Layers of Meaning Representation in a Dependency Tradition

Jan Hajič
Charles University, Prague
Faculty of Mathematics and Physics
Computer Science School
Institute of Formal and Applied Linguistics (ÚFAL)
& LINDAT/CLARIN Research Infrastructure

Adapted from previous presentations about PDT and PDT-AMR comparison by J. Hajič, Z. Urešová, E. Hajičová and J. Mírovský.

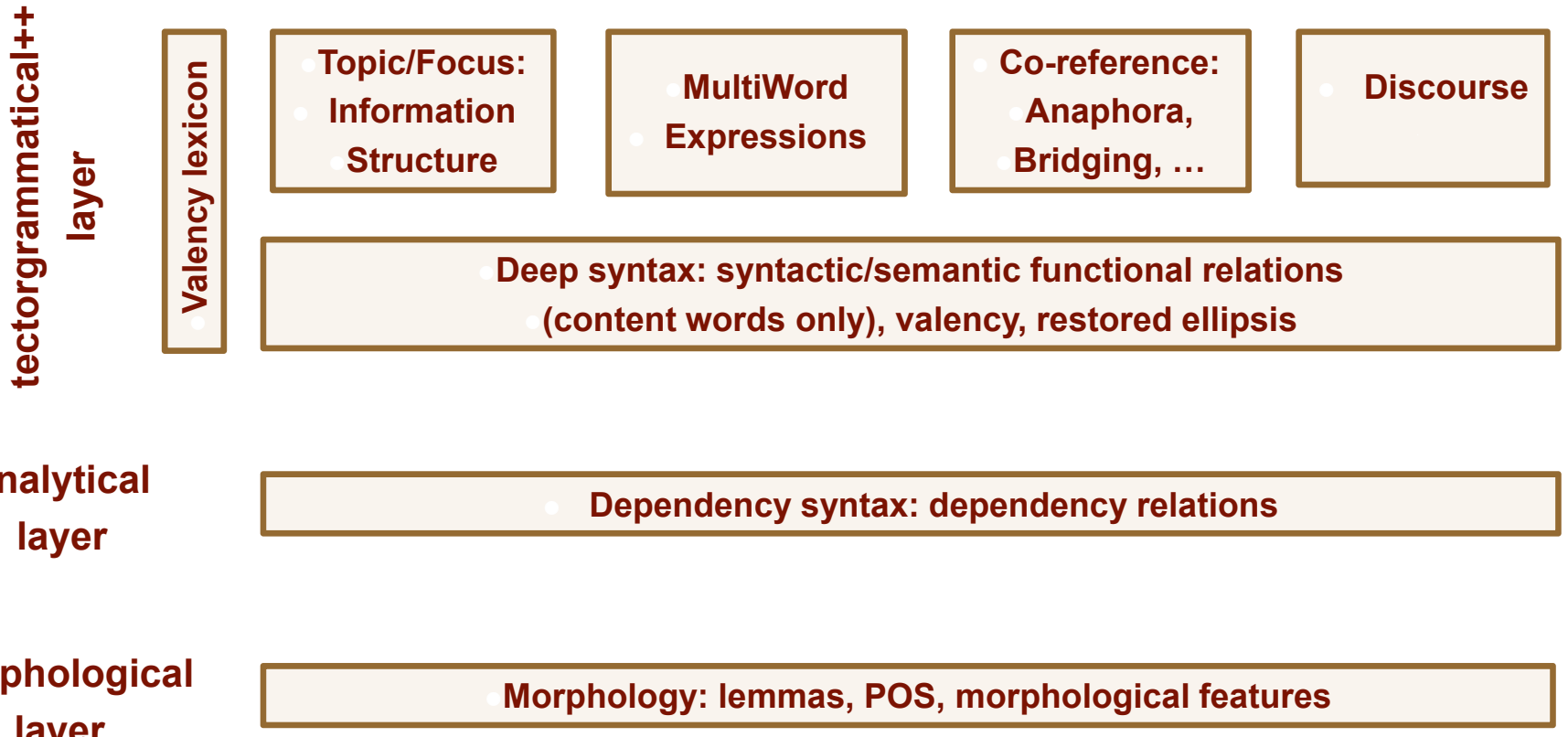
Outline

- Prague Dependency Treebank Family
  - The FGD theory
  - Data: Prague Dependency Treebank (PDT)
- Three fundamental layers of annnotation
  - Morphology, syntax, deep syntax (tectogrammatics)
- Tectogrammatical layer
  - Core structure, deep dependency relations, valency
  - Coreference, information structure, discourse
- Comparison to AMR, Cross-lingual comparison
- Summary

