The Prague Dependency Treebank and Valency Annotation (part 2)

Jan Hajič
Institute of Formal and Applied Linguistics
School of Computer Science
Faculty of Mathematics and Physics
Charles University, Prague
Czech Republic

PDT – Syntactic Annotation (tutorial part 2)

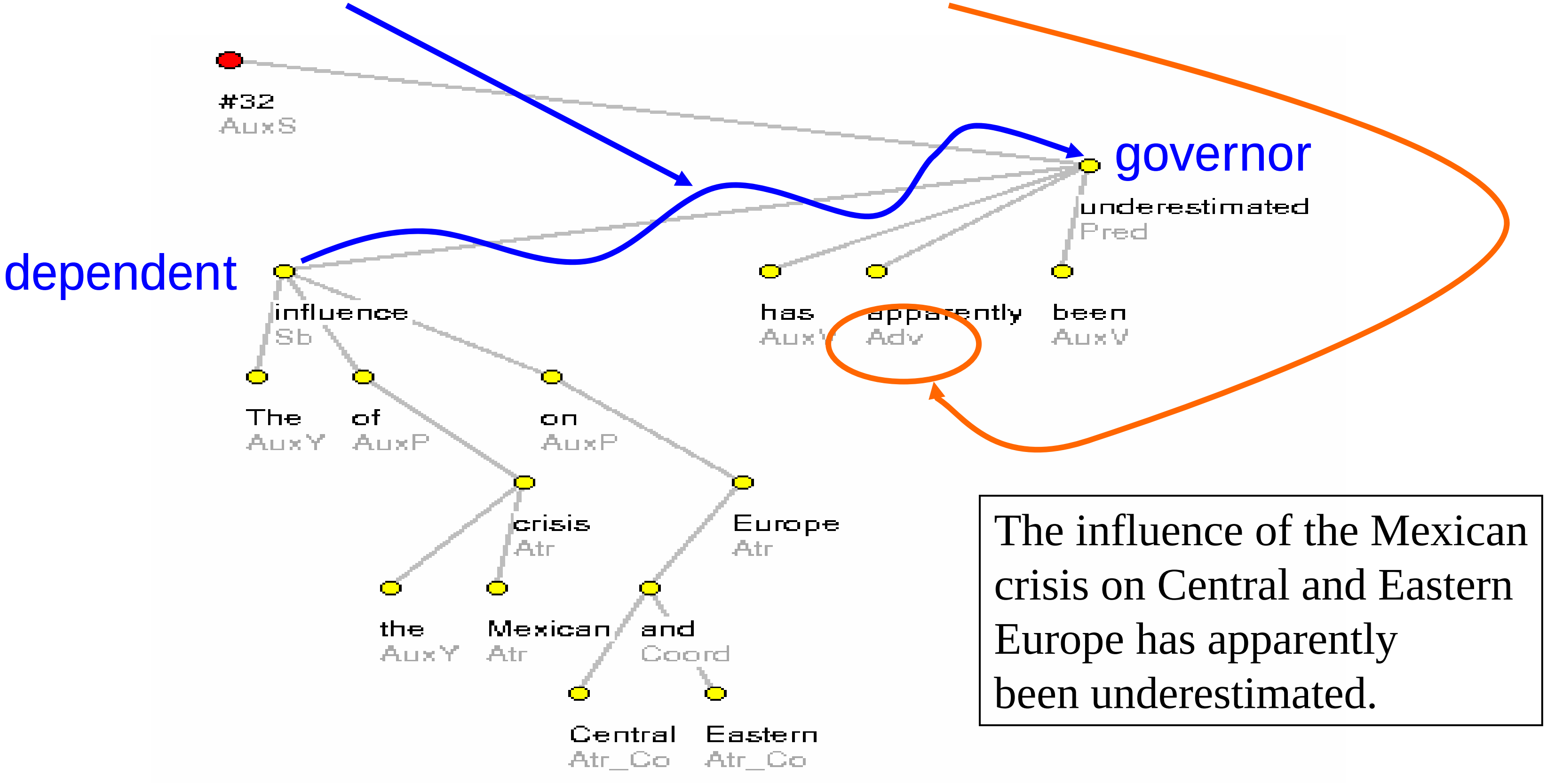
- Surface syntax annotation
  - Dependency surface syntax
  - Comparable to Penn Treebank annotation
    - Convertible: dependency ↔ parse trees
- Deep syntactic/semantic annotation
  - Dependency trees
  - Different topology
  - High level of generalization and formalization
  - Many node attributes

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



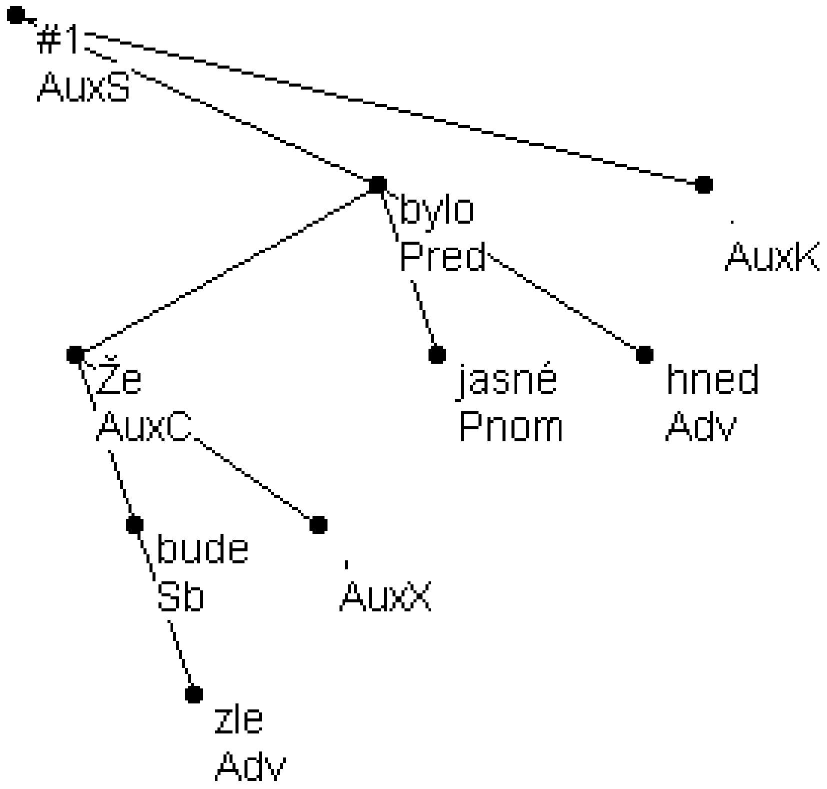
The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated.

