



# Prague Dependency Treebank and Functional Generative Description

Markéta Lopatková

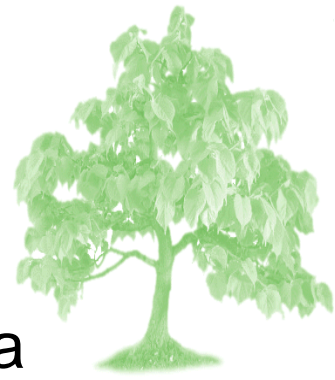
Institute of Formal and Applied Linguistics, MFF UK

[lopatkova@ufal.mff.cuni.cz](mailto:lopatkova@ufal.mff.cuni.cz)

---

---

# Prague Dependency Treebank



~ application of the FGD theory on the large set of data

<http://ufal.mff.cuni.cz/pdt2.0/>

<https://ufal.mff.cuni.cz/pdt3.0/>

1. data

2. tools

3. documentation:

- Guide, <http://ufal.mff.cuni.cz/pdt2.0/>
- manuals for individual layers  
<https://ufal.mff.cuni.cz/pdt3.0/documentation>
- survey of data formats and tools
- releases: 2.0 (2006), 2.5 (2011), 3.0 (2013)

---

# Prague Dependency Treebank (cont.)



4 layers:

- word layer (w-layer)
- morphological layer (m-layer)
- analytical layer (a-layer)
- tectogrammatical layer (t-layer)

***layers of annotation***

layers of description	t,a,m-layer				a,m-layer
	train	dtest	etest	total	total
# documents	2 536	316	316	3 168	2 170
# sentences	38 737	5 228	5 477	<b>49 442</b>	<b>38 538</b>
# tokens	652 700	87 988	92 669	833 357	671 490

---

# Prague Dependency Treebank (cont.)



- stand-off annotation
- manual annotation  
with a massive post-annotation consistency checking
- formats and tools:
  - TrEd ... tree editor and viewer (Pajas, xxxx)  
<http://ufal.mff.cuni.cz/~pajas/tred/index.html>
  - PML data format (XML-based format )  
<http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pml/index.html>
  - PML-TQ ... search tool  
<http://ufal.mff.cuni.cz/~pajas/pmltq/>
- more during the practical sessions

---

# PDT: w-layer



- layer of source texts (1991-1995)
  - Lidové noviny (daily newspapers)
  - Mladá fronta Dnes (daily newspapers)
  - Českomoravský Profit (business weekly)
  - Vesmír (scientific journal)
- part of the Czech National Corpus
- a sequence of **tokens** (word forms and punctuation marks)
- including errors, typing errors, bad segmentation, ...

---

# PDT: m-layer



- the sequence of tokens divided into sentences
- errors are corrected
- annotation:
  - ***morphological lemma***
  - ***morphological tag***
  - id
  - reference to w-layer
  - form (corrections: spelling errors, incorrectly split or joined words, ...)
- manually annotated (parallel annotation)

# PDT: m-layer



Některé kontury problému se však po oživením Havlovým projevem zdají být jasnější .  
[Some contours of the problem seem to be clearer after the resurgence by Havel's speech.]

Form	Lemma	Morphological tag
Některé	některý	PZFP1-----
kontury	kontura	NNFP1-----A----
problému	problém	NNIS2-----A----
se	se_^(zvr._zájmeno/částice)	P7-X4-----
však	však	J^-----
po	po-1	RR--6-----
<b>oživení</b>	oživení_^(*3it)	NNNS6-----A----
Havlovým	Havlův_;S_^(*3el)	AUIS7M-----
projevem	projev	NNIS7-----A----
zdají	zdat	VB-P---3P-AA---
být	být	Vf-----A----
jasnější	jasný	AAFP1-----2A----
.	.	Z:-----



---

# PDT: a-layer

- dependency tree
- one token from m-layer ~ one node incl. prepositions, punctuation ... plus technical root
- relations ~ edges  
dependency, coordination, punctuation, ...
- linear ordering ~ surface word order
- annotation:
  - ***analytical function*** (afun)
  - ***linear order***
  - is\_member
  - is\_parenthesis\_root
  - id
  - reference to m-layer

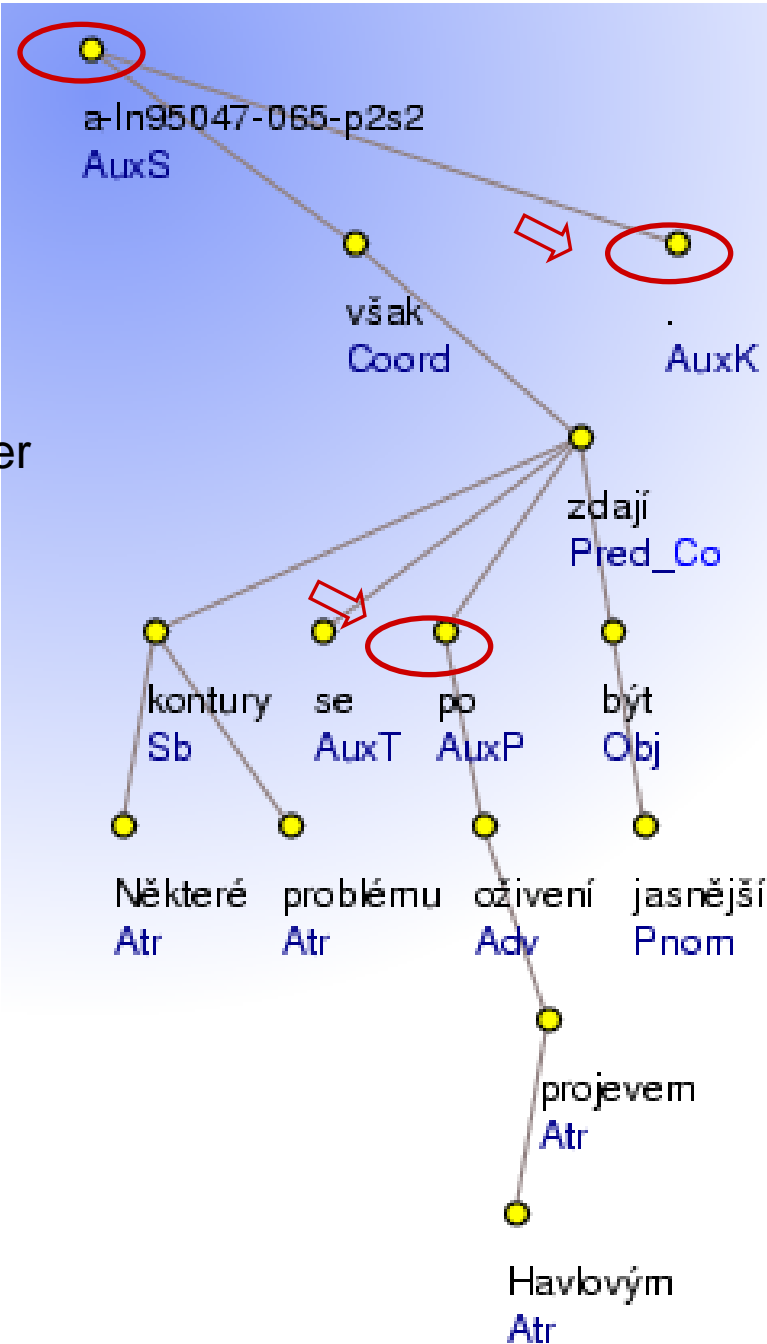
} coordination, apposition, parenthesis



# PDT: a-layer

*Některé kontury problému se však po oživením Havlovým projevem zdají být jasnější .*

[Some contours of the problem seem to be clearer after the resurgence by Havel's speech.]



---

# PDT: t-layer



- tectogrammatical tree structure ~ dependency tree
  - nodes for auto-semantic/lexical words only  
syn-semantic/functional words as attributes of lexical words  
(plus technical root)
  - ellipses as nodes
  - edges ~ relations (dependency, coordination, others)
  - link to a valency lexicon for verbs and (certain types of) nouns
- topic-focus articulation (TFA)
  - linear ordering ~ deep word order
  - contextually bounded and unbounded nodes
- coreference

---

# PDT: t-layer (basic attributes)

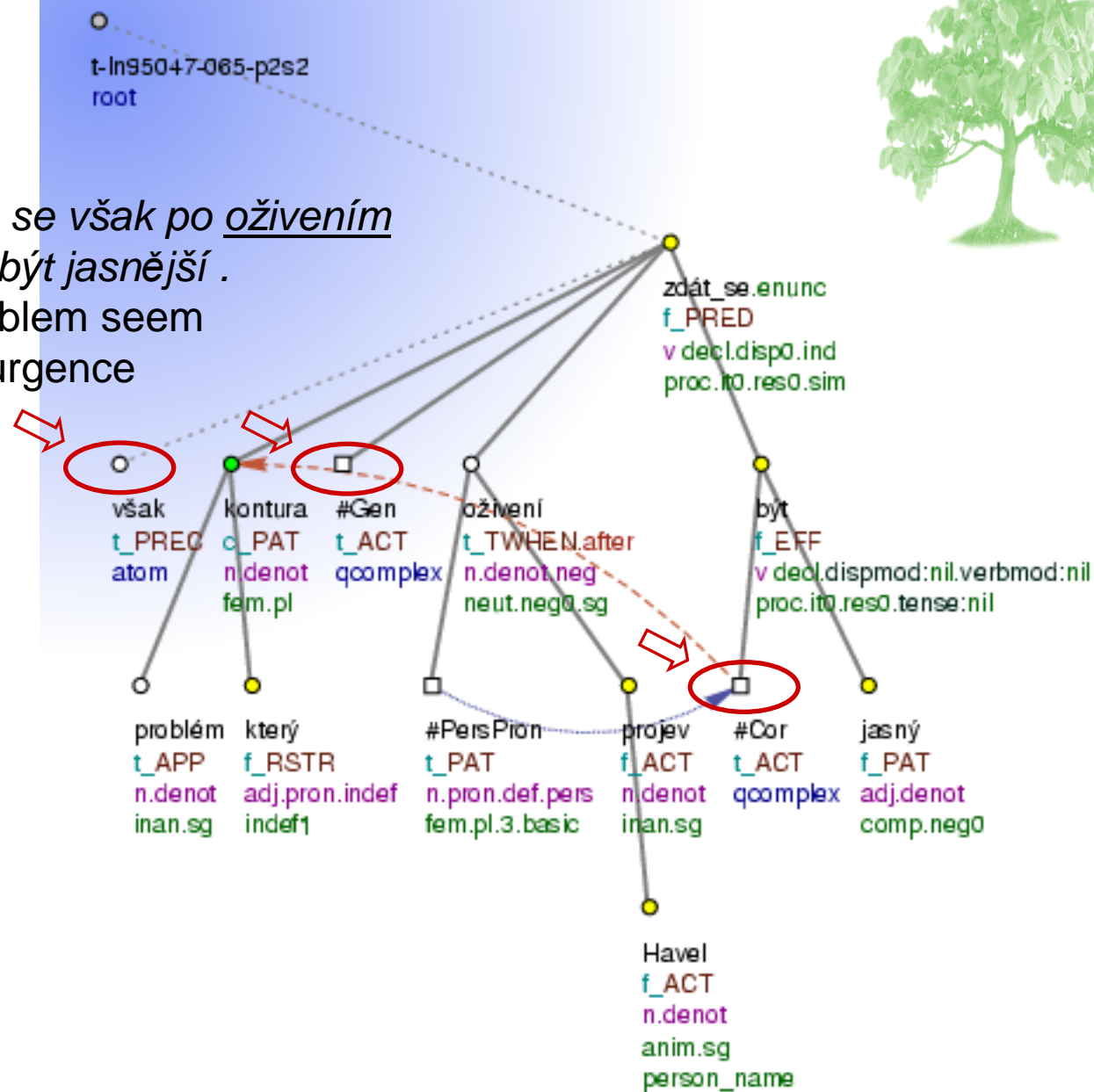


- tectogrammatical tree structure
  - **t-lemma**
  - **functor**
  - **grammatemes** (16 attributes starting with the prefix gram )
  - is\_member
  - is\_parenthesis\_root
  - id
  - reference to a-layer
  - ...
- topic-focus articulation (TFA)
  - deepord
  - tfa
- coreference
  - coref\_text.rf
  - coref\_gram.rf
  - ...