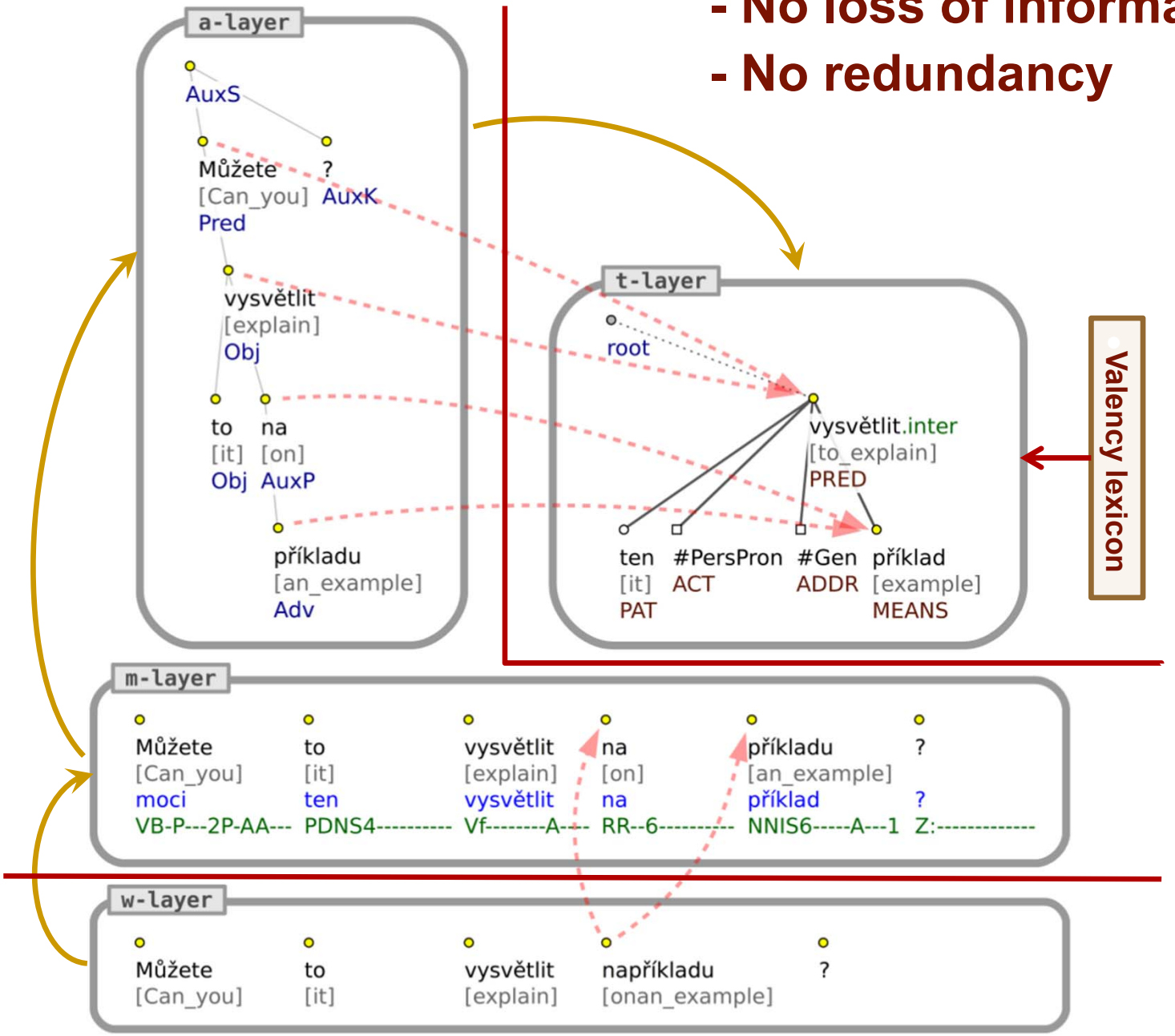
Prague Dependency Treebank(s) (PDT)

- Manual annotation of Czech, written texts
  - Morphology
  - Surface (dependency) syntax
  - Deep syntax/semantics (“tectogrammatics”)
  - Information structure, Discourse, Coreference (incl. bridging)
  - MWE, word senses
- Charles University in Prague, ÚFAL
  - ~60 people, since 1996
  - Latest version: http://ufal.mff.cuni.cz/pdt3.5/
- Purpose
  - apply and test machine learning methods
  - test and preserve the linguistic theory
- Additional treebanks, same annotation style
  - Parallel Czech-English, Spoken Czech, Spoken English, Arabic

Representation Layers



- No loss of information
- No redundancy



Abstraction / formalization

Morphological Attributes

Ex.: nejnezajímavějším
"(to) the most uninteresting"