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

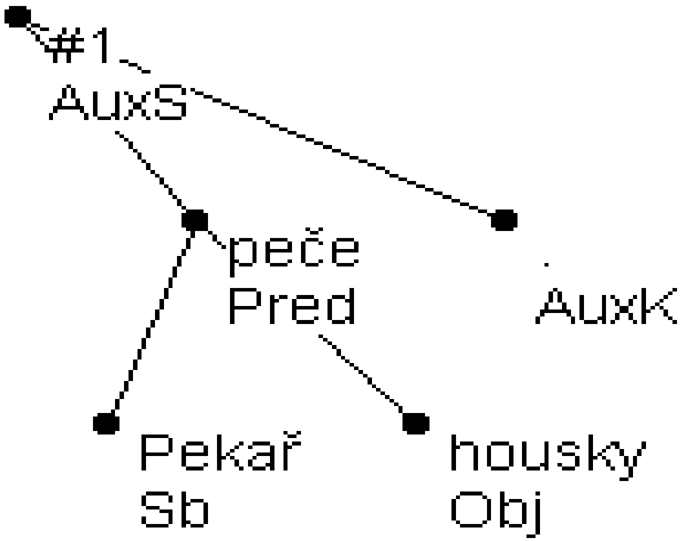
Example

- *lit.* That it will go wrong, (that) was clear immediately.
  - Že bude zle, bylo jasné hned.