# PDT: t-layer

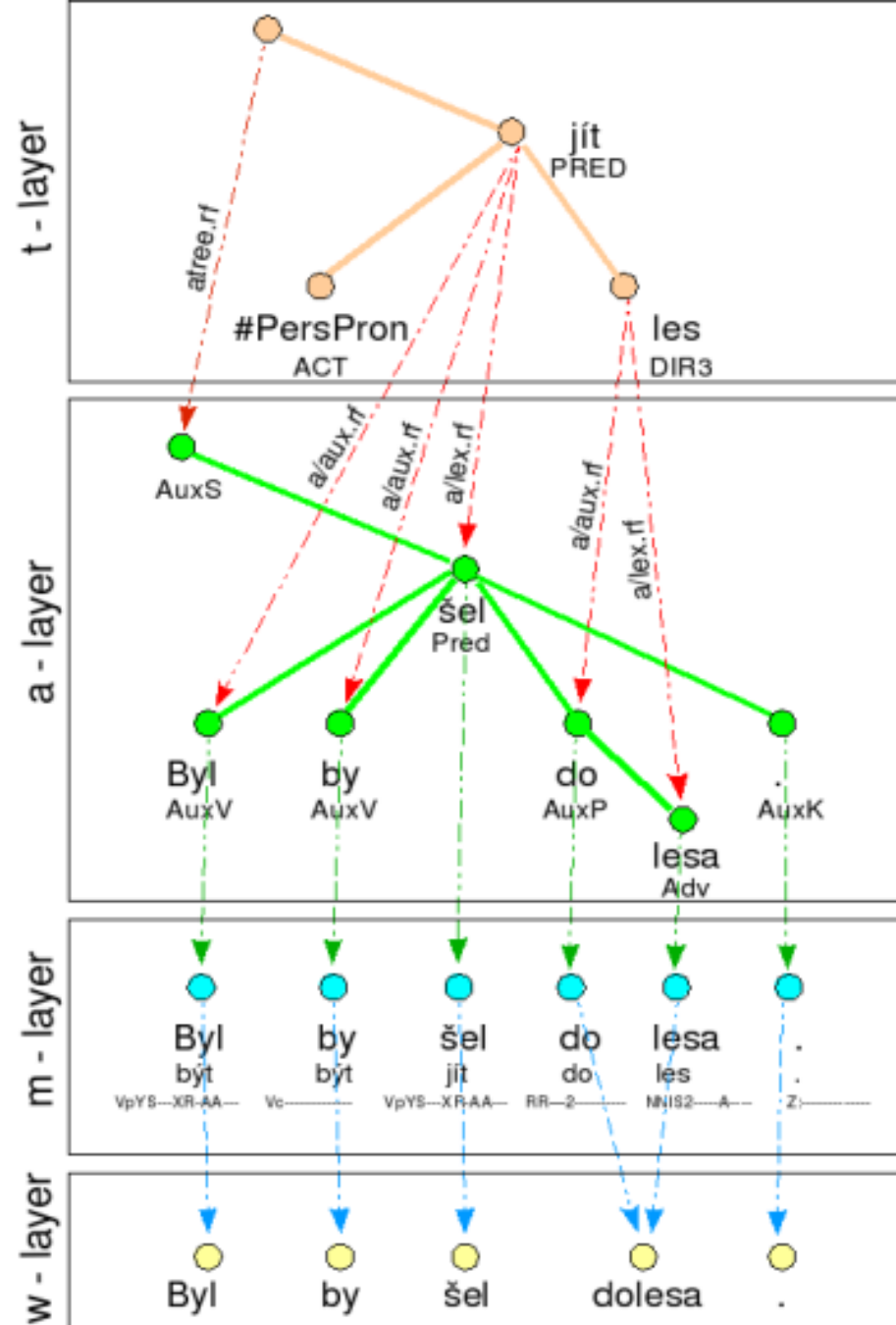
*Některé kontury problému se však po oživením Havlovým projevem zdají být jasnější .*

[Some contours of the problem seem to be clearer after the resurgence by Havel's speech.]

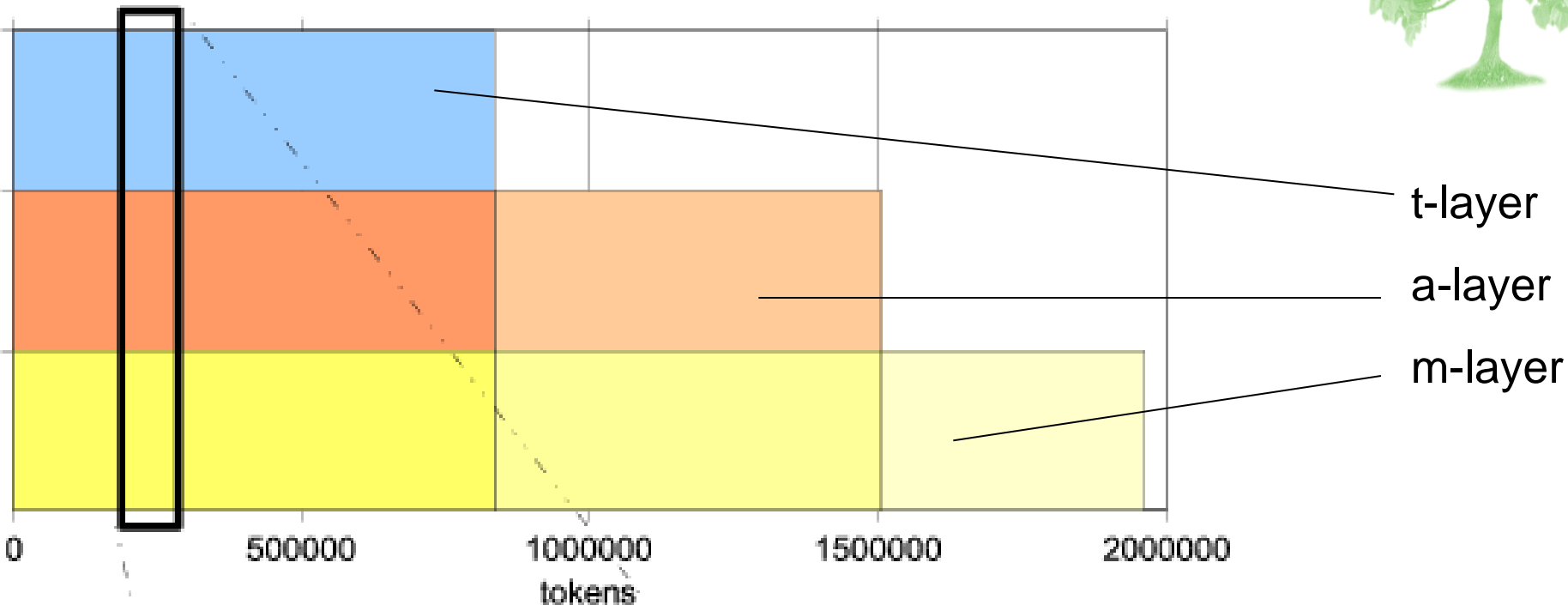


# Linking the layers

- references *from a higher layer to a lower layer*:
  - t-layer → a-layer
  - a-layer → m-layer
  - m-layer → w-layer
- 1:1** correspondence between nodes of the *m-* and *a-layers*