Tag: 13 POS + morph. features

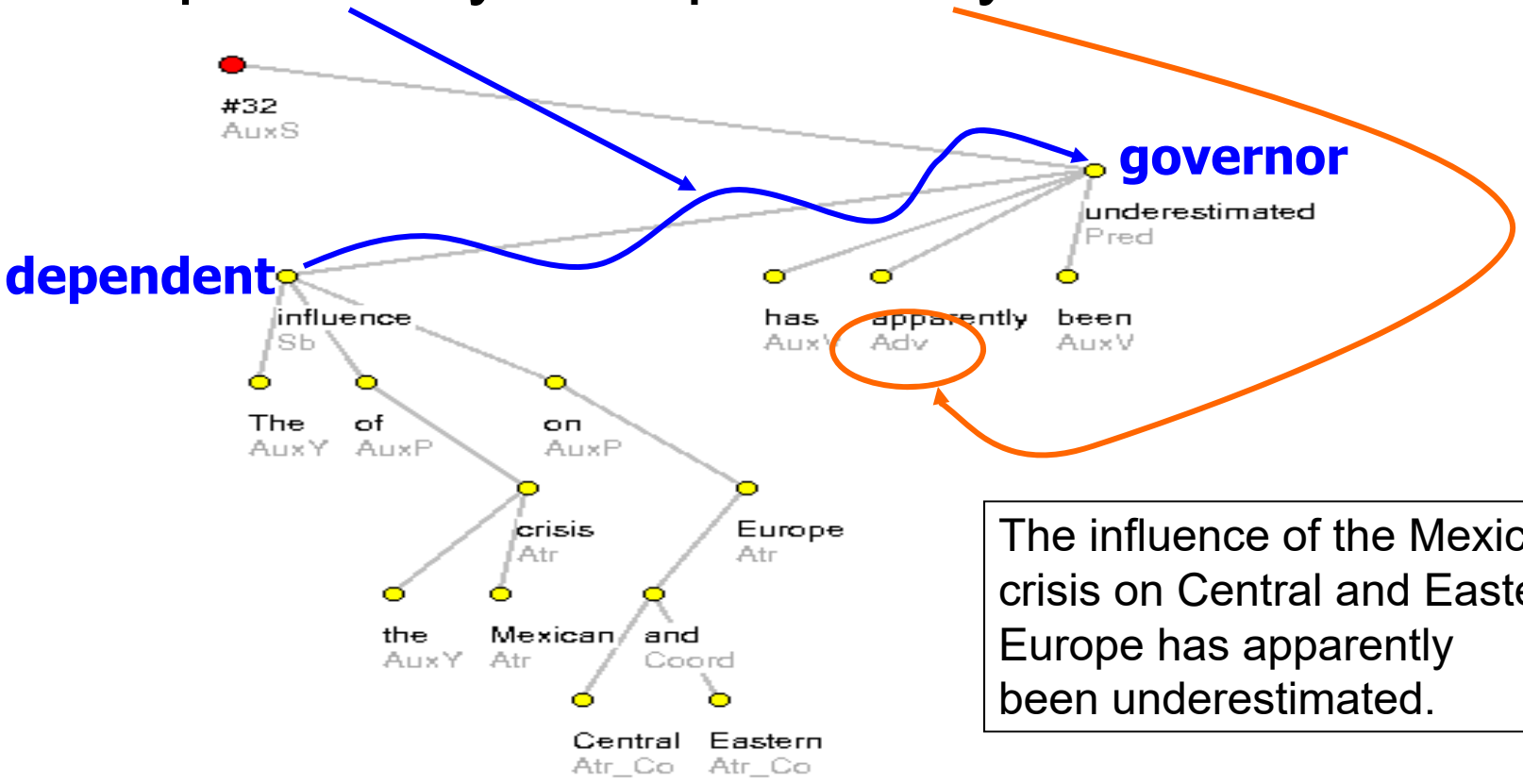
Example: AAFP3----3N----

Lemma: POS-unique identifier

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

Dependency Syntax

- Dependency + Dependency Relation

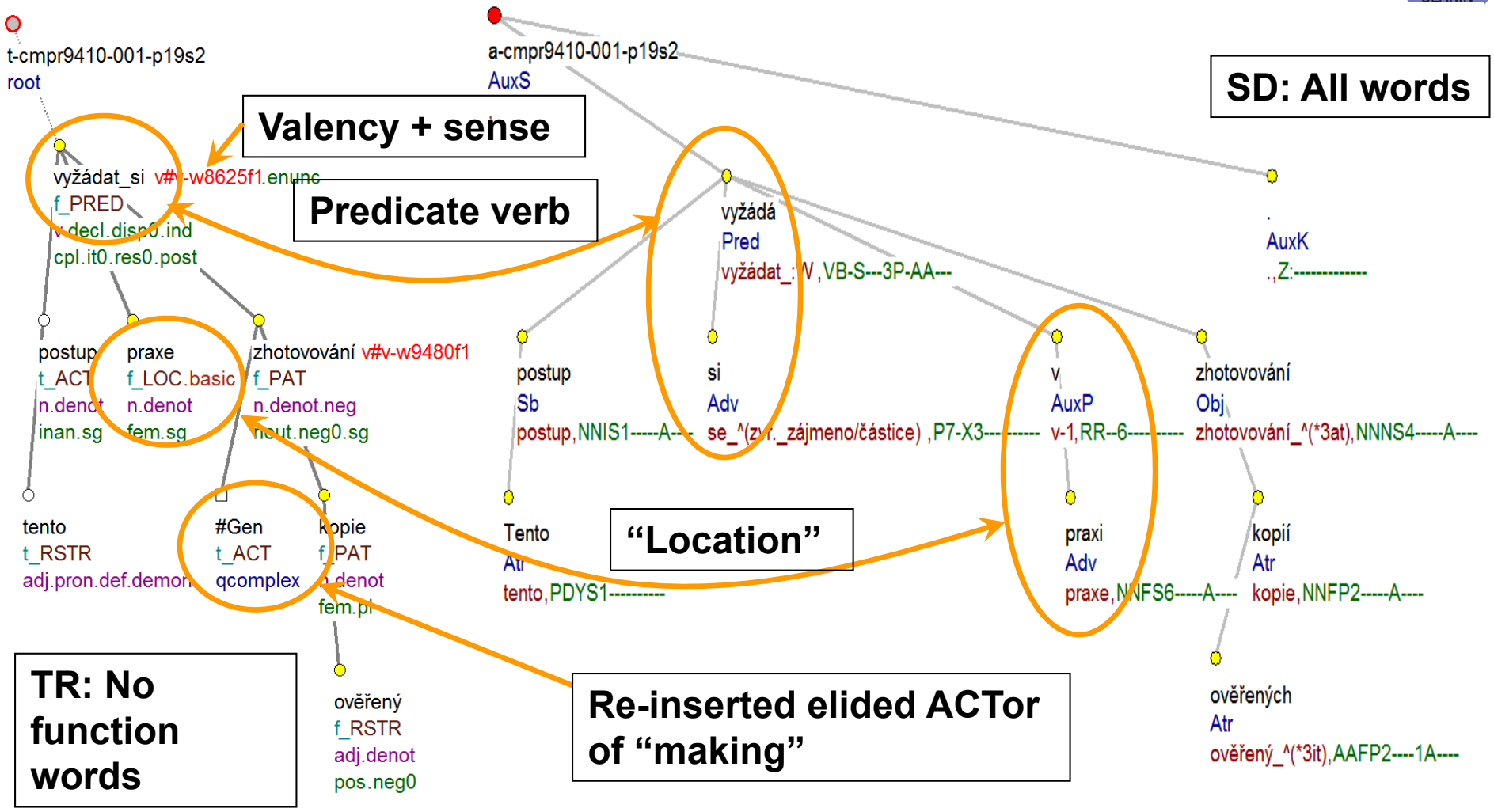


The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated.

Tectogrammatical Meaning Representation

- "Underlying" (deep) syntax
  - 5 sublayers (integrated):
    - dependency structure, (detailed) functors
      - valency annotation
    - topic/focus and deep word order
    - coreference (grammatical, textual, bridging, …)
    - discourse (Penn Discourse Treebank style)
    - all the rest (grammatemes):
      - detailed "functors"
      - underlying gender, number, MWEs, Wordnet senses...
  - Total: >40 features (vs. 5 at m-layer, 2 at a-layer)

Structure, “deep” dependency relations, link to valency lexicon



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

Valency in the PDT

Main principles:

- Every **"autosemantic"** word
  - subcategorization requirements
- Expressed in the **valency frame** of the word
- Valency slots labeled by functors

[type of dependency]: inner participants (~arguments)