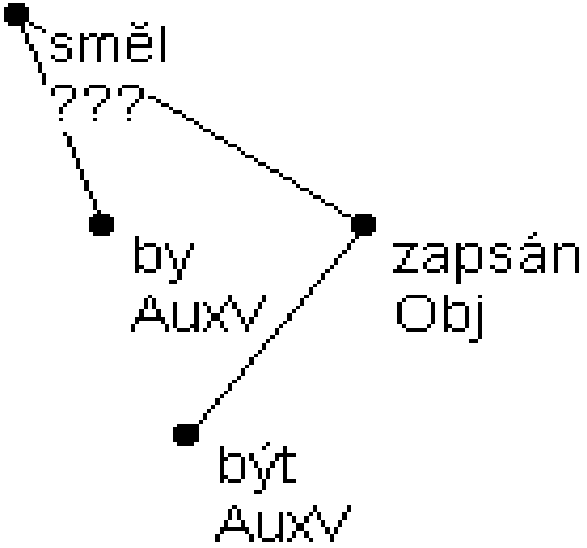
Surface Syntax Example

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



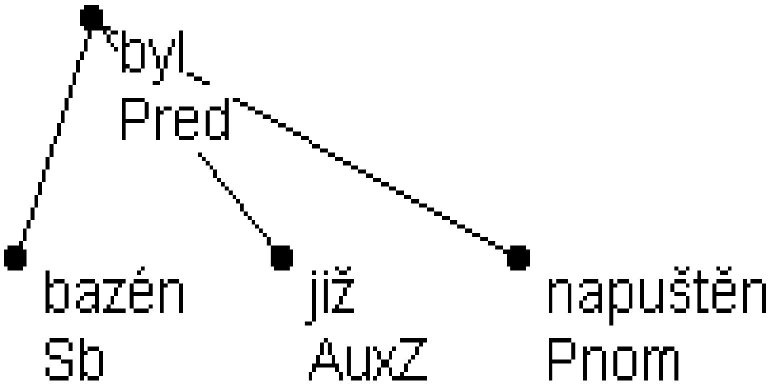
Surface Syntax Example

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



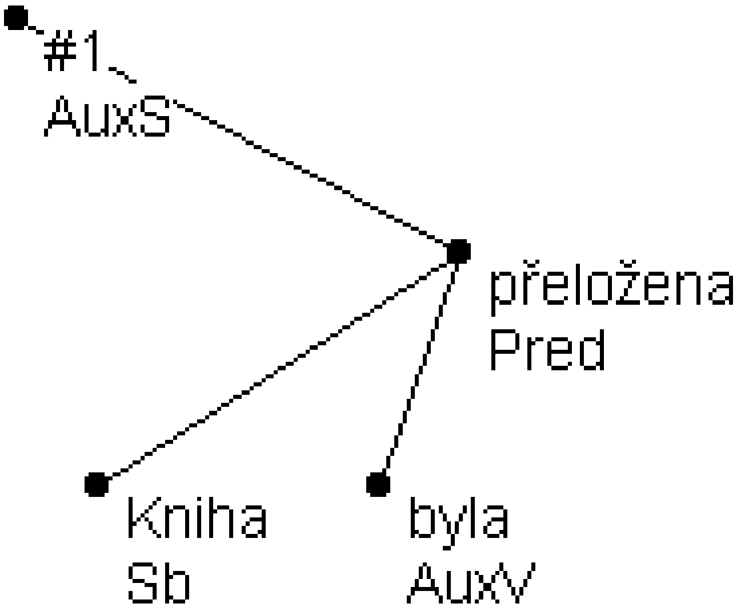
Surface Syntax Example

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



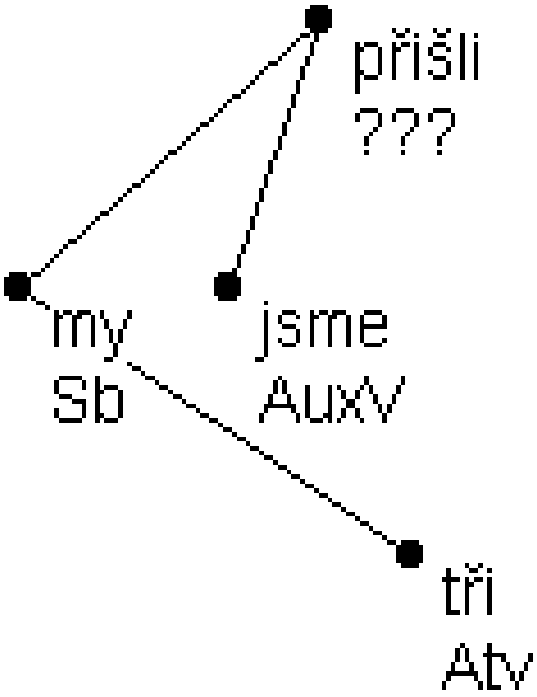
Surface Syntax Example

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



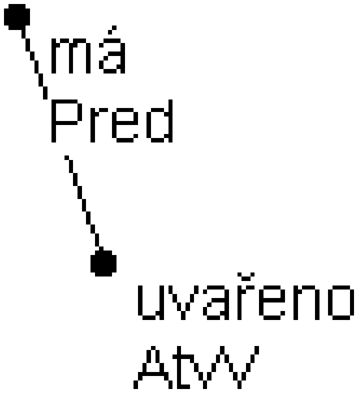
Surface Syntax Example

- Complement
  - we (are) came three
  - my jsme přišli tři



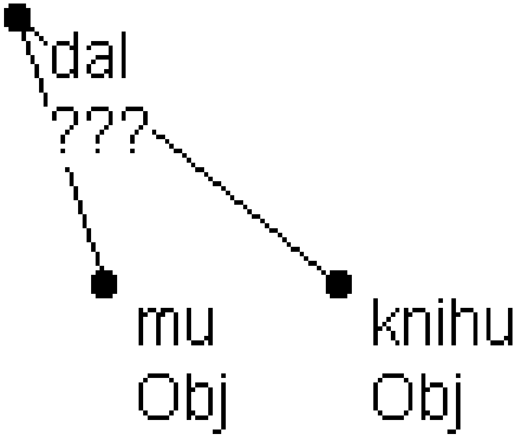
Surface Syntax Example

- Complement when NP is missing
  - (he) has cooked [his meals]
  - má uvařeno



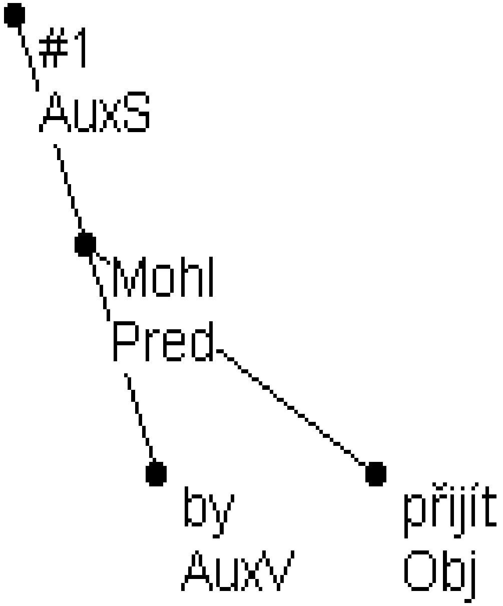
Surface Syntax Example

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



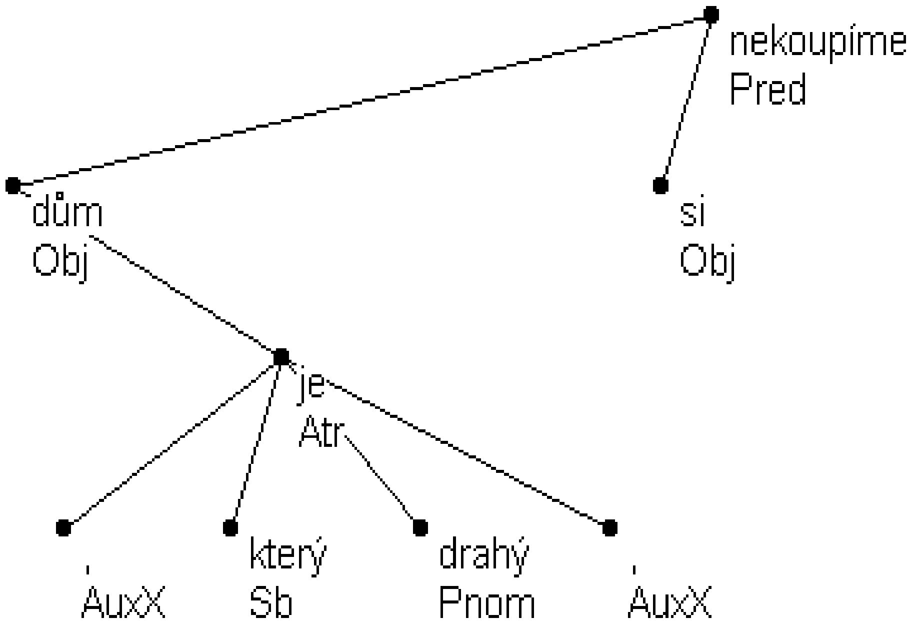
Surface Syntax Example

- Object used for infinitive of analytical verb forms
  - (he) Could come
  - Mohl by přijít



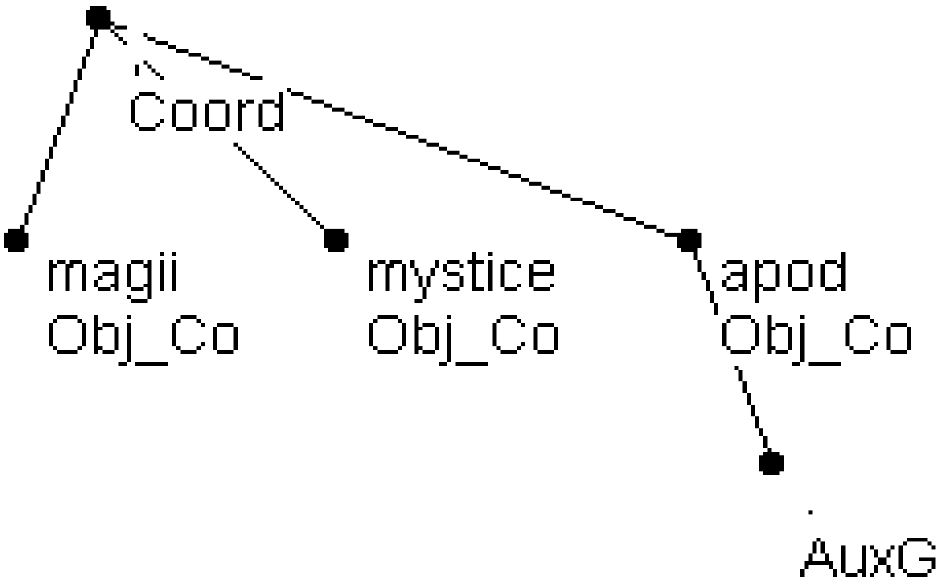
Surface Syntax Example

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



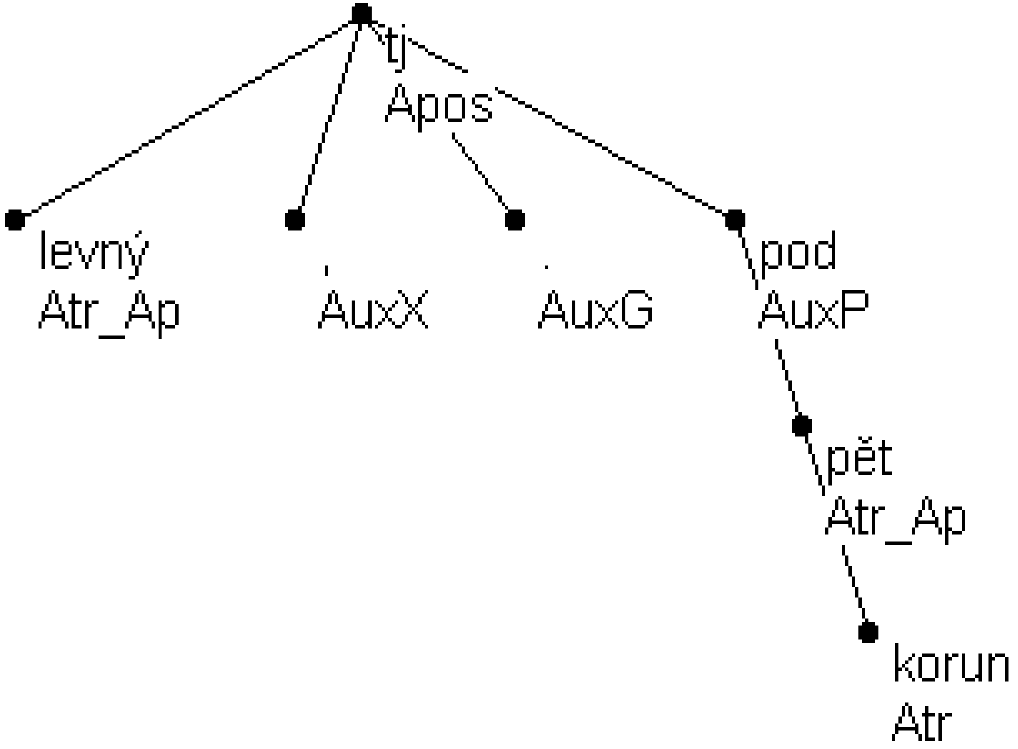
Surface Syntax Example

- Coordination
  - ... (to) magic, mystic(,) etc.
  - ... magii , mystice apod.