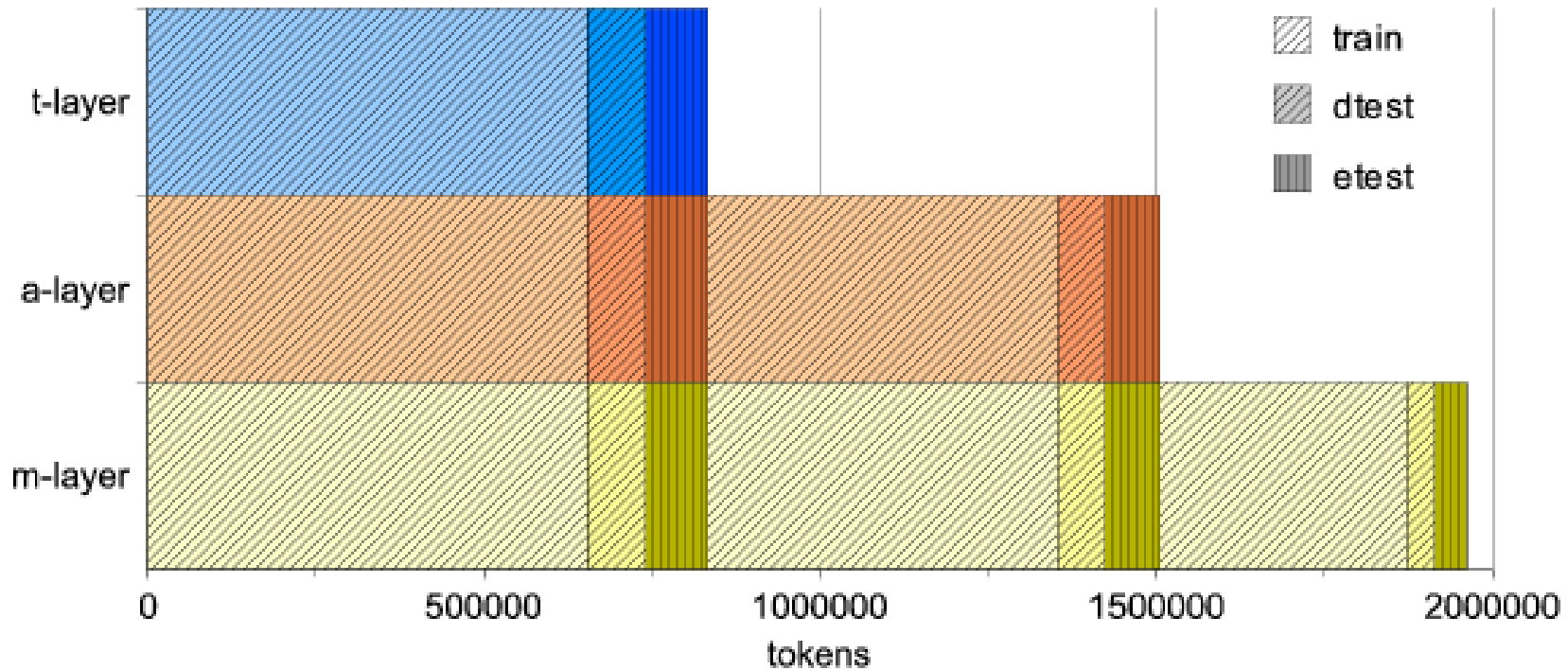
# Division of the data to layers



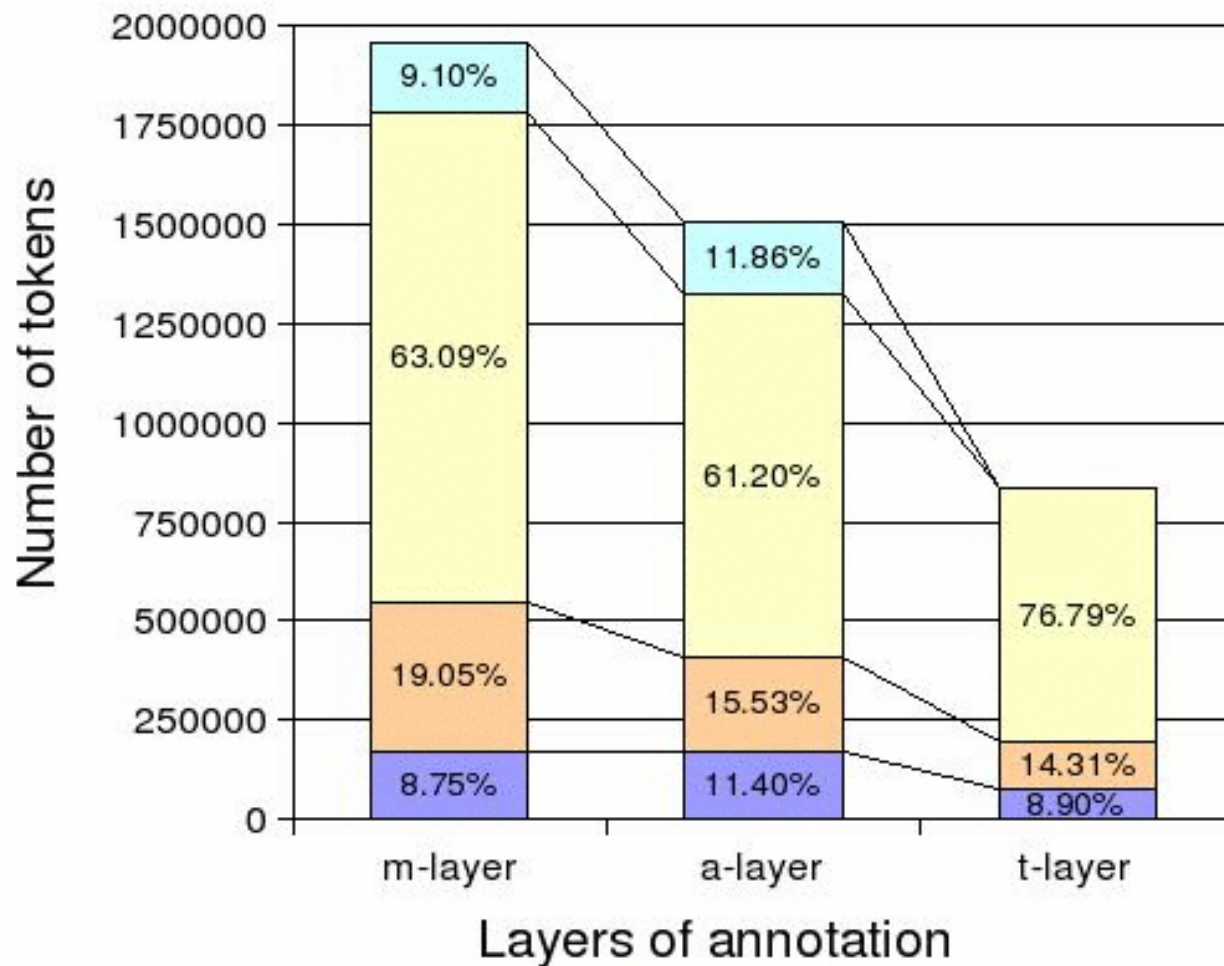
In94206_1.m.gz + In94206_1.w.gz	In94206_1.a.gz	In94206_1.t.gz
In94206_2.m.gz + In94206_2.w.gz	In94206_2.a.gz	In94206_2.t.gz
In94206_3.m.gz + In94206_3.w.gz	In94206_3.a.gz	In94206_3.t.gz

---

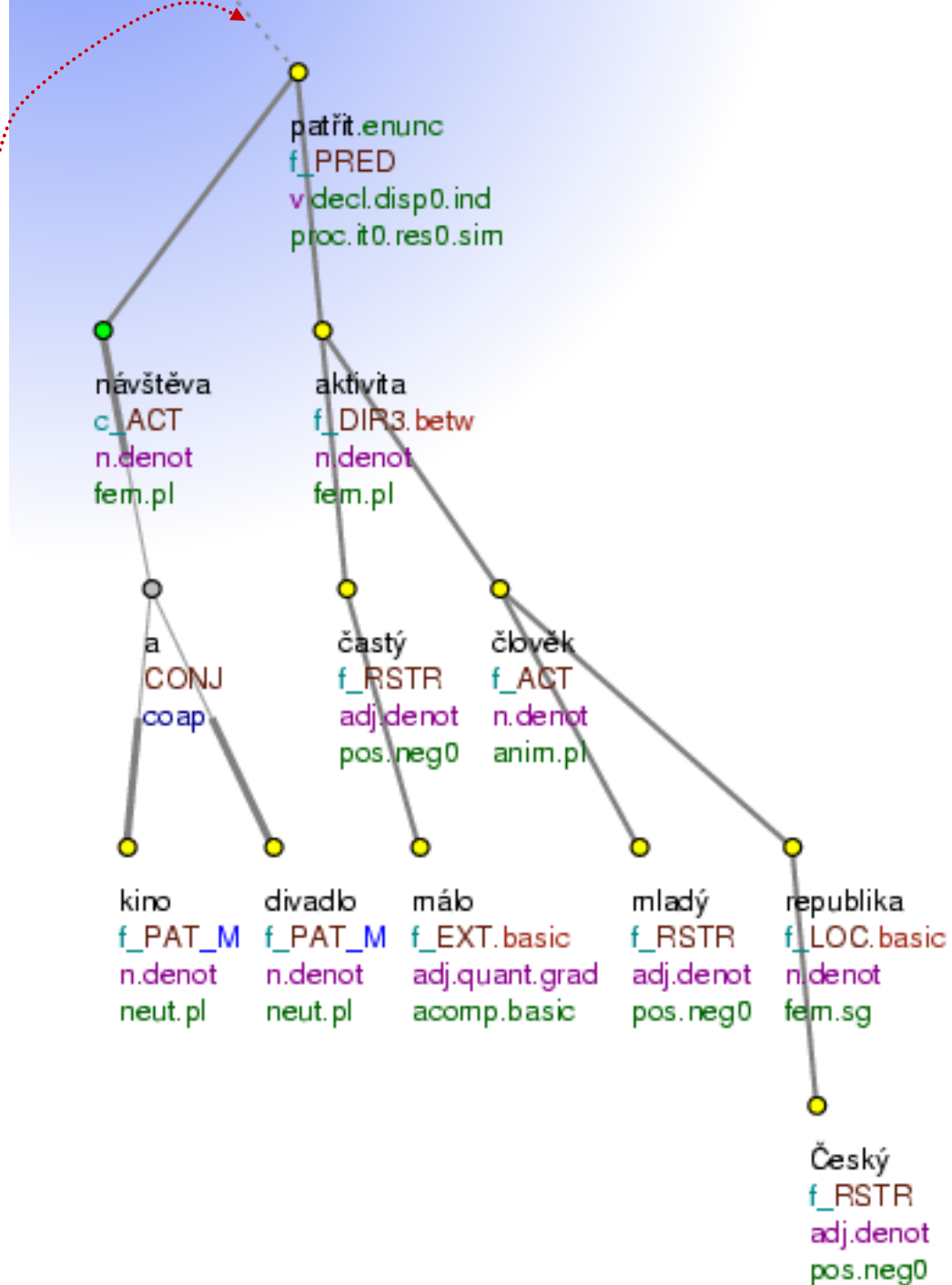
# Division of the data into training and test sets