free modifications (~adjuncts)

[governing-verb specific]: obligatory vs. optional

- Each valency frame ~ one sense of the verb
  - …with the usual caveats (polysemy, formal problems, …)
- [Argument shifting [criterion for distinguishing arguments]]

A Valency Frame in PDT-Vallex

Structure:

	obligatory	optional
argument		
adjunct		

Contents:

- functor (dependency relation)
- obligatoriness
- surface form

one meaning of the word → one valency frame *(… almost always, except for formal representation difficulties)*

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

CzEngClass: verbal synonym classes

- In progress, ~200 classes so far (Coling'18)
  - Based on semantic role mapping to valency slots

"complain"	Complainer	Addressee	Complaint
complain	ACT	ADDR	PAT
gripe	ACT	ADDR	PAT
grumble	ACT	ADDR	PAT
brblat	ACT	LOC	PAT
postěžovat si	ACT	ADDR	PAT
stěžovat si	ACT	ADDR	PAT & EFF

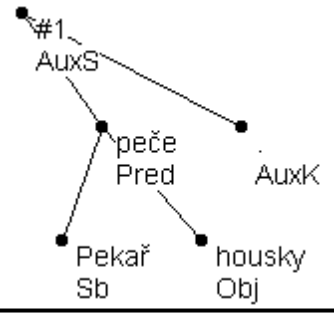
He.ACT complained to her.ADDR that her son lies. PAT

He.ACT complained to her.ADDR about her son.PAT that he lies.EFF

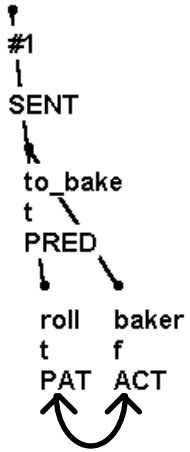
Information Structure: Topic/Focus

- Example:

Syntactic dep. tree:



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

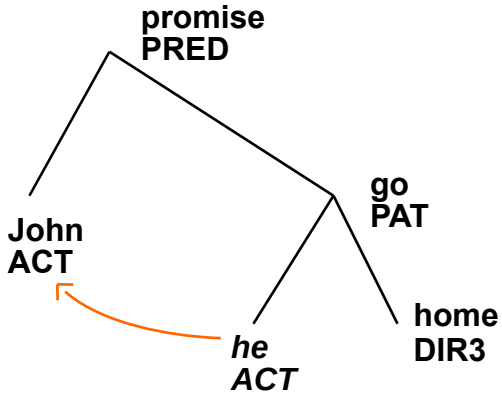


Information Structure: Deep Word Order

- 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

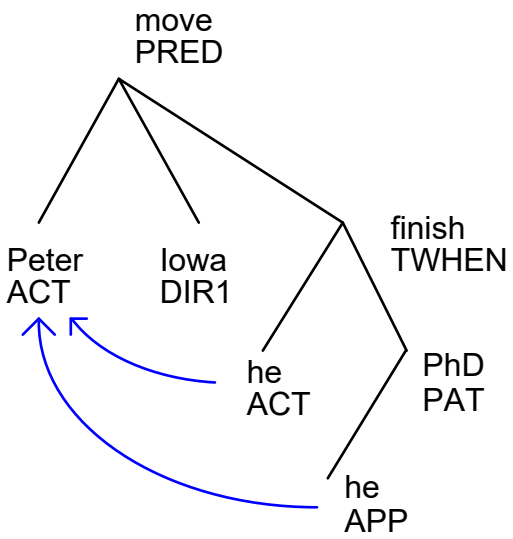
Coreference

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



Coreference

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

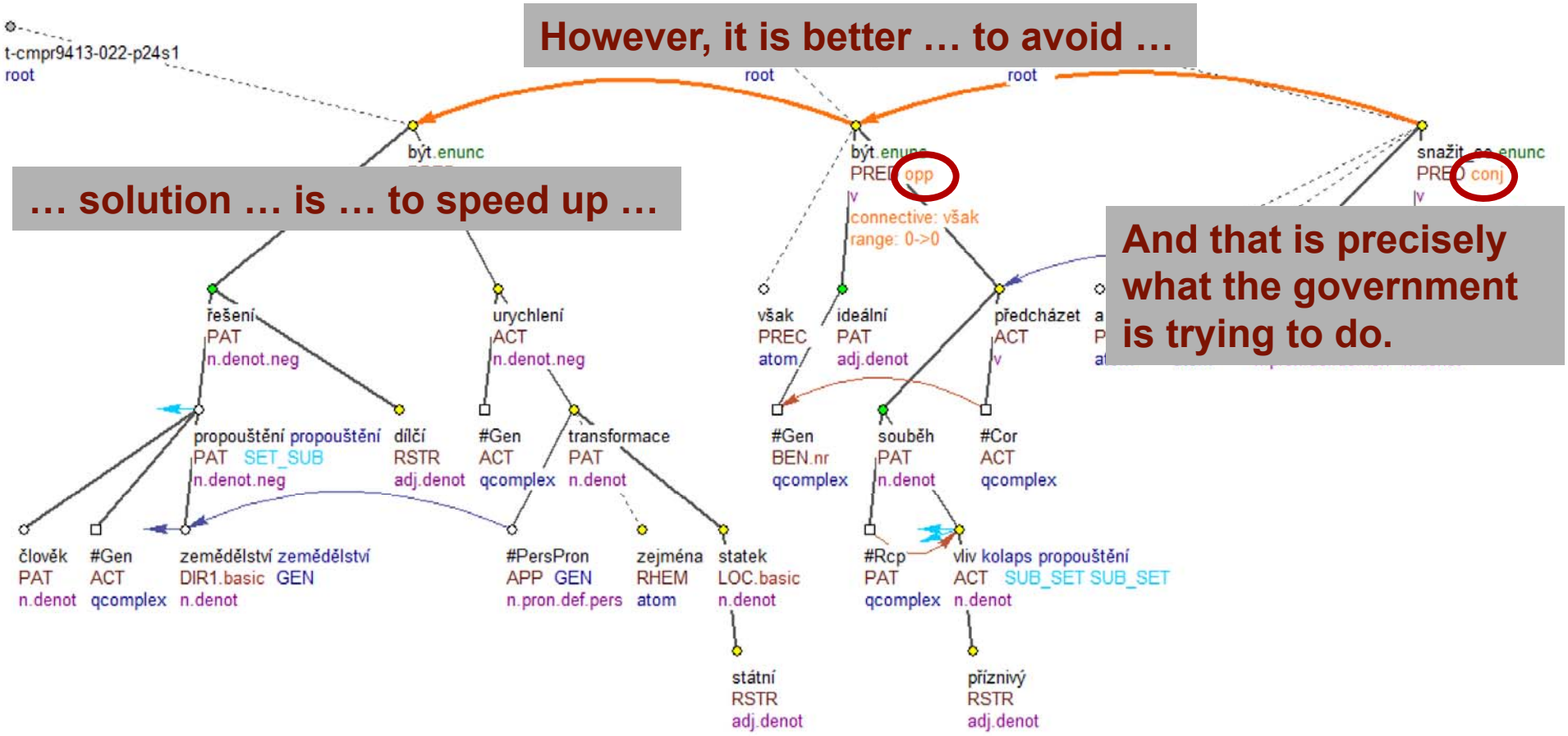


Coreference

- Bridging
  - Ex.: After the accident, they had to repair the front of the car. But the doors were intact.