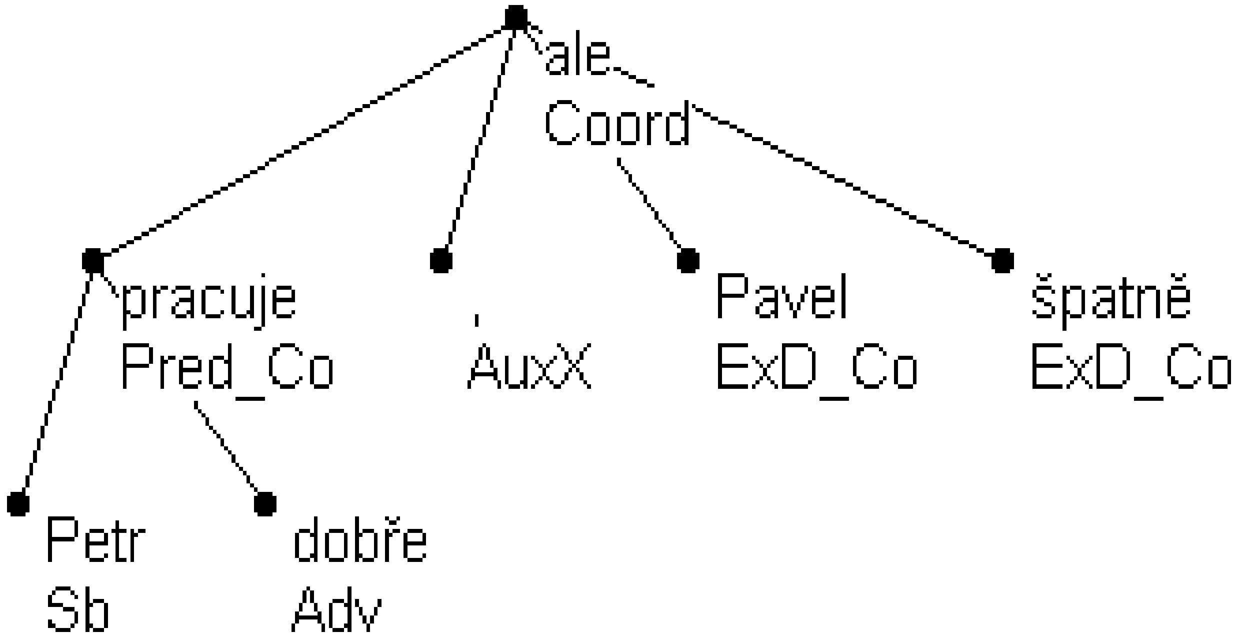
Surface Syntax Example

- Apposition
  - cheap, i.e. under 5 crown
  - levný , tj. pod   5 korun



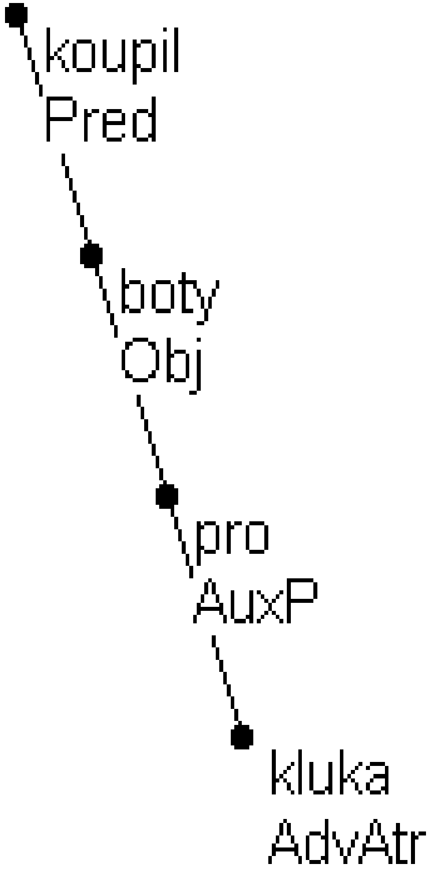
Surface Syntax Example

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



Surface Syntax Example

- Variants (equality)
    - (he) bought shoes for boy
    - koupil boty pro kluka



Using the Results: Parsing

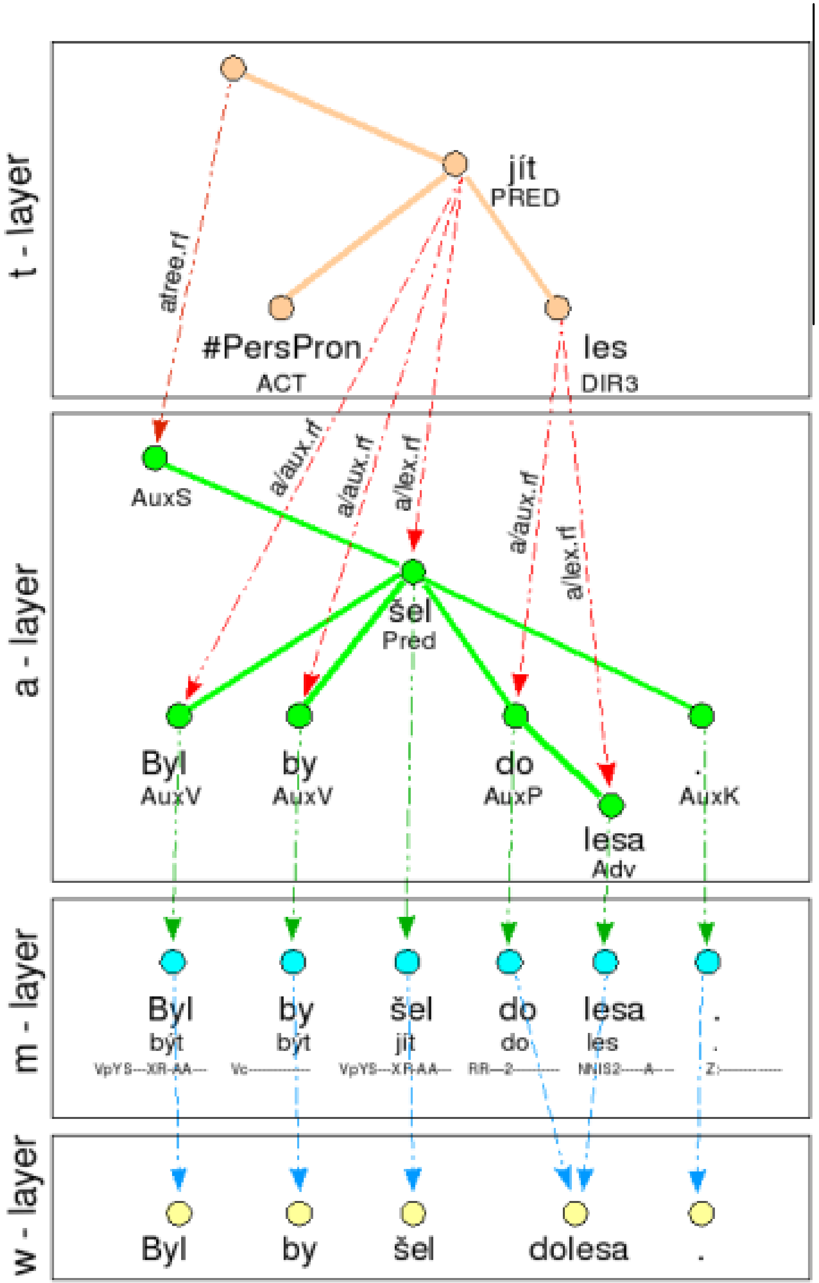
- Several parsers of Czech
  - Analytical layer dependency syntax
  - Trained on PDT 1.0 dat, 1.2 mil. words
- Collins (98), Charniak (00), Žabokrtský (02), Ribarov (04), Nivre (05), Zeman(05), McDonald (05)
- Best results (accuracy: percent of correct dependencies):
  - 84-85% for a single parser, > 86% for a combination