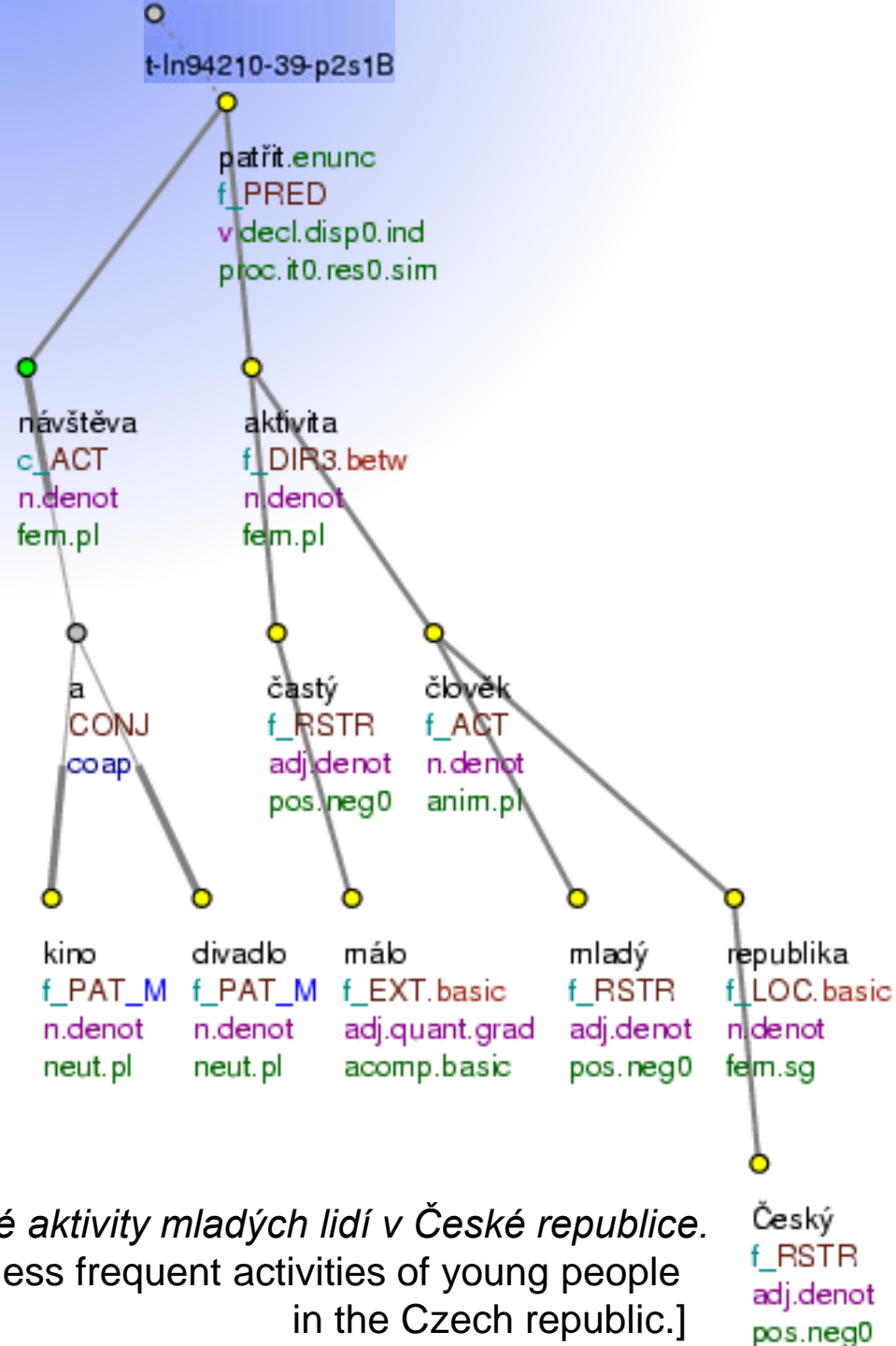
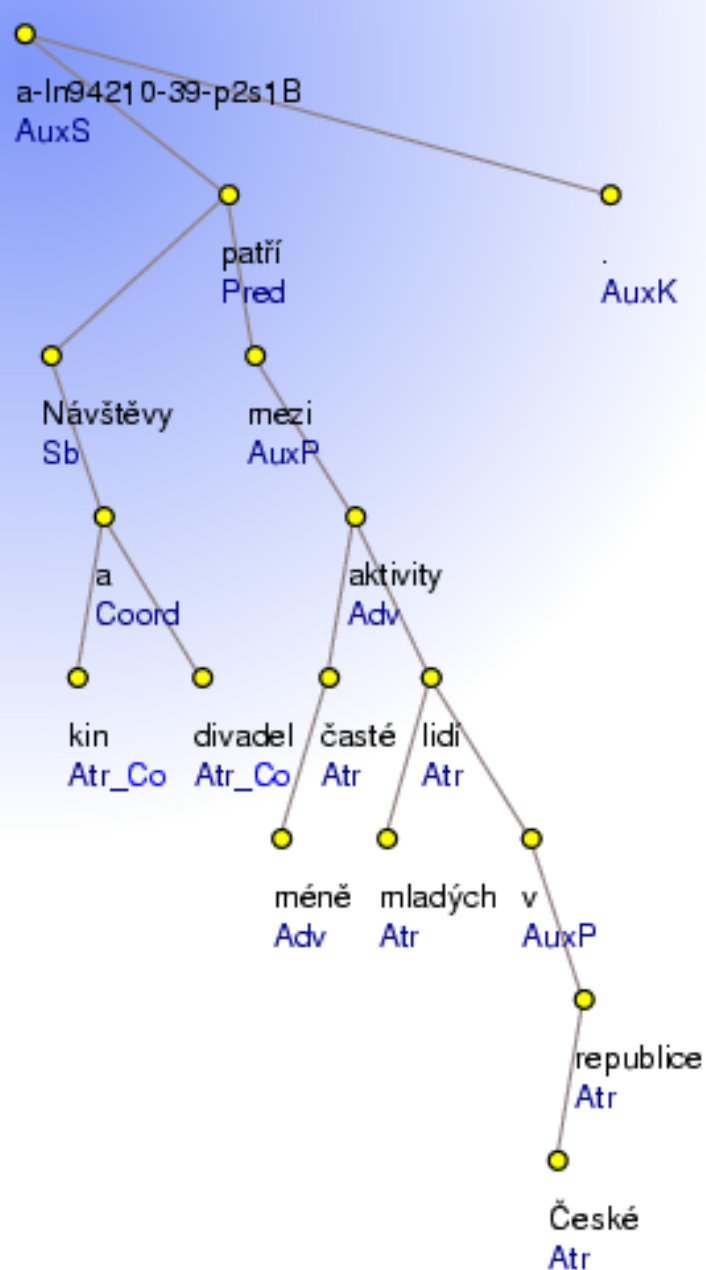


# Number of tokens from the particular sources

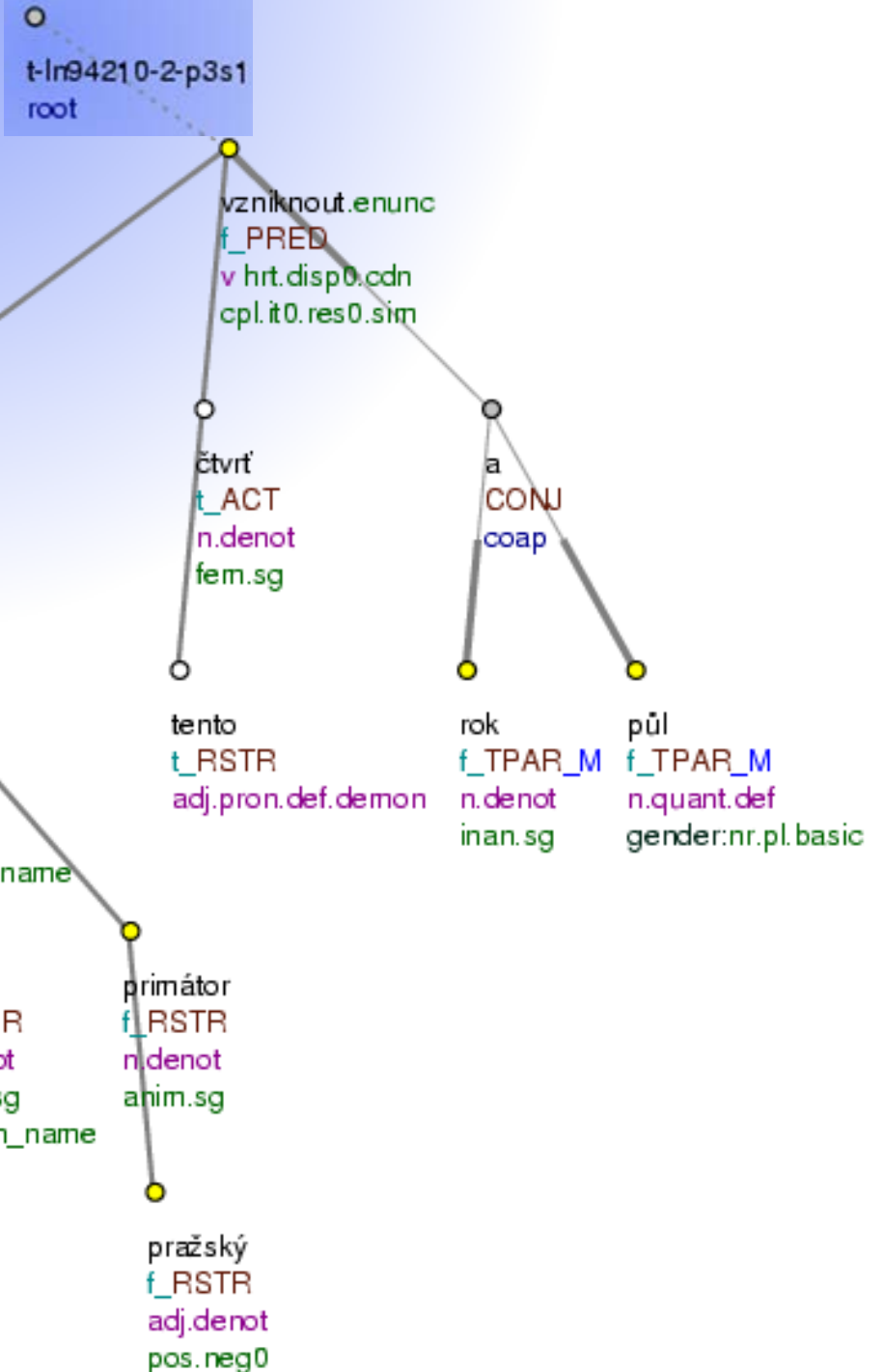
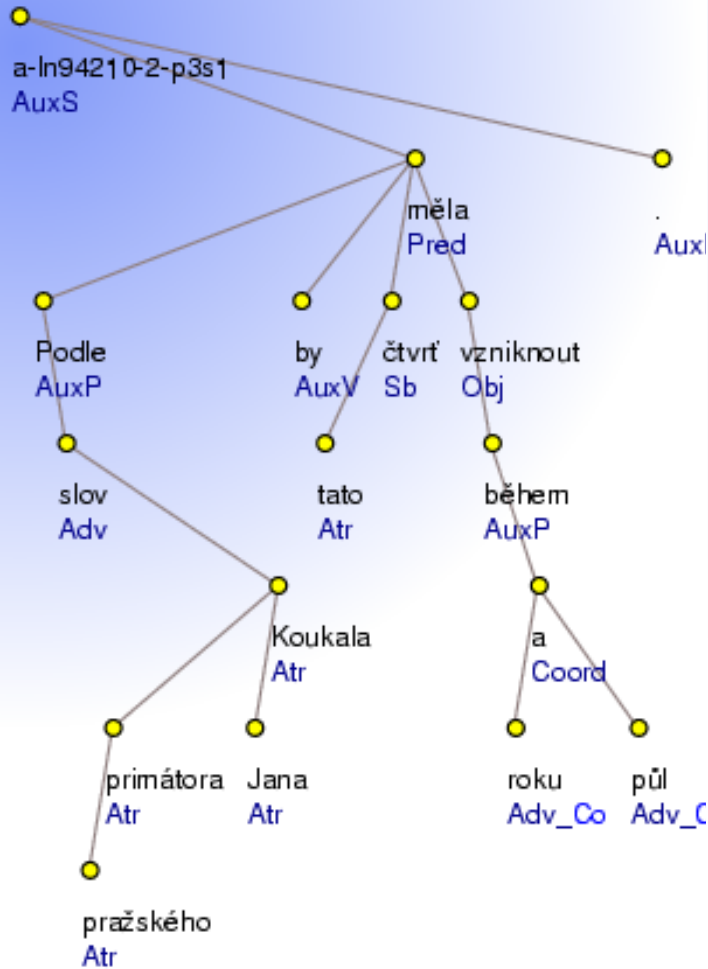