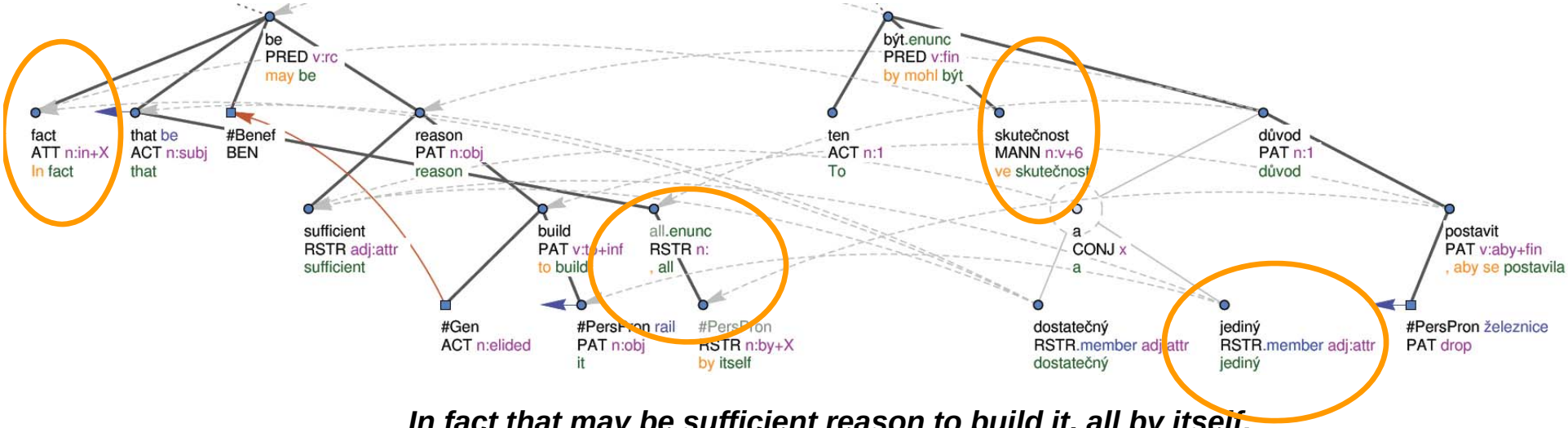
  - Subtypes:
    - whole x part
    - set x element(s) of a set
    - object x function (team – coach)
    - pragmatic contrast (this year – last year)
    - specific relations (author – piece of work)

Discourse annotation
(~ Penn Discourse Treebank)



Cross-lingual Comparison: TR

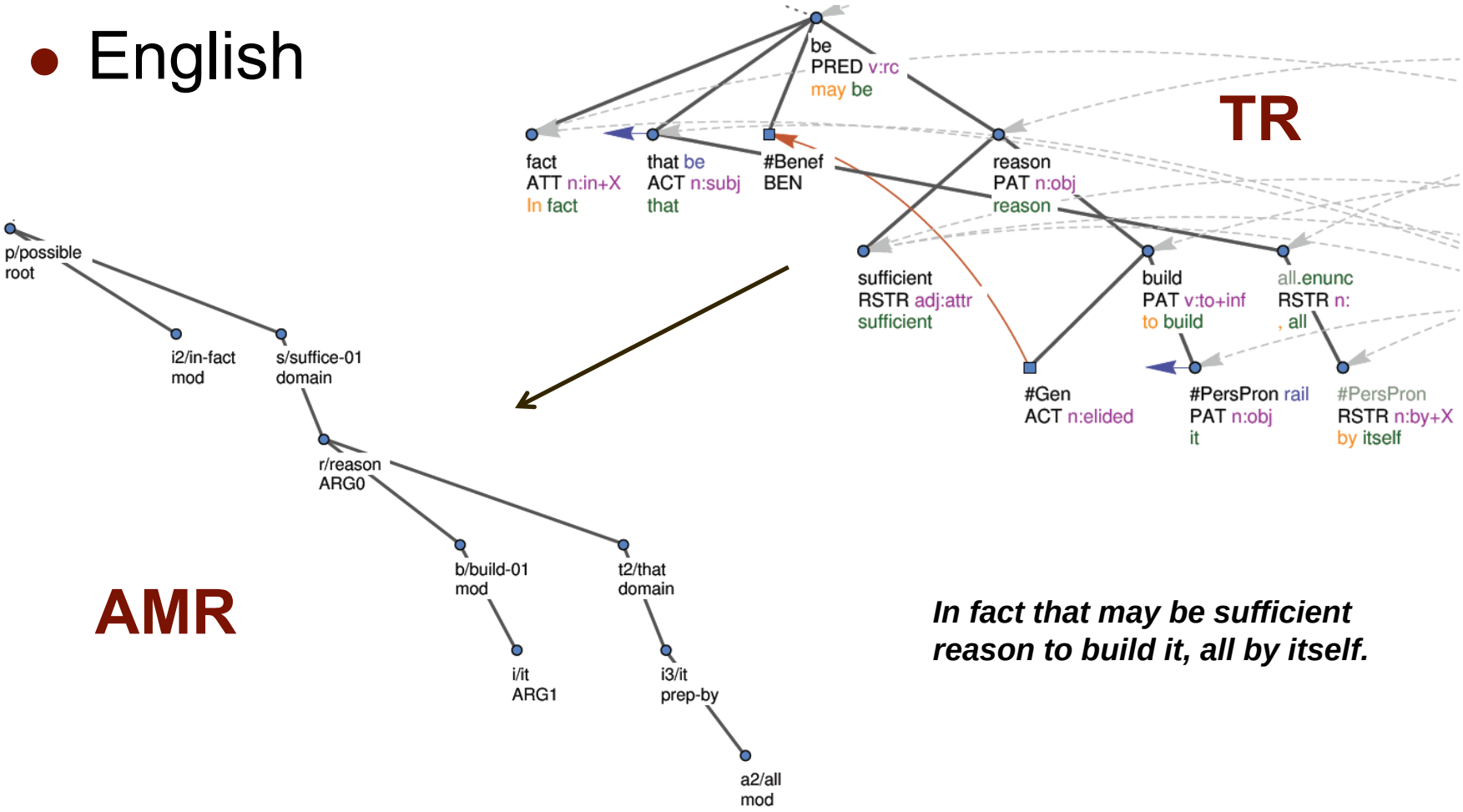
- Example from the Czech-English parallel corpus (PCEDT, WSJ translation to Czech)