The Prague Markup Language (Intro only – see P. Pajas, p. 6)

- XML-based, UTF-8 coding used
- Stand-off annotation
  - strict hierarchical scheme
  - 4 files for each annotated document ~ 4 layers of annotation
- Can capture intermediate annotation
  - e.g., ambiguous analysis after morphological preprocessing
- Lexical resources linked in
  - valency lexicon referenced from t-layer data

XML Annotation Layers

- Strictly top-down links
- w+m+a can be easily "knitted"
- API for cross-layer access (programming)
- PML Schema / Relax NG
- [With slight modification, can be used for spoken data (audio as layer "-1")]



The Prague Markup Language Example

- m-layer data, linked to w-layer:

```
<m id="m-tr/_12941_01_00013.fs-s1w4">
  <src.rf>manual</src.rf>
  <w>
    <dest.rf>w#w-tr/_12941_01_00013.fs-s1w4</dest.rf>
    <trans>basic</trans>
  </w>
  <form>pocházela</form>
  <lemma>pocházet_:T</lemma>
  <tag>VpQW---XR-AA---</tag>
</m>
<m id="m-tr/_12941_01_00013.fs-s1w5">
  ...
```

Pointer to w-layer

PDT Annotation Layers

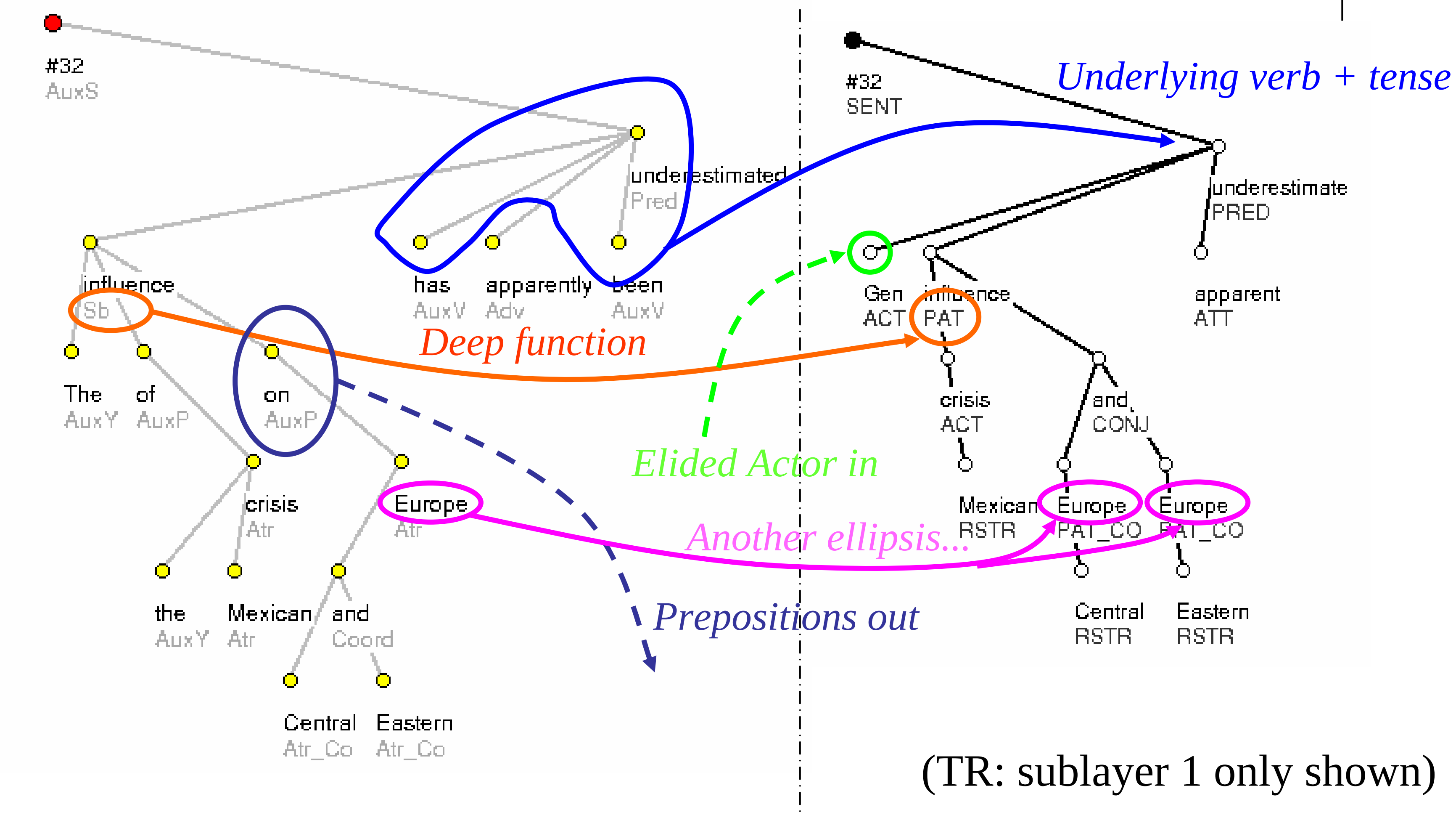
- L0 (w) Words (tokens)
  - automatic segmentation and markup only
- L1 (m) Morphology
  - Tag (full morphology, 13 categories), lemma
- L2 (a) Analytical layer (surface syntax)
  - Dependency, analytical dependency function
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), 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)

Analytical vs. Tectogrammatical

annotation (TR: sublayer 1 only)



Underlying verb + tense

Deep function

Elided Actor in

Another ellipsis...