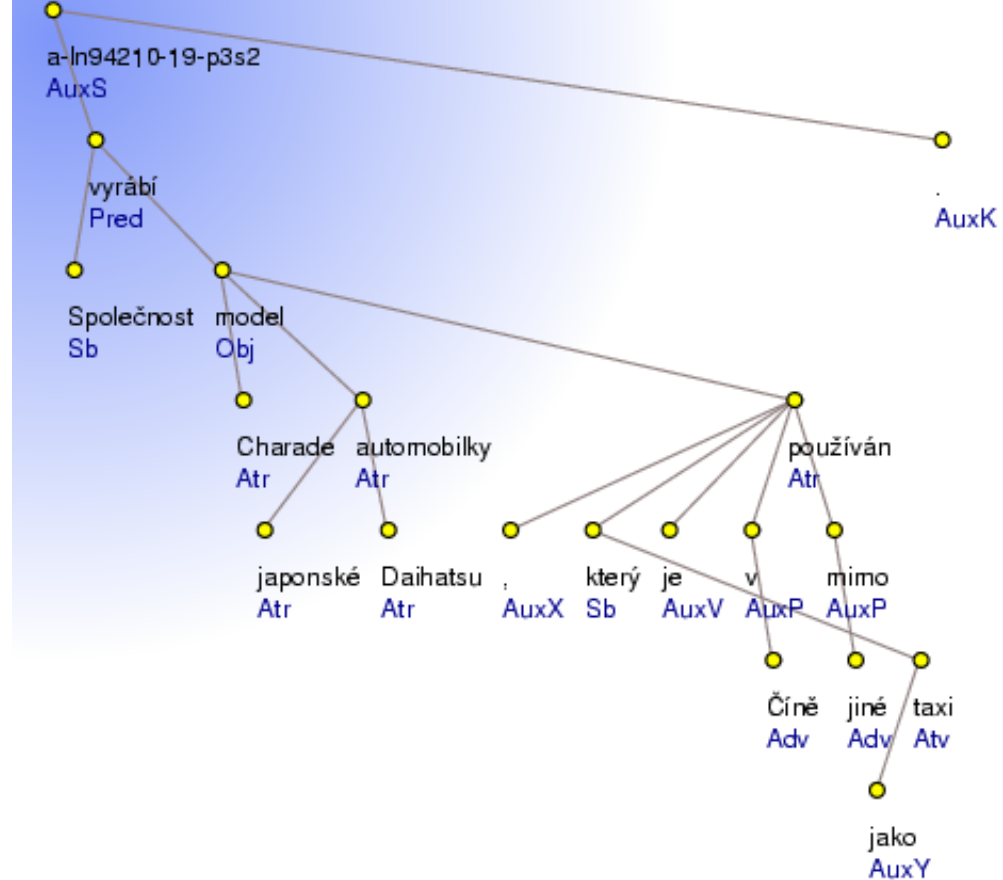
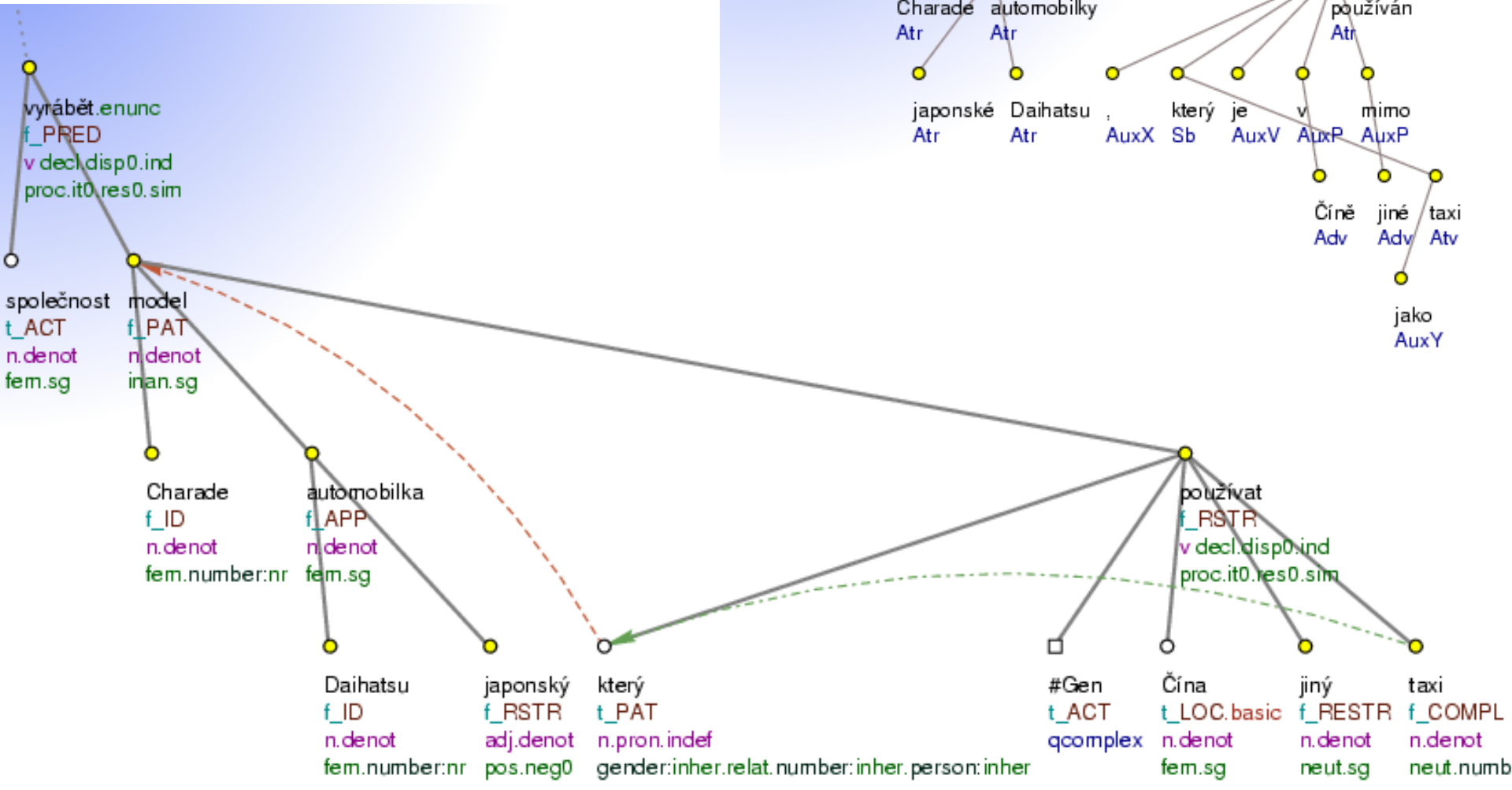
*Návštěvy kin a divadel patří mezi méně časté aktivity mladých lidí v České republice.*  
 [Attending cinemas and theaters belongs to less frequent activities of young people  
 in the Czech republic.]



*Podle slov pražského primátora Jana Koukala  
by tato čtvrť měla vzniknout během roku a půl.*  
[In the words of the city's mayor Jan Koukal,  
this quarter should arise in a year and a half.]

*Společnost vyrábí model Charade japonské automobilky Daihatsu, který je v Číně používán mimo jiné jako taxi.*

[The company produces the Charade model of the Japanese car factory Daihatsu, which is used in China also as a taxi.]



---

# Differences between FGD and PDT





# Differences between FGD and PDT

## FGD

- tectogrammar/deep syntax
- surface syntax
- morphematics
- morphonology
- phonology

## PDT

- t-layer (tectogrammatical l.)
- a-layer (analytical l.)
- m-layer (morphological l.)
- w-layer (word layer)

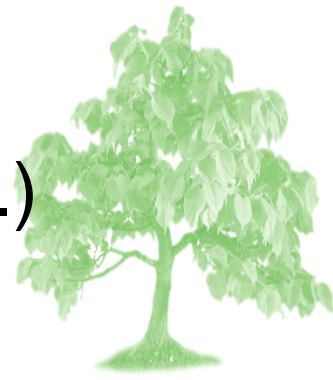
***structural layers***

## reasons

- analysis vs. synthesis/generation ➡ richer information
- technical reasons (financial, temporal restrictions, implementation)

---

# Differences between FGD and PDT (cont.)

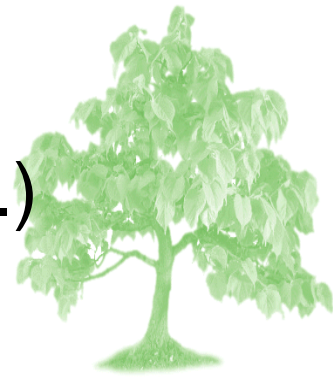


*morphematics* (FGD) vs. *m-layer* (PDT)

- morphemes for individual words are grouped
- grammatical categories ~ morphological tags
- annotated text is divided into sentences

---

# Differences between FGD and PDT (cont.)



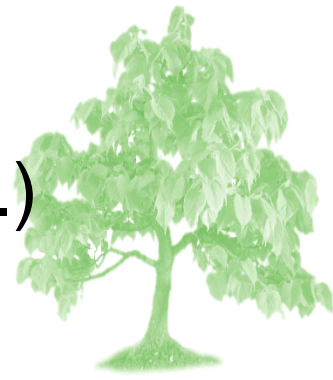
## *structural layers*

- technical root
- connecting constructions for coordination and apposition in PDT



---

# Differences between FGD and PDT (cont.)



## 1. *surface syntax* (FGD) vs. *a-layer* (PDT)

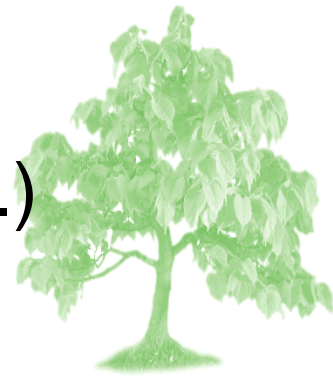
- each token of m-layer is represented by a node (incl. prepositions, auxiliary verbs, punctuation, ...)

(vs. units corresponding to formemes)



edges for non-dependency relations  
(other than coordination/apposition)

- function words (e.g., auxiliary verbs) usually below respective lexical words
- exception: prepositions, subordinating conjunctions as parents of lexical words



---

# Differences between FGD and PDT (cont.)

## 1. *surface syntax* (FGD) vs. *a-layer* (PDT)

- each token of m-layer is represented by a node (incl. prepositions, auxiliary verbs, punctuation, ...)