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

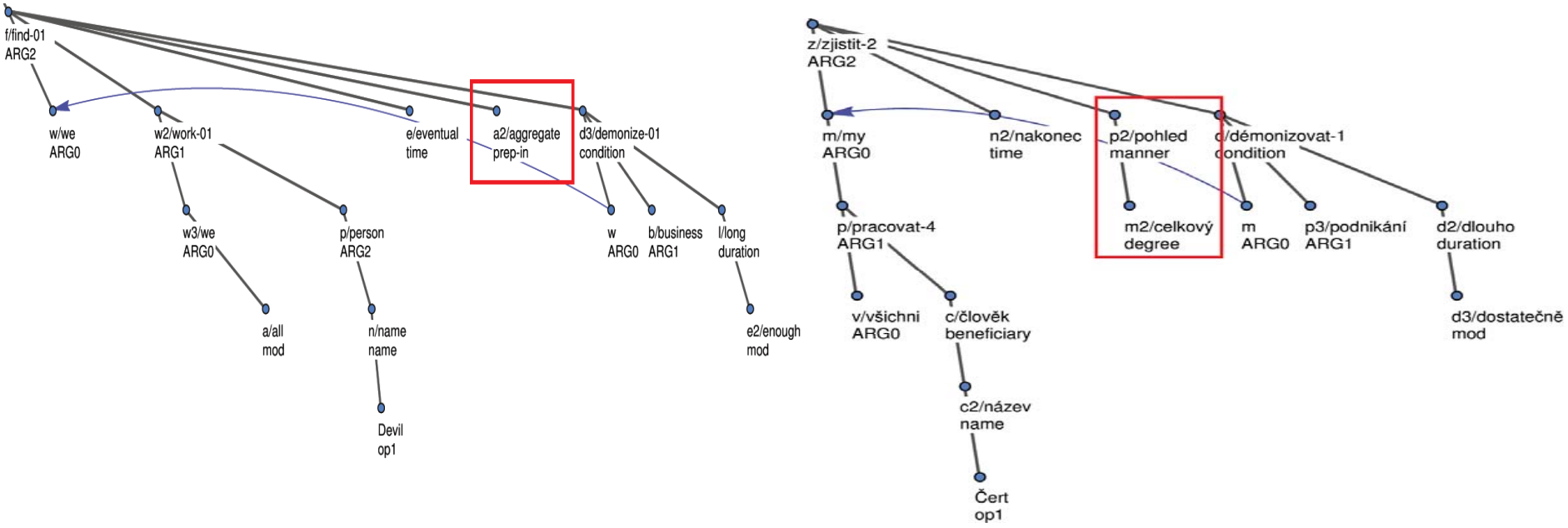
PDT Tectogrammatical representation vs. AMR