Prepositions out

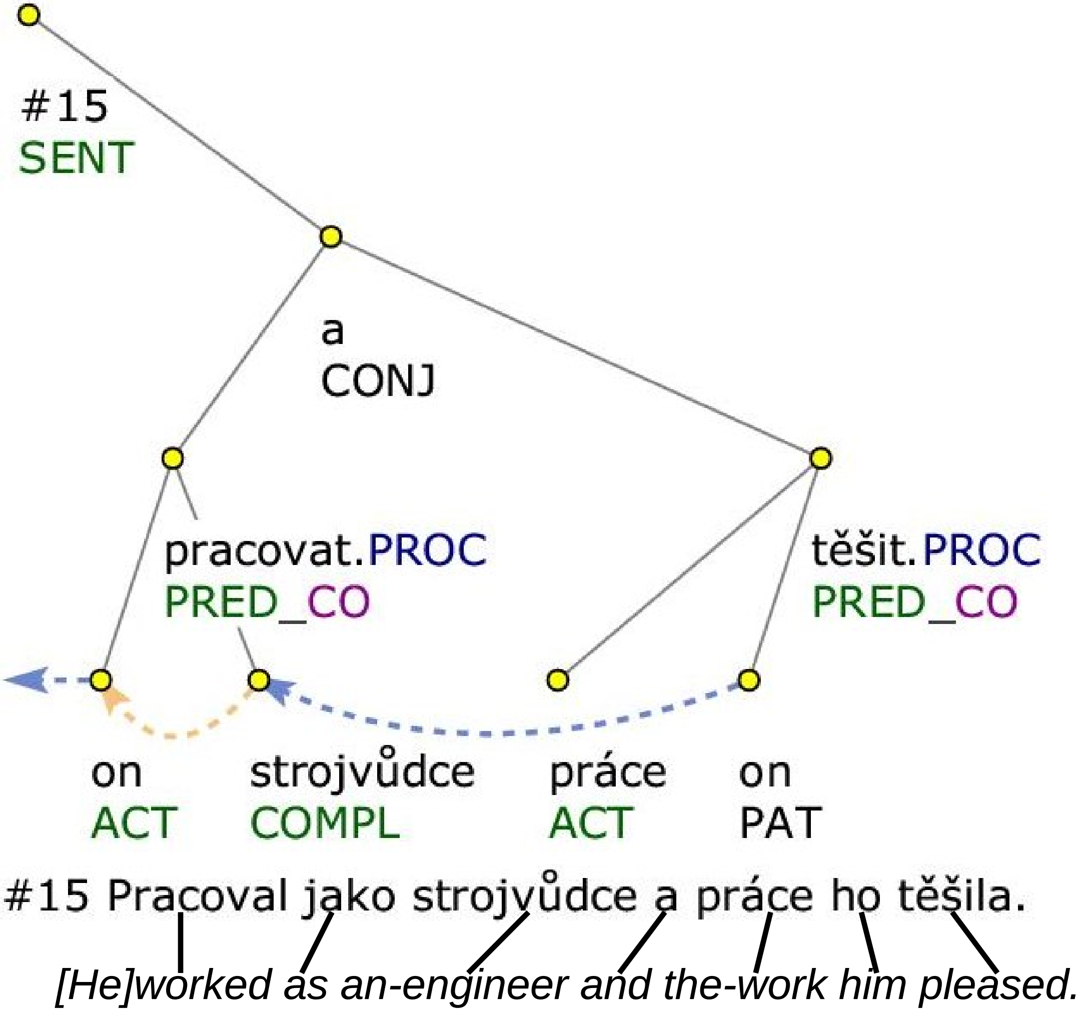
(TR: sublayer 1 only shown)

Layer 3: Tectogrammatical

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

Example - TR

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


#15
SENT
a
CONJ
pracovat.PROC
PRED_CO
těšit.PROC
PRED_CO
on
ACT
strojvůdce
COMPL
práce
ACT
on
PAT
#15 Pracoval jako strojvůdce a práce ho těšila.
[He]worked as an-engineer and the-work him pleased.

Dependency Structure

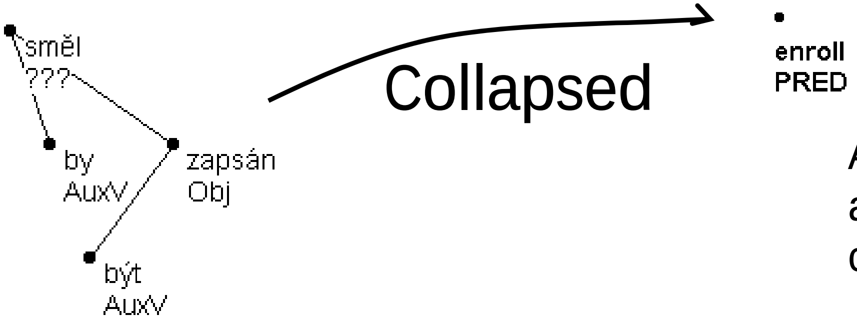
- Similar to the surface (Analytical) layer...
  ...but:
  - certain nodes deleted
    - auxiliaries, non-autosemantic words, punctuation
  - some nodes added
    - based on word (mostly verb, noun) valency
    - some ellipsis resolution
  - detailed dependency relation labels (functors)

Tectogrammatical Functors

- “Actants”: $\overbrace{\text{ACT, PAT,}}^{\text{syntactic}}\ \overbrace{\text{EFF, ADDR, ORIG}}^{\text{semantic}}$
  - 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

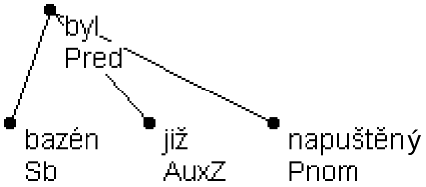


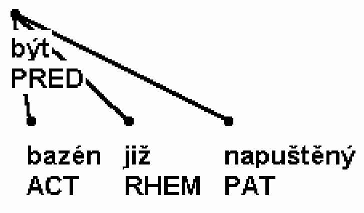
Collapsed

Additional attributes (grammatemes): conditional + “allow”

Tectogrammatical Example

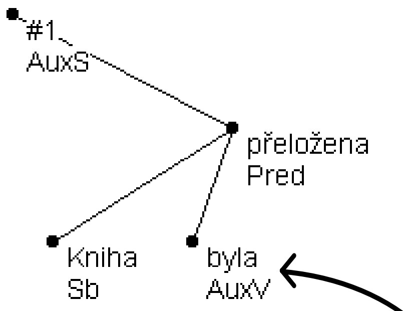
- Predicate with copula (state)
  - (the) pool has-been already filled
  - bazén byl již napuštěný

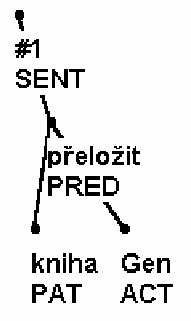




Tectogrammatical Example

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



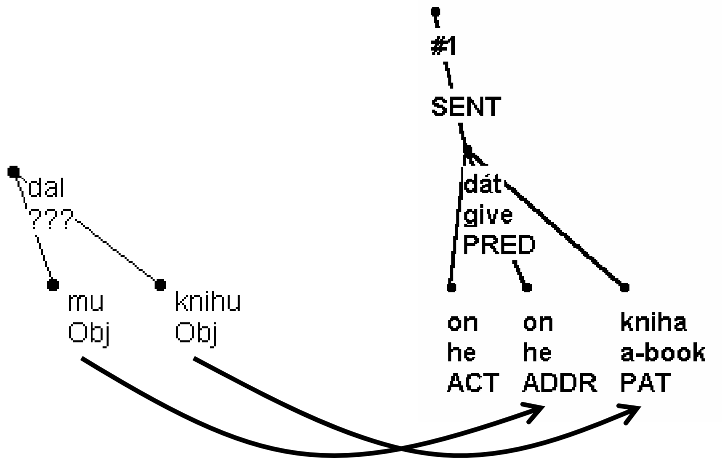


Disappeared

Added

Tectogrammatical Example

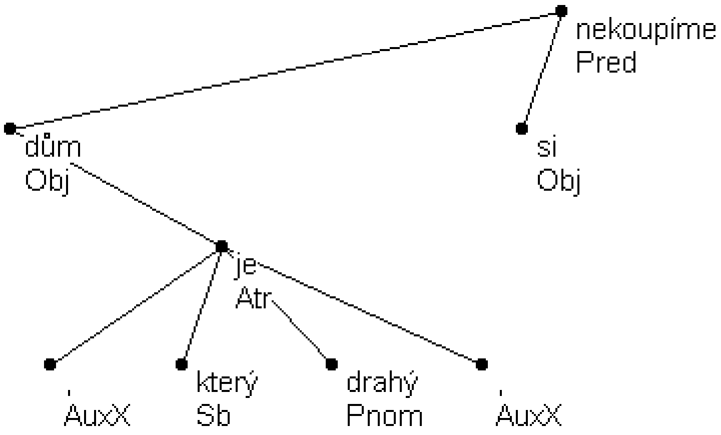
- Object
  - (he) gave him a-book
  - dal mu knihu

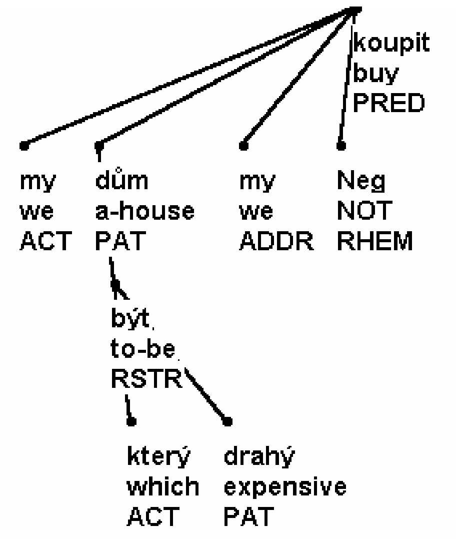


Obj goes into ACT, PAT, ADDR, EFF or ORIG based on governor’s valency frame

Tectogrammatical Example

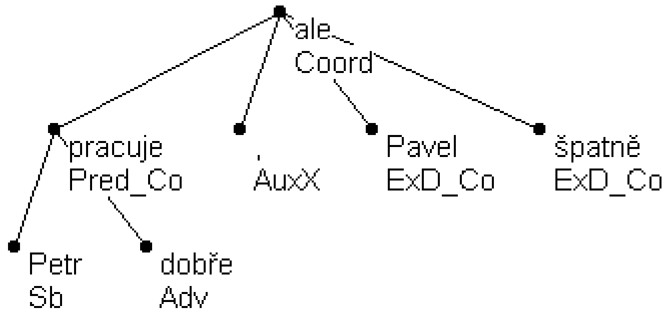
- Relative clause (embedded)
  - (a) house, which is expensive, (we) (to-ourselves) will-not-buy
  - dům , který je drahý , si nekoupíme

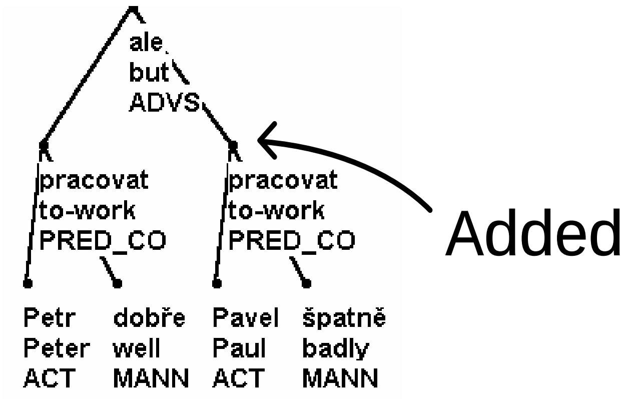




Tectogrammatical Example

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





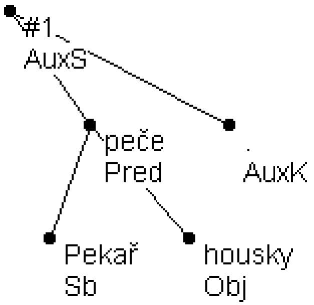
Layer 3: Tectogrammatical

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

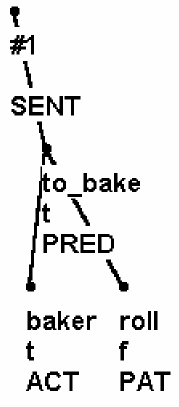
Deep Word Order, Topic/Focus (intro only: see E. Hajičová, p.3)

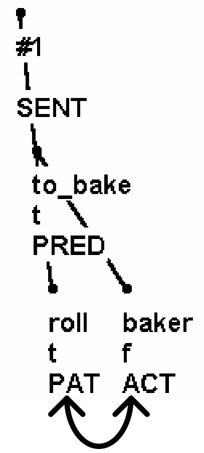
- Example:

Analytical dep. tree:



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





Deep Word Order Topic/Focus

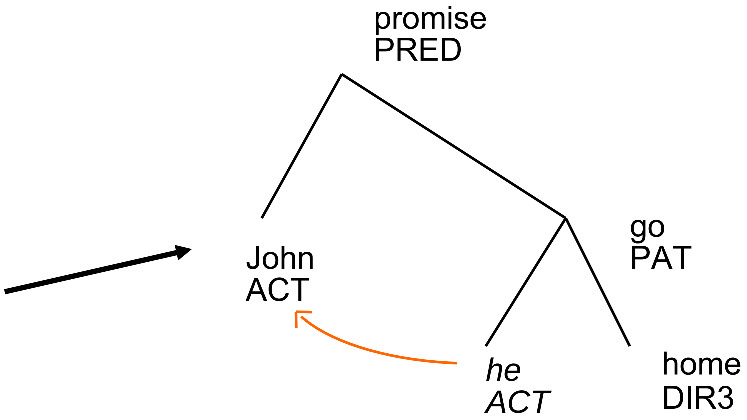
- Deep word order:
  - from “old” information to the “new” one (left-to-right) at every level (head included)
  - projectivity by definition (almost...)
    - i.e., partial level-based order -> total d.w.o.
- Topic/focus/contrastive topic
  - attribute of every node (t, f, c)
  - restricted by d.w.o.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

Layer 3: Tectogrammatical

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

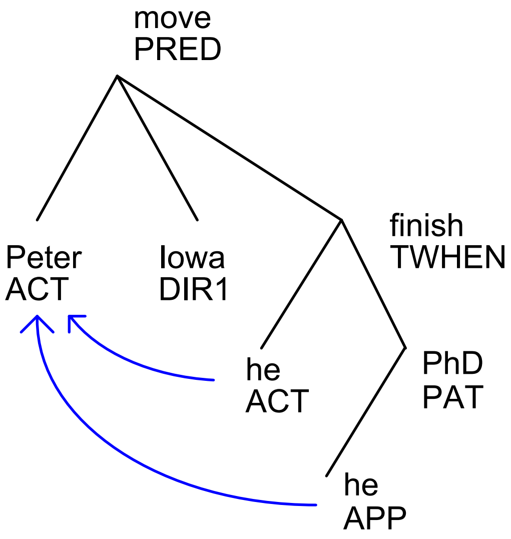
Coreference (intro only: see E. Hajičová p.3)

- 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.