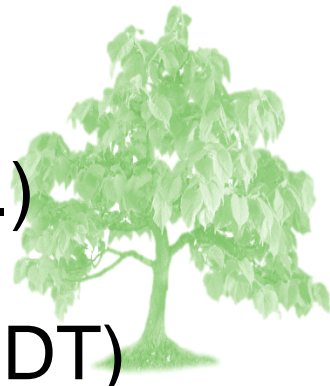
(vs. units corresponding to formemes)

⇒ edges for non-dependency relations  
(other than coordination/apposition)

- function words (e.g., auxiliary verbs) usually below respective lexical words
  - exception: prepositions, subordinating conjunctions as parents of lexical words
  - ellipses: elided words are not restored at a-layer
- ⇒ a word modifying an elided word as a child of the 'lowest' ancestor

---

# Differences between FGD and PDT (cont.)

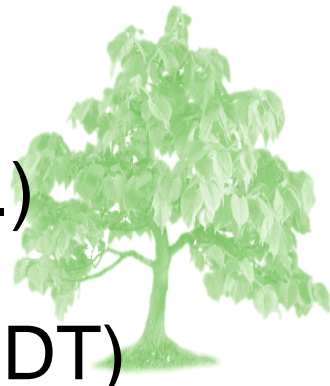


## 2. *deep/tectogram*. syntax (FGD) vs. *t-layer* (PDT)

- core vs. periphery
  - specific constructions (direct speech, comparison)
- edges for non-dependency relations
  - syntactically unclear expressions
  - list structures
  - phrasemes
- info on the (non)realization in the surface sentence (is\_generated)

---

# Differences between FGD and PDT (cont.)



## 2. *deep/tectogram*. syntax (FGD) vs. *t-layer* (PDT)

- core vs. periphery
  - specific constructions (direct speech, comparison)
- edges for non-dependency relations
  - syntactically unclear expressions
  - list structures
  - phrasemes
- info on the (non)realization in the surface sentence (is\_generated)
- topic-focus articulation
- coreference
  - relative/ interrogative pronouns, personal pronouns (3<sup>rd</sup> person)
  - grammatical control, complement

---

# Other treebanks: Prague dependency family



Prague Dependency Treebank 1.0 (2001), 2.0 (2006)

---

# Other treebanks: Prague dependency family



## Prague Dependency Treebank

1.0 (2001); 2.0 (2006); 2.5 (2011); 3.1 (2013)

<http://ufal.mff.cuni.cz/pdt2.5/>

## Czech Academic Corpus 1.0 (2006), 2.0 (2008)

[http://ufal.mff.cuni.cz/rest/CAC/cac\\_20.html](http://ufal.mff.cuni.cz/rest/CAC/cac_20.html)

- morphological annotation (652 000 tokens, 32 000 sentences)
- analytical annotation (493 000 tokens, 25 000 sentences)
- both written and spoken language
- manually annotated

## Prague Dependency Treebank of Spoken Czech

<http://ufal.mff.cuni.cz/pdtsl/>

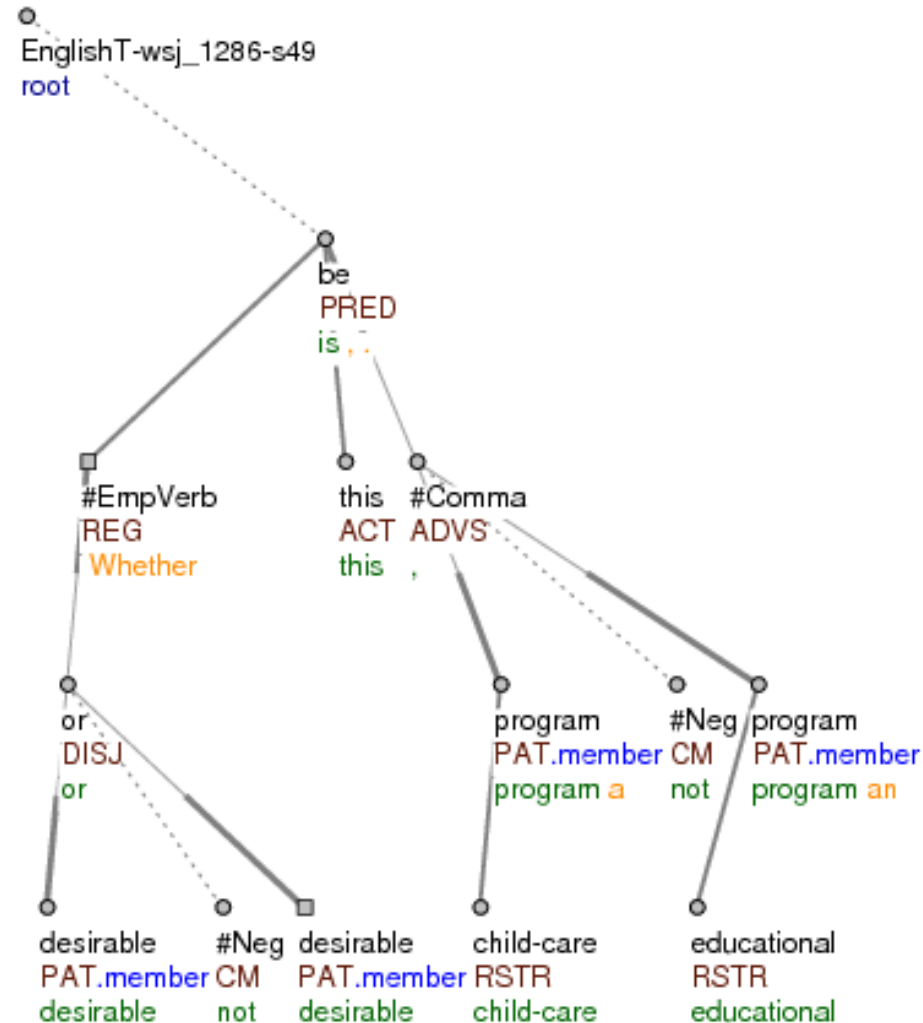
# Other treebanks: Prague dependency family



## Prague English Dependency Treebank 1.0 (2009)

- texts from the Wall Street Journal (Penn Treebank III)
- adaptation of the PDT-like annotation scheme to English
- tectogrammatical annotation
- 12 440 annotated and checked trees

*Whether desirable or not,  
this is a child-care program,  
not an educational program.*  
(Wall Street Journal 1286/49)



---

# Other treebanks: Prague dependency family



## Prague Czech-English Dependency Treebank 1.0 (2004)

- Penn Treebank data (Wall Street Journal, 21 600 English sentences)
- human translators
- automatic conversions of Penn Treebank annotation into PDT-like annotation scheme (m-, a- and t-layers)
- plain text from Reader's Digest 1993-1996 (50 000 sentences)
- test data:
  - 515 sentence pairs
  - manually annotated on tectogrammatical level, Czech and English
  - retranslated from Czech to English by 4 different translation companies

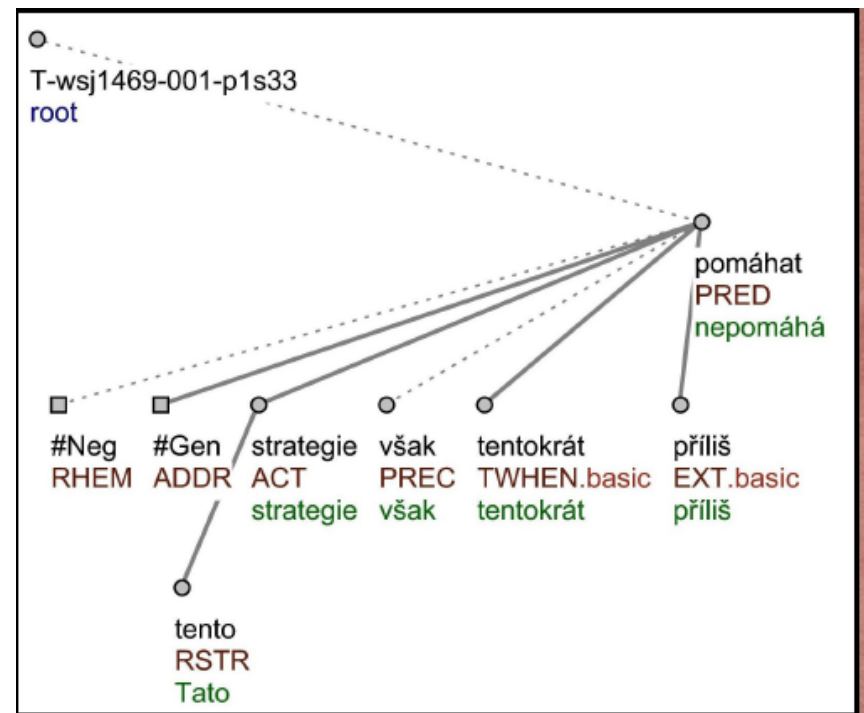
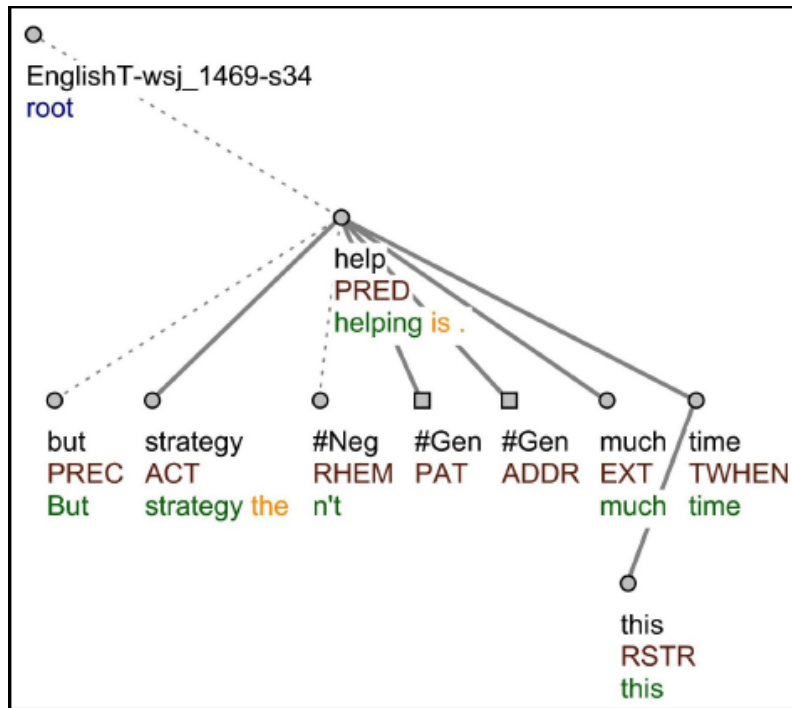


# Other treebanks: Prague dependency family

## Prague Czech-English Dependency Treebank 2.0



- Penn Treebank data
- manually annotated data (49 000 sentences)
- <http://ufal.mff.cuni.cz/pcedt2.0/>