- English



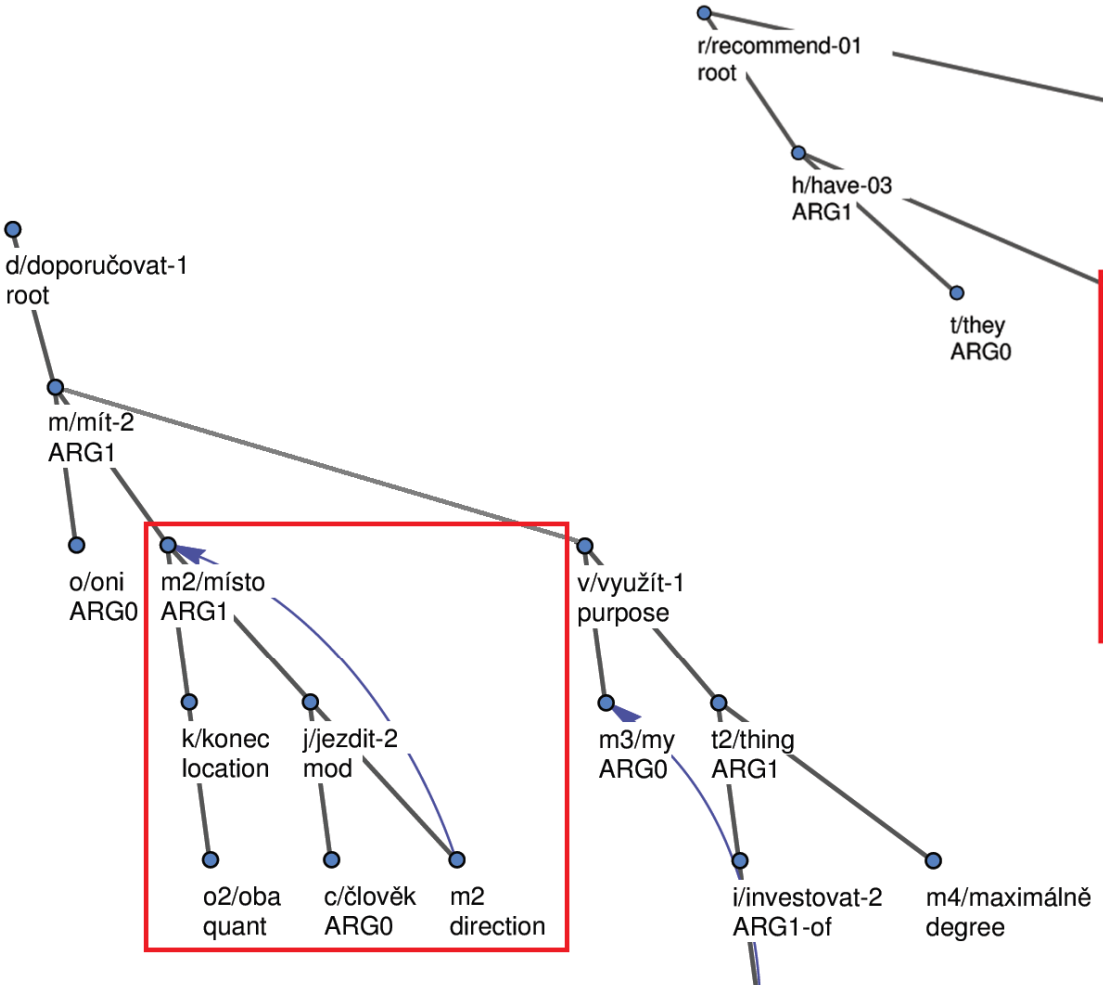
In fact that may be sufficient reason to build it, all by itself.

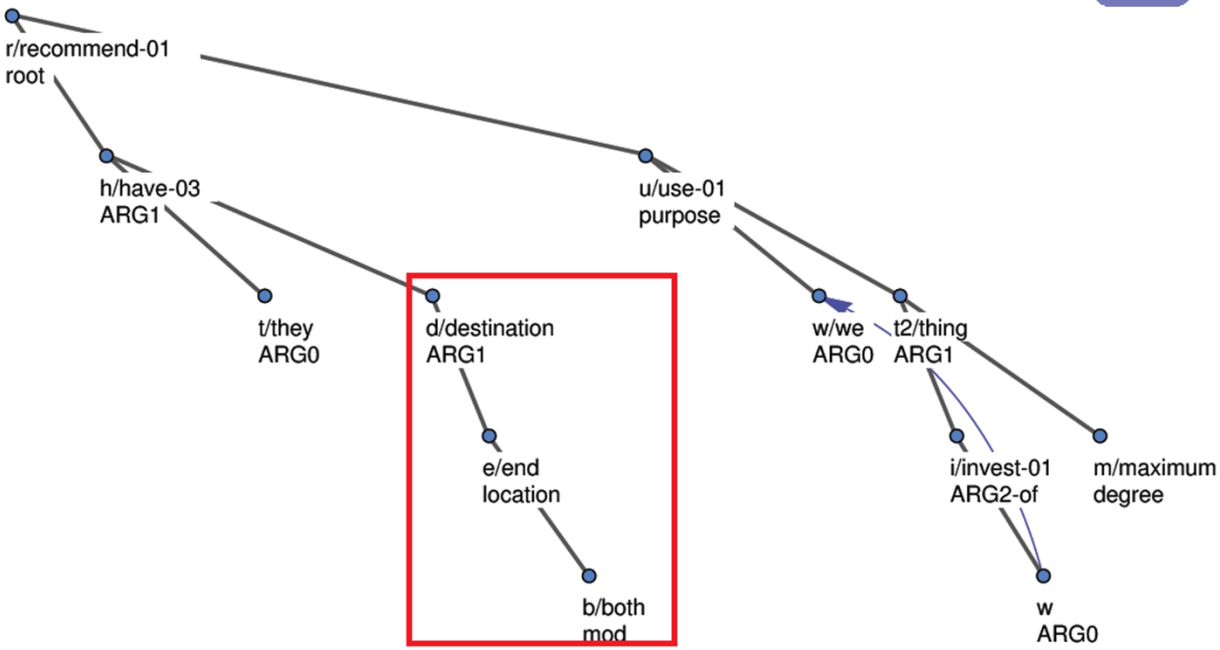
Cross-lingual comparison: AMR



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

Cross-lingual comparison: 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*

Cross-lingual comparison: AMR

- 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)

Thank you!

https://ufal.mff.cuni.cz/pdt3.5

https://lindat.mff.cuni.cz/services/PDT-Vallex/

https://lindat.mff.cuni.cz/services/EngVallex/

https://lindat.mff.cuni.cz/services/CzEngVallex/

Supported by LINDAT/CLARIN, LM2015071 by MEYS, and OP VVV LINDAT/CLARIN, ESIF Structural Funds and MEYS.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY