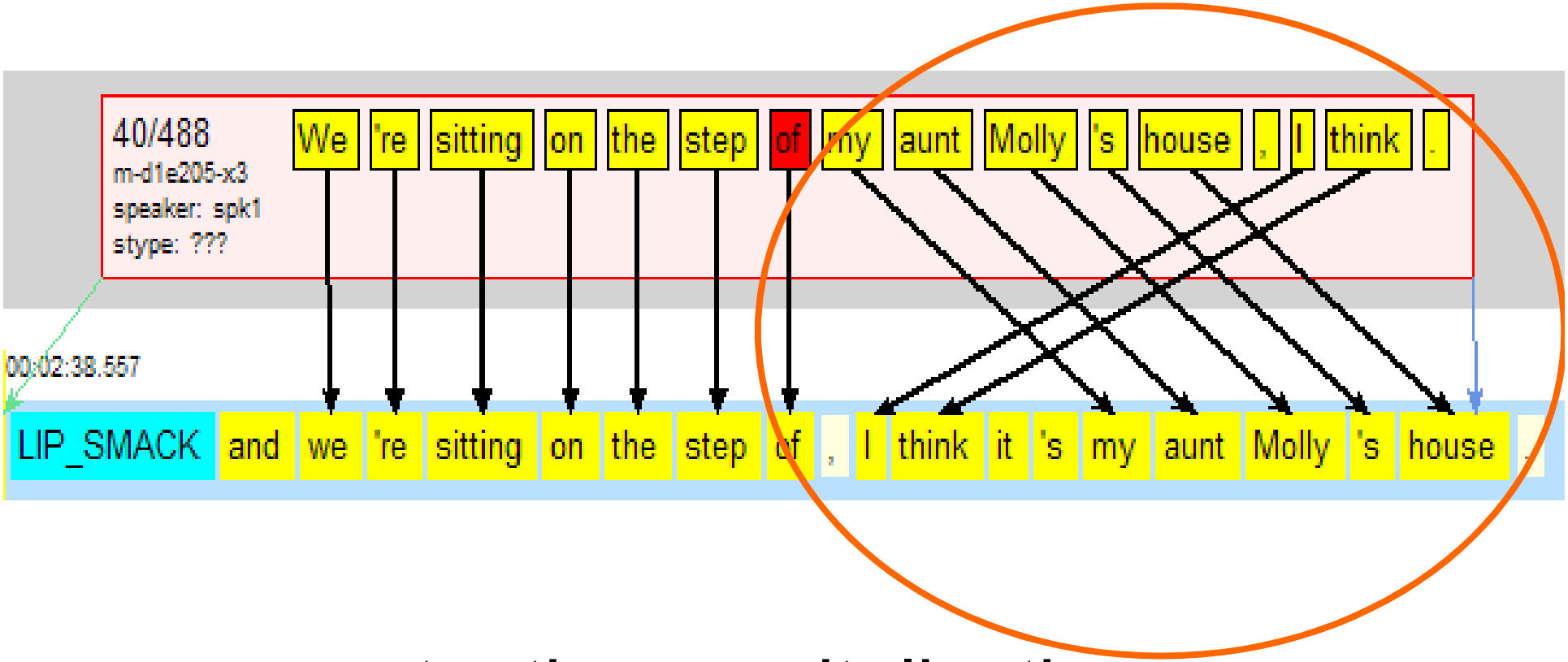
Segment splitting



(reconstruction) segment = sentence

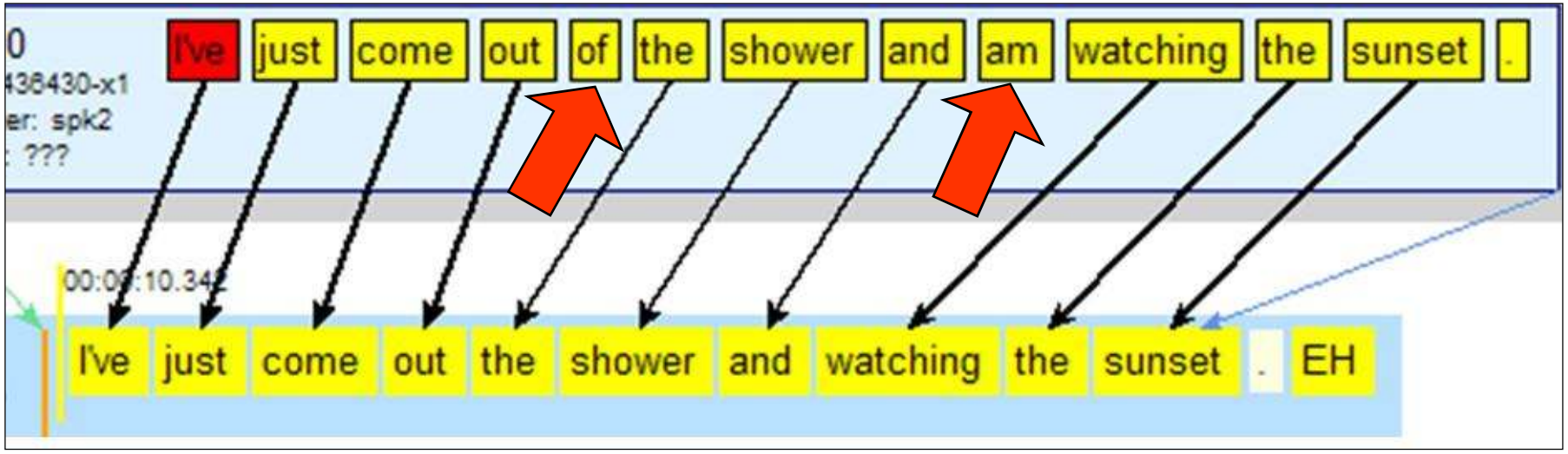
(transcript) segment ~ non-silence span

Word order, deletions, insertions, ...



..., punctuation, capitalization, ...

Function word insertion



Czech Example

Original transcription:
ale taky důvod byl ten že škodováci byli hrozně rádi bejvali kdybych tam byla mohla nastoupit k ni - k nim jako do zaměstnání

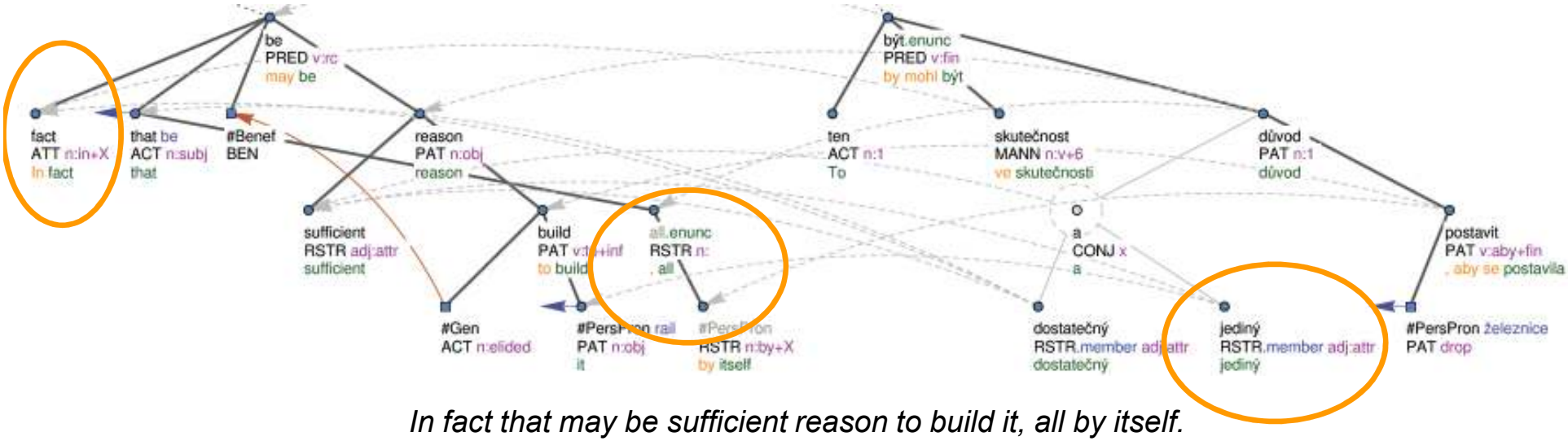
Recovering punctuation and capitalization:
Ale taky důvod byl ten , že škodováci byli hrozně rádi bejvali , kdybych tam byla , mohla nastoupit k ni k nim jako do zaměstnání

Automatic reconstruction:
Ale taky důvod byl ten , že škodováci bývali byli hrozně rádi , kdybych tam byla , mohla nastoupit k ní , k nim jako do zaměstnání

Reference reconstructions:
(a) Ale důvod byl taky ten , že "škodováci" by bývali byli hrozně rádi , kdybych tam k nim mohla nastoupit do zaměstnání .
(b) Důvod byl ale taky ten , že škodováci by bývali byli hrozně rádi , kdybych tam byla mohla nastoupit k nim jako do zaměstnání .