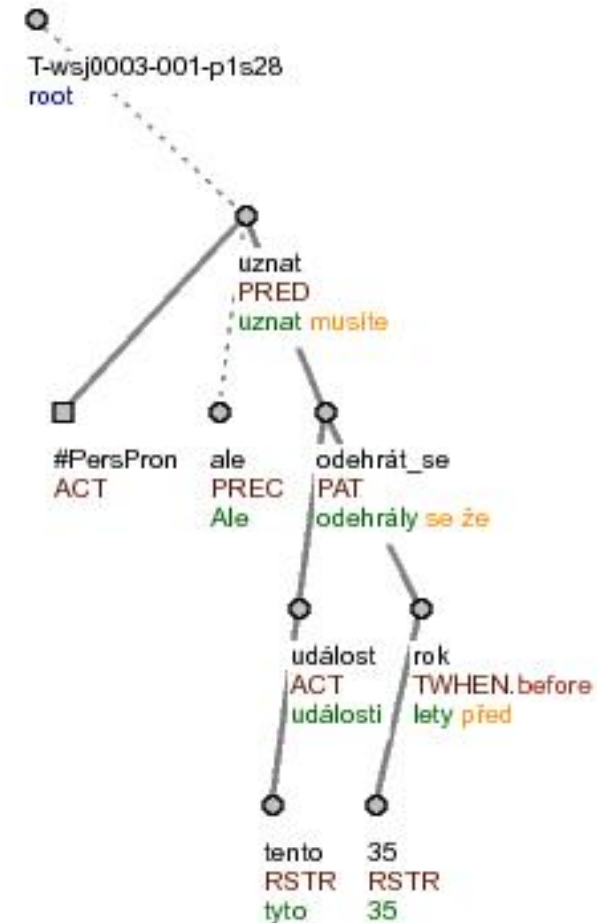
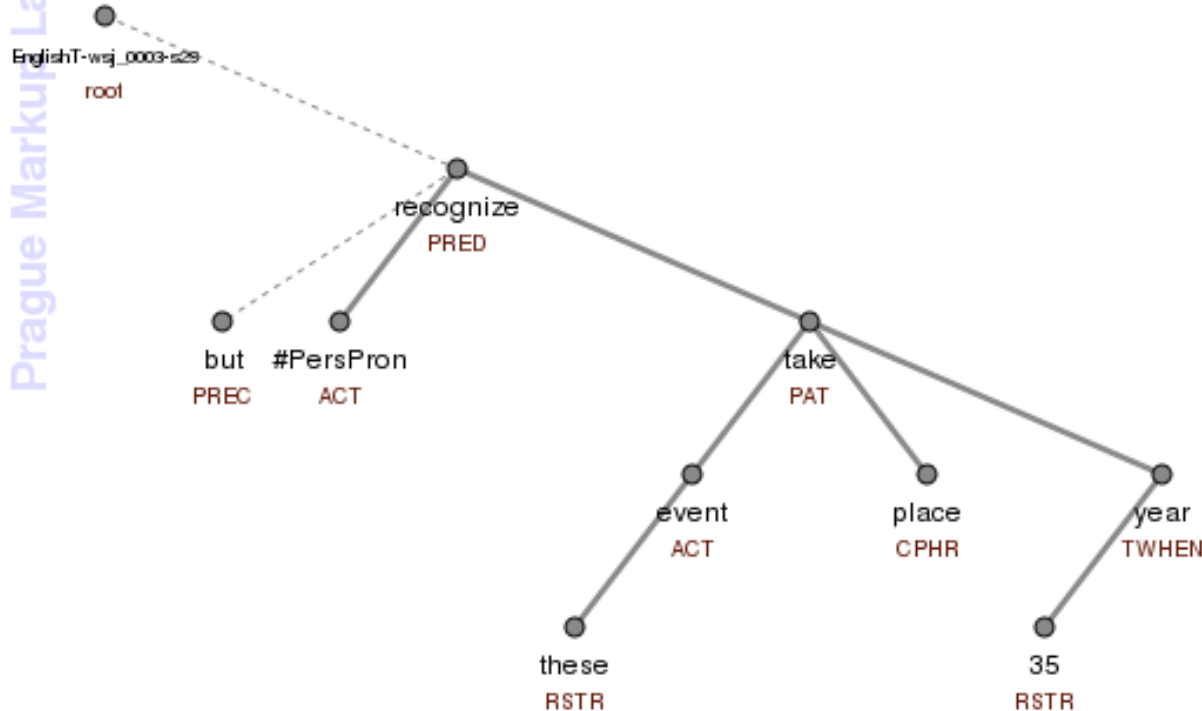
*But the strategy isn't helping much this time.*

*Tato strategie však tentokrát příliš nepomáhá .*

# Prague Czech-English Dependency Treebank

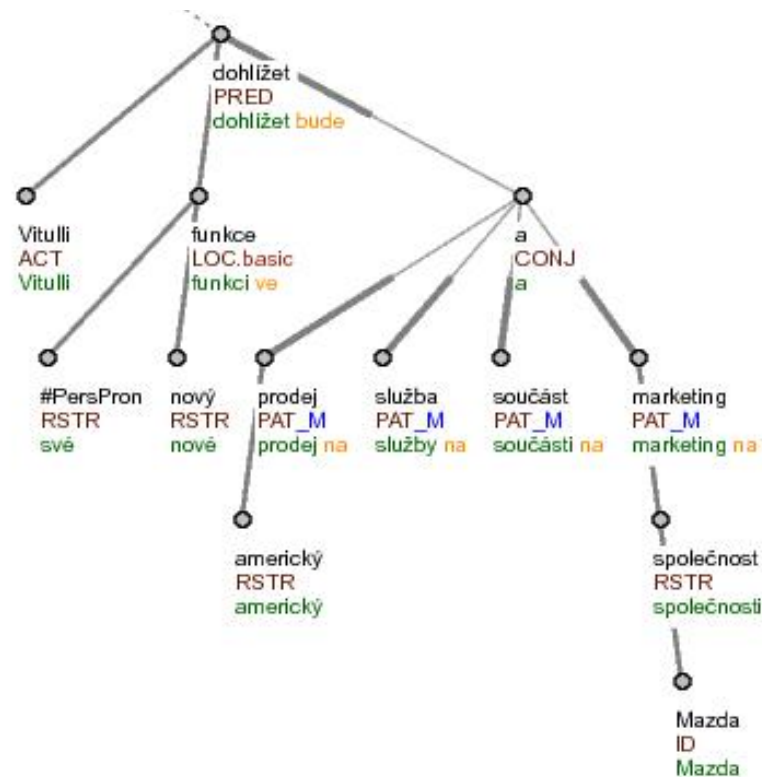
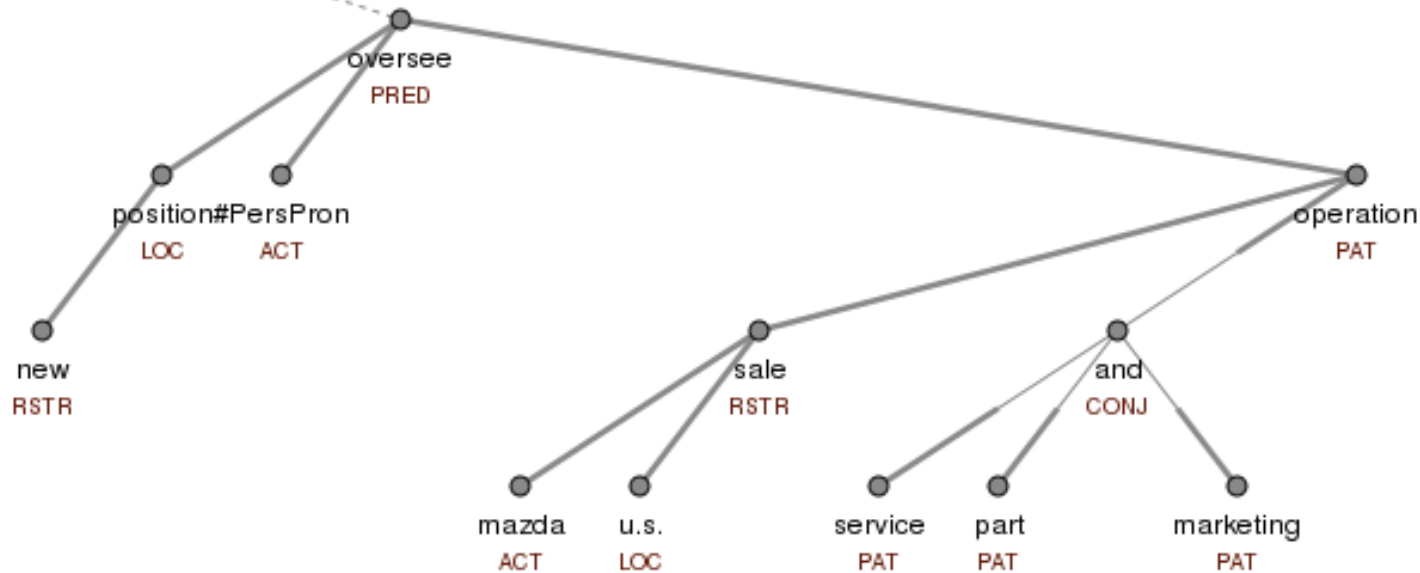