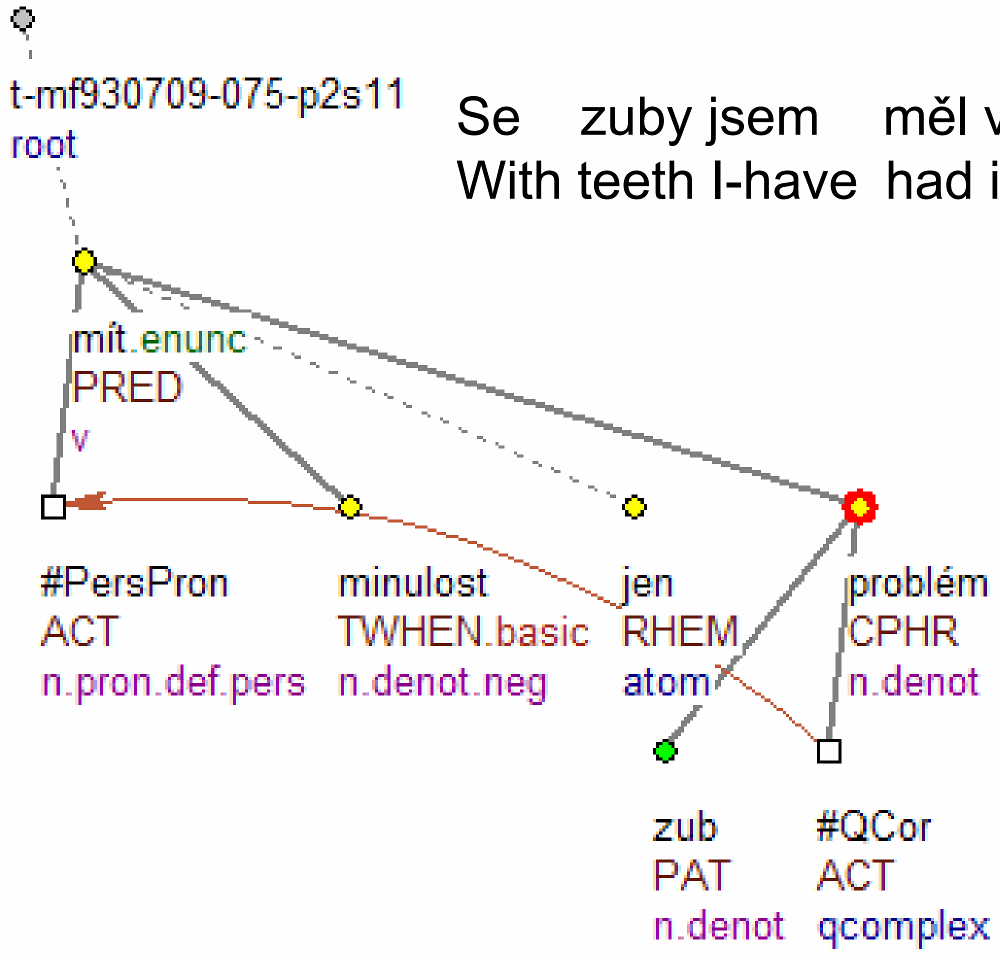
Layer 3: Tectogrammatical

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

Grammatemes (intro only: see Z. Žabokrtský p. 3)

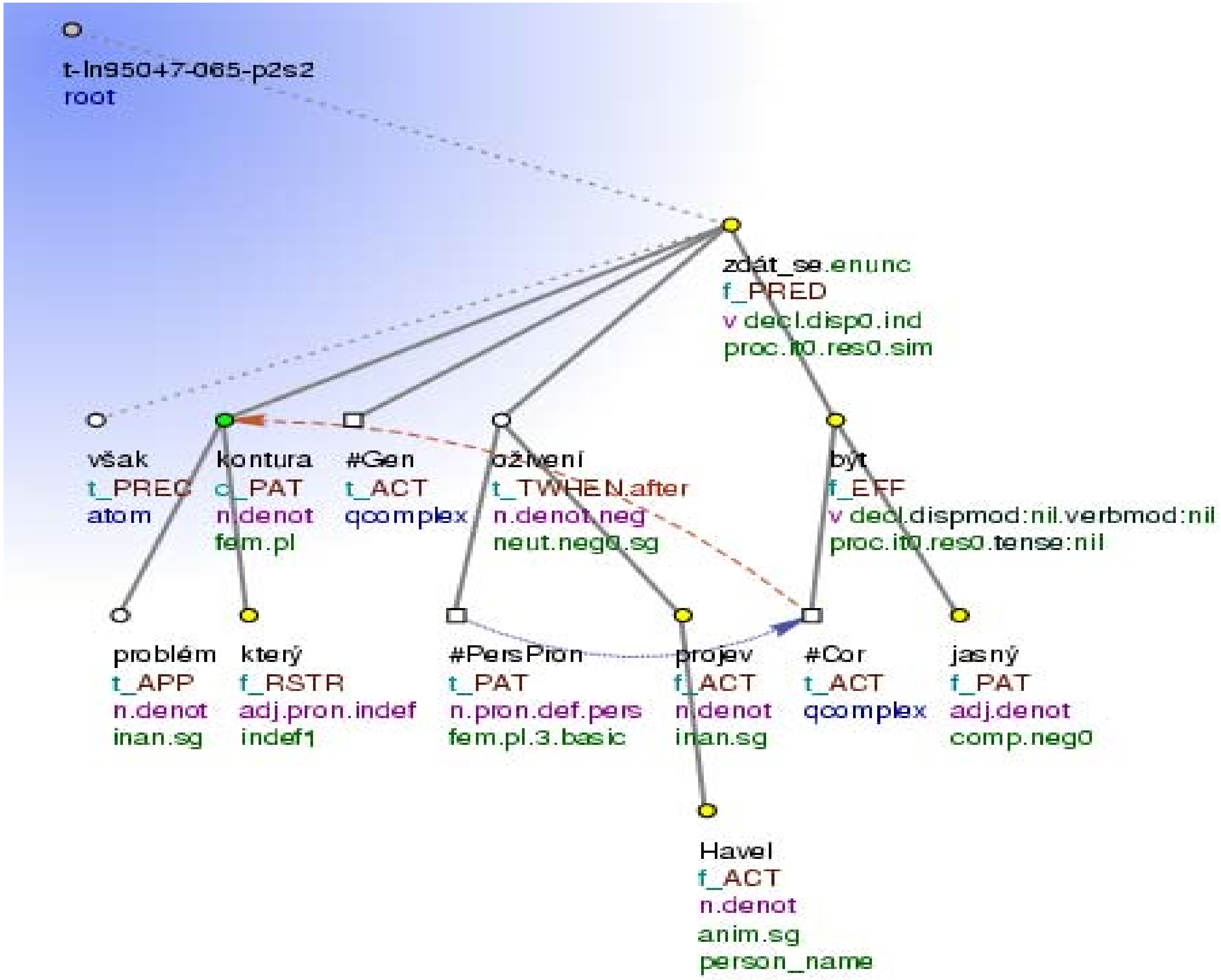
- Detailed functors (subfunctors)
  - only for some functors:
    - TWHEN: before/after
    - LOC: next-to, behind, in-front-of, ...
    - also: ACMP, BEN, CPR, DIR1, DIR2, DIR3, EXT
- Lexical (underlying)
  - number (SG/PL), tense, modality, degree of comparison, ...
  - strictly only where necessary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

Example - simplified view



Se zuby jsem měl v minulosti jen problémy.
With teeth I-have had in the-past only problems.

Fully Annotated Sentence



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