Czech-English Structural Correspondence

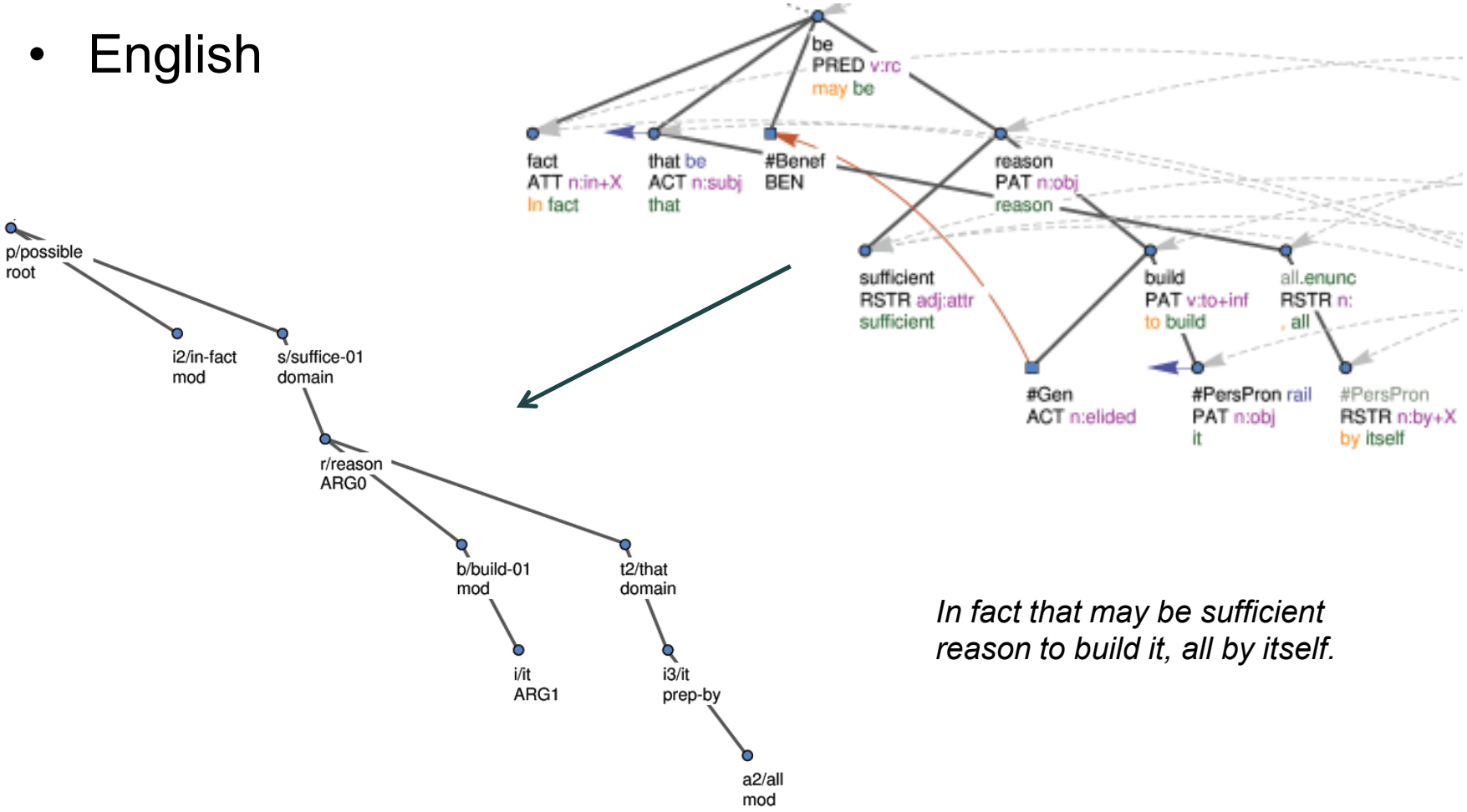
- Tectogrammatical representation



In fact that may be sufficient reason to build it, all by itself.

TR vs. AMR

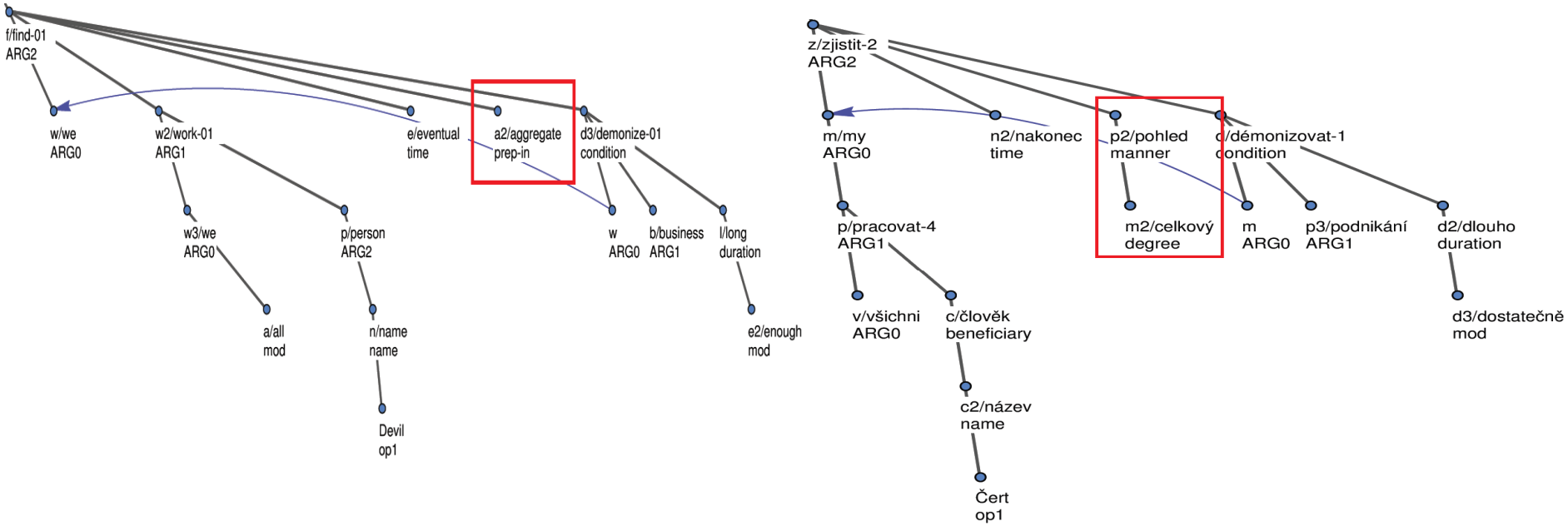
- English



In fact that may be sufficient reason to build it, all by itself.

Czech – English Structural Correspondence

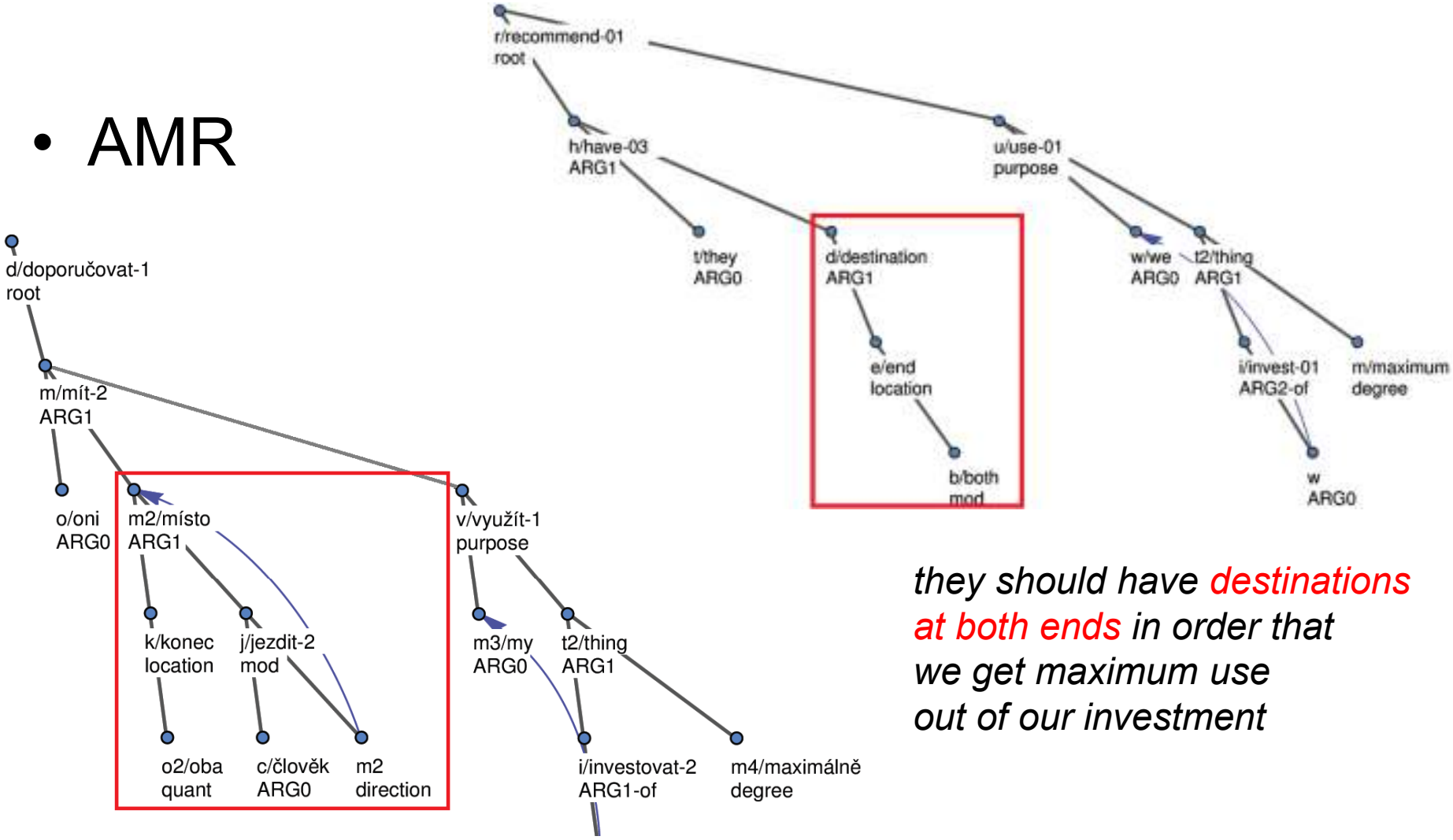
- AMR



but in the aggregate if we demonize business long enough we will eventually find out we all work for the Devil .

Structural Differences

- AMR

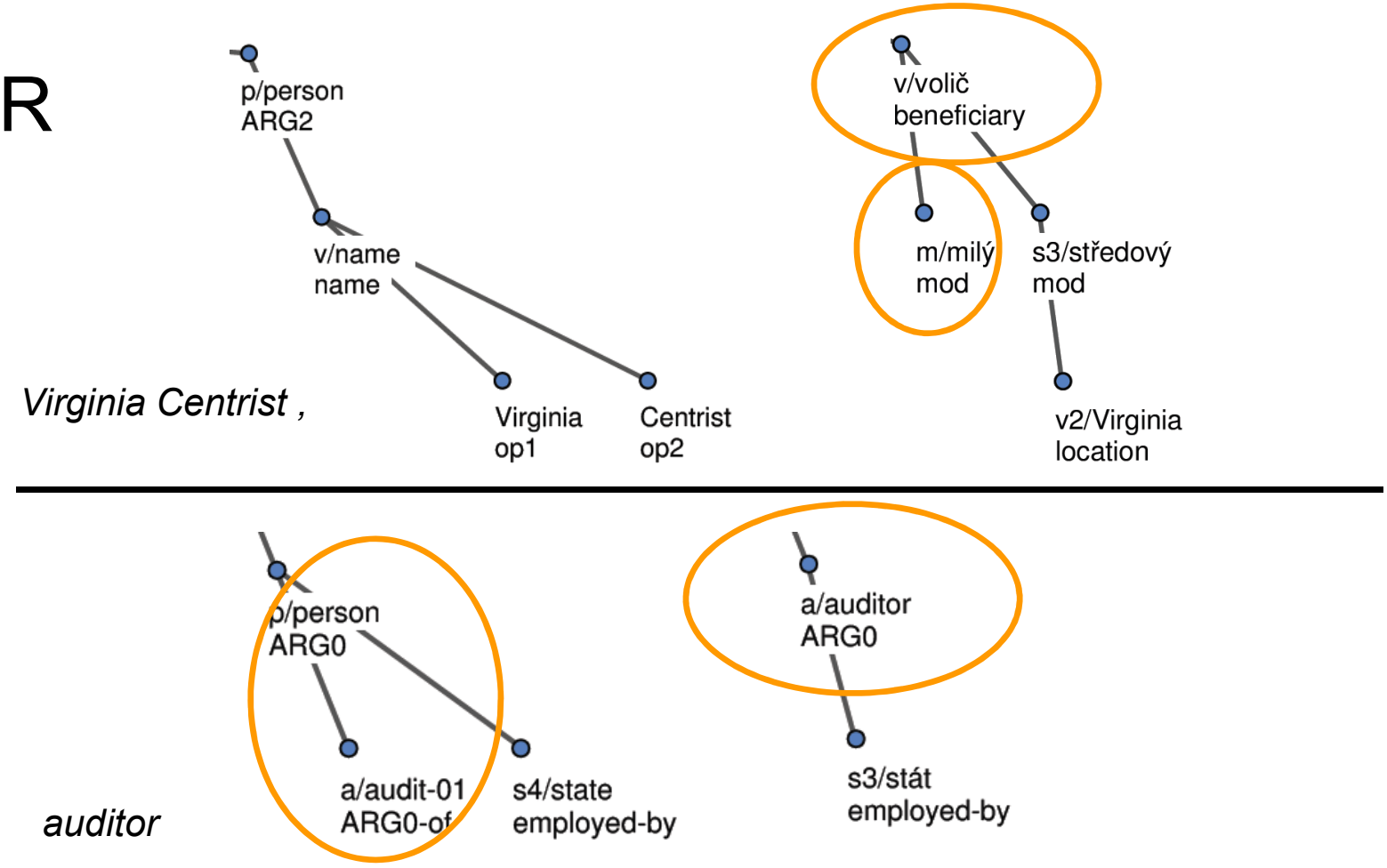


they should have destinations at both ends in order that we get maximum use out of our investment

na/*at* obou/*both* koncích/*ends* místa/*places*,
kam/*where* lidé/*people* jezdí/*go*

Structural Differences

- AMR



The AMR Differences in Numbers

- 100 sentences annotated (1215 AMR nodes)
  - Differences (manually) classified

Same structure	Different substructures	Local difference only	Relation differences	Reference differences
29 (sents)	193 (subgraphs)	92 (subgraphs)	331 (nodes)	37 (nodes)
of 100	of approx. 800[2]	of 193 (all diffs)	of 1215 Cz nodes	of 1215 Cz nodes
29 %	approx. 25 %[2]	47.7 %	27.2 %	3.0 %

Table 1: Number and percentages of differences in the annotated data

- Disregard local differences?
  - ... +18 sentences would match structurally
  - 29 + 18 = 47 (almost half)

Pointers

- PDT 2.0 (the “Original”)
  - http://ufal.mff.cuni.cz/pdt2.0
- PCEDT
  - http://ufal.mff.cuni.cz/pcedt (1.0, now obsolete)
  - http://ufal.mff.cuni.cz/~toman/pcedt (preview)
- PEDT
  - English side of PCEDT, additional: NE, coreference
  - http://ufal.mff.cuni.cz/~toman/pedt (preview)
- PADT (Arabic, morphology + surface syntax)
  - http://ufal.mff.cuni.cz/padt
- PDTSC (spoken Czech , multiple speech reconstruction annotation)
  - http://ufal.mff.cuni.cz/~toman/pdtsc (preview)
- PDTSE (spoken English, multiple speech reconstruction annotation)
  - http://ufal.mff.cuni.cz/~toman/pdtse (preview)
- LDC catalog numbers:
  - LDC2006T01 (PDT 2.0), LDC2004T23 (PADT 1.0), LDC2004T25 (PEDT 1.0)
- CoNLL 2009 shared task (7 languages, surface syntax + predicate arguments only)
  - http://ufal.mff.cuni.cz/conll2009-st