Prague Markup Language



*Ale musíte uznat, že se tyto události odehrály před 35 lety.*

*But you have \*-1 to recognize that these events took place 35 years ago.*

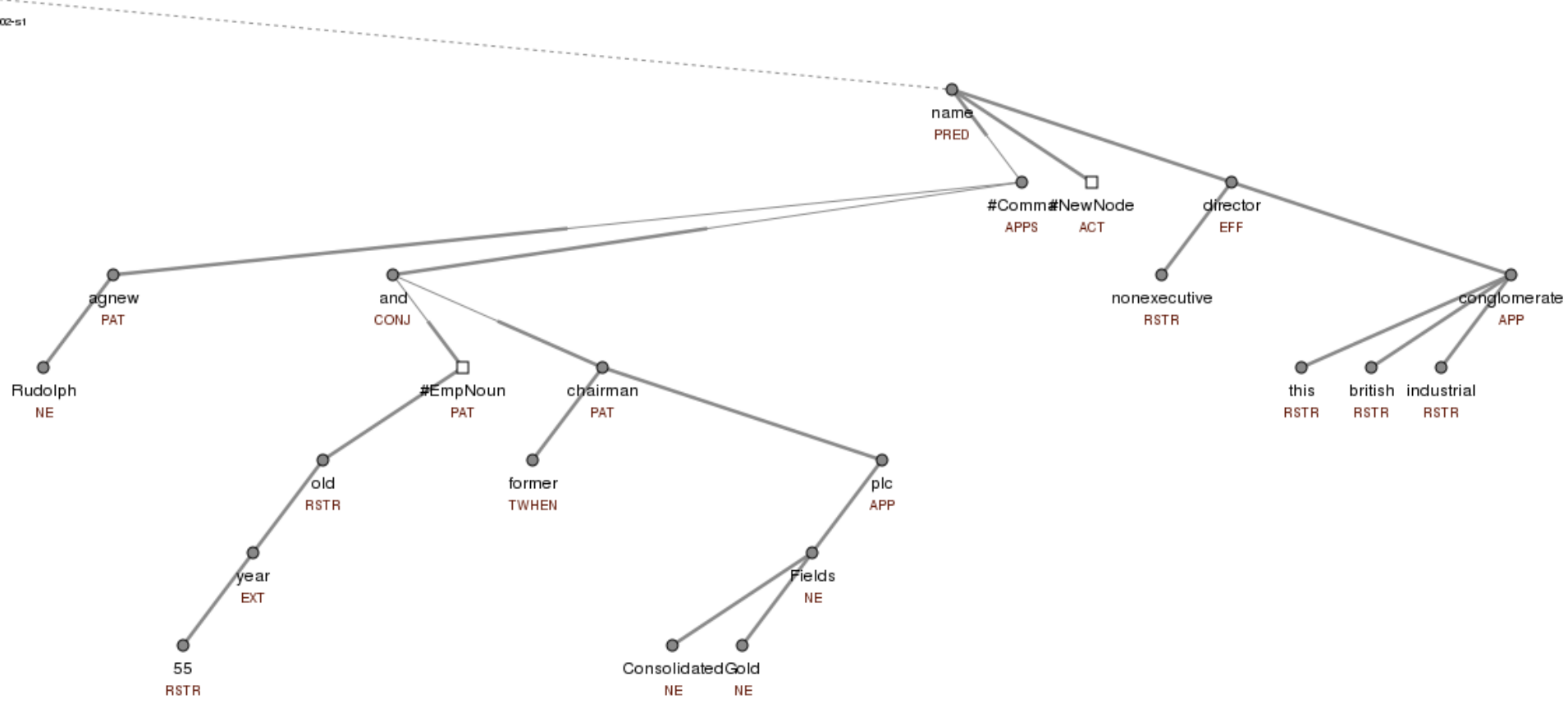
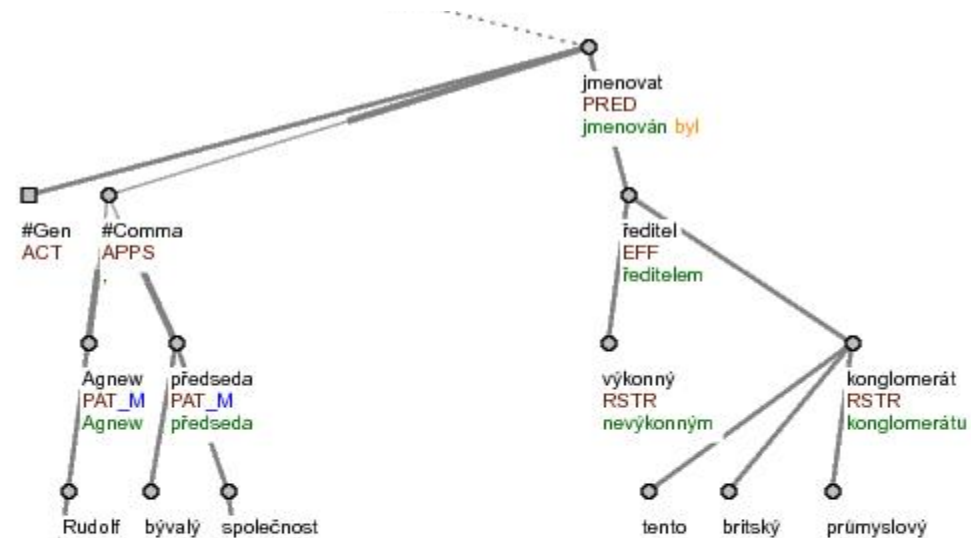


*In the new position he will oversee Mazda 's U.S. sales, service, parts and marketing operations .*

*Vitulli bude ve své nové funkci dohlížet na americký prodej, služby, součásti a marketing společnosti Mazda.*

*Pětapadesátiletý Rudolf Agnew, bývalý předseda společnosti Consolidated Gold Fields PLC, byl jmenován nevýkonným ředitelem tohoto britského průmyslového konglomerátu.*

*Rudolph Agnew , 55 years old and former chairman of Consolidated Gold Fields PLC , was named \*-1 a nonexecutive director of this British industrial conglomerate.*



# Other treebanks: Prague dependency family

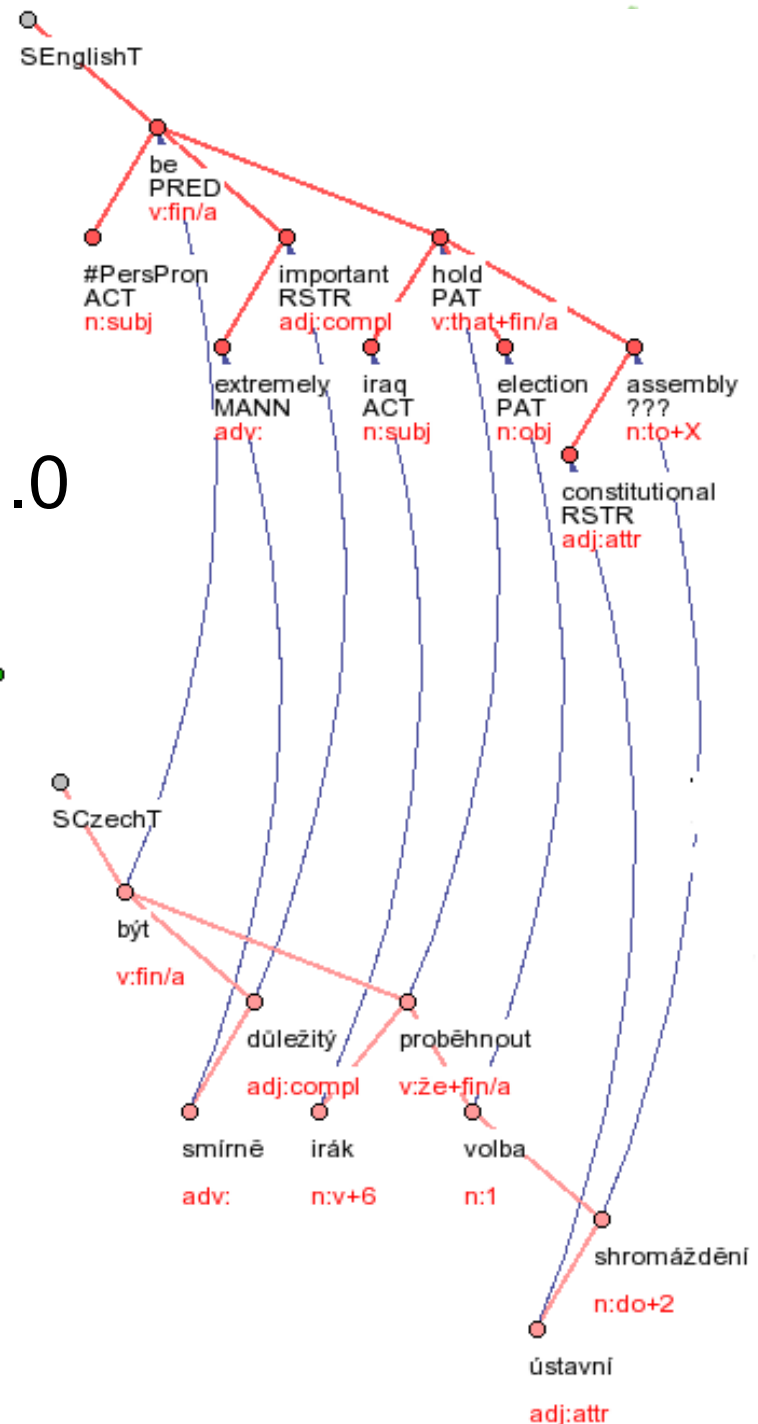
## Czech-English Parallel Corpus 1.0

(~15.0 M parallel sentences )

<http://ufal.mff.cuni.cz/czeng/>

- collected automatically
- annotated automatically
- European laws, subtitles, technical documentation, electronic books, newspapers, ...

*It is extremely important that Iraq held elections to a constitutional assembly.*





# Prague Arabic Dependency Treebank 1.0 (2004)

[http://ufal.mff.cuni.cz/padt/PADT\\_1.0/docs/index.html](http://ufal.mff.cuni.cz/padt/PADT_1.0/docs/index.html)

- Functional Arabic Morphology
- analytical layer  
(about 130 000 tokens)
- tectogrammatical layer

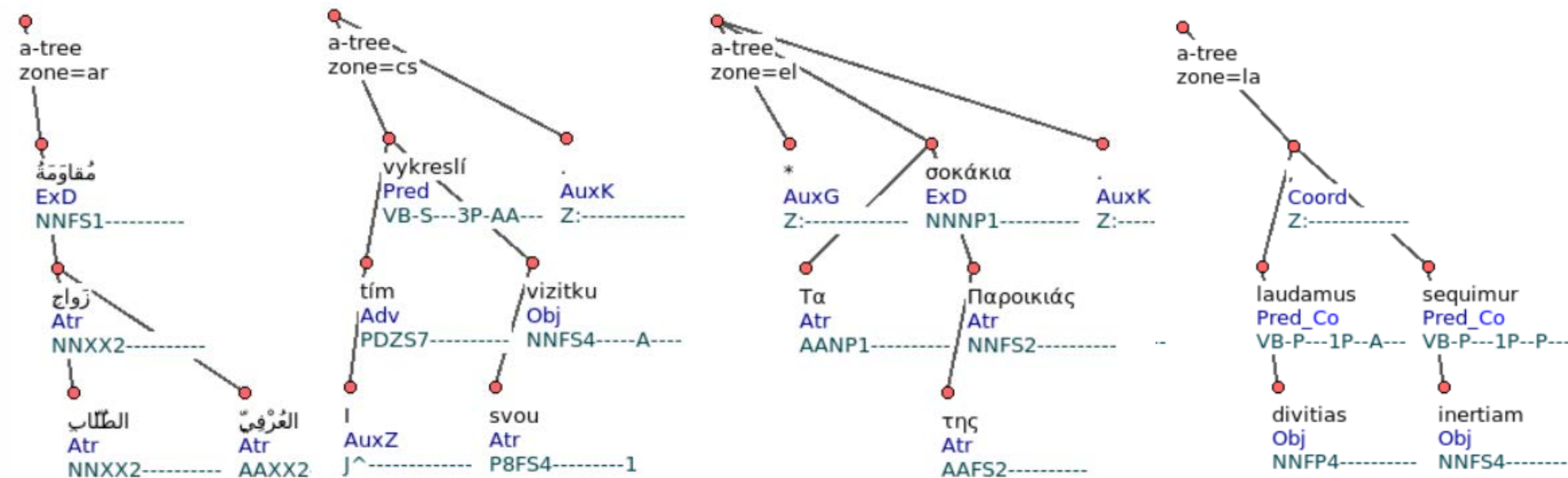


# Other treebanks: Prague dependency family



**HamleDT** ~ a compilation of existing dependency treebanks  
(or dependency conversions of other treebanks),  
transformed so that they all conform to the same annotation style  
29 treebanks

<http://ufal.mff.cuni.cz/hamledt/>





---

# References



- Sgall, P., Hajičová, E., Panevová, J. (1986) *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.
- Hajičová, E., Panevová, J., Sgall, P. (2002) *Úvod do teoretické a počítačové lingvistiky*, sv. I. Karolinum, Praha.
- PDT guide <http://ufal.mff.cuni.cz/pdt2.0/>
- PDT documentation
- Štěpánek, J. (2006) Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat). PhD thesis, MFF UK.