Prague Dependency Treebank: Introduction

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz

Prague Dependency Treebank

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

Prague Dependency Treebank

http://ufal.mff.cuni.cz/prague-dependency-treebank

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

- data
- tools
  - TrEd … graphical editor and interface for creating queries (practical lectures) http://ufal.mff.cuni.cz/tred/
- documentation: http://ufal.mff.cuni.cz/pdt3.5/documentation
  - Guide, http://ufal.mff.cuni.cz/pdt2.0/
  - manuals for individual layers
  - survey of data formats and tools

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	**49 442**	**38 538**
# 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 2000, …)
    http://ufal.mff.cuni.cz/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----------
oživení	*oživení_^(*3it)*	NNNS6-----A----
Havlovým	*Havlův_;S_^(*3el)*	AUIS7M---------
projevem	*projev*	NNIS7-----A----
zdají	*zdát*	VB-P---3P-AA---
být	*být*	Vf--------A----
jasnější	*jasný*	AAFP1----2A----
.	*.*	Z:----------

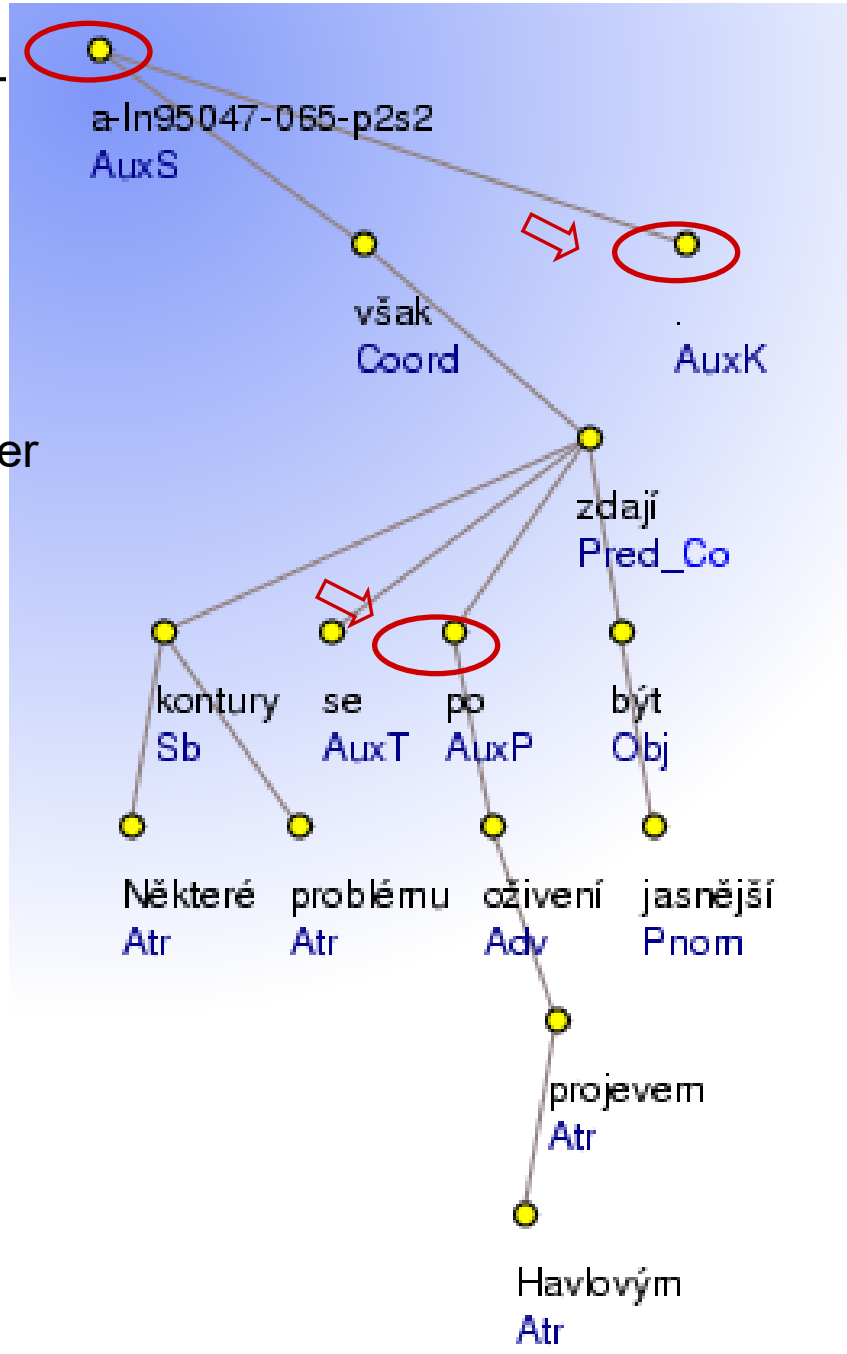
PDT: a-layer

- 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

  } coordination, apposition, parenthesis
  - id
  - reference to m-layer

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

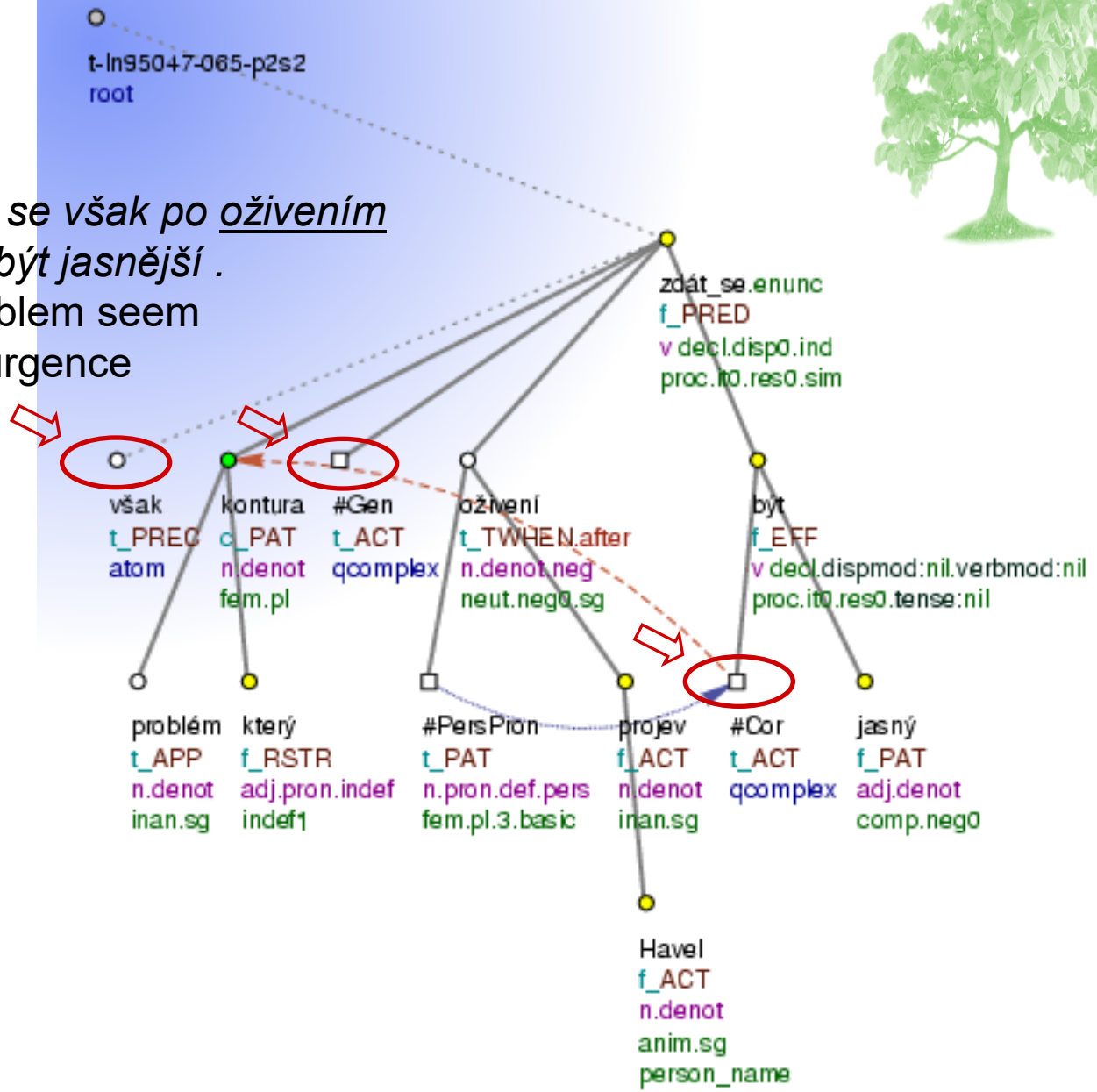
- tectogrammatical tree structure ~ dependency tree
  - nodes for auto-semantic/lexical words only
    syn-semantic/functional words as attributes of lexical words
    (plus technical root)
  - ellipses as nodes
  - edges ~ relations (dependency, coordination, others)
  - link to a valency lexicon for 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
  - ...

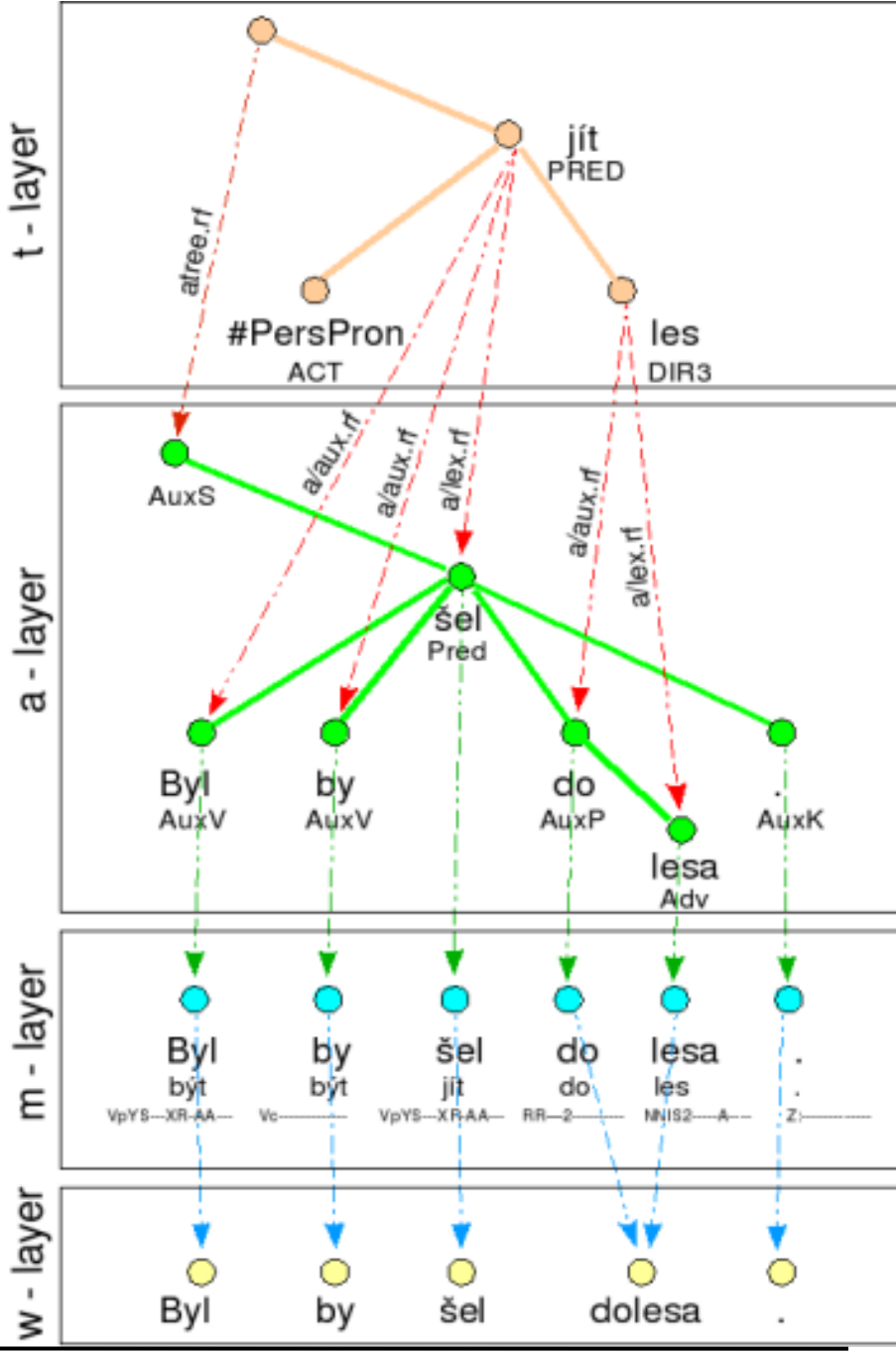
PDT: t-layer

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.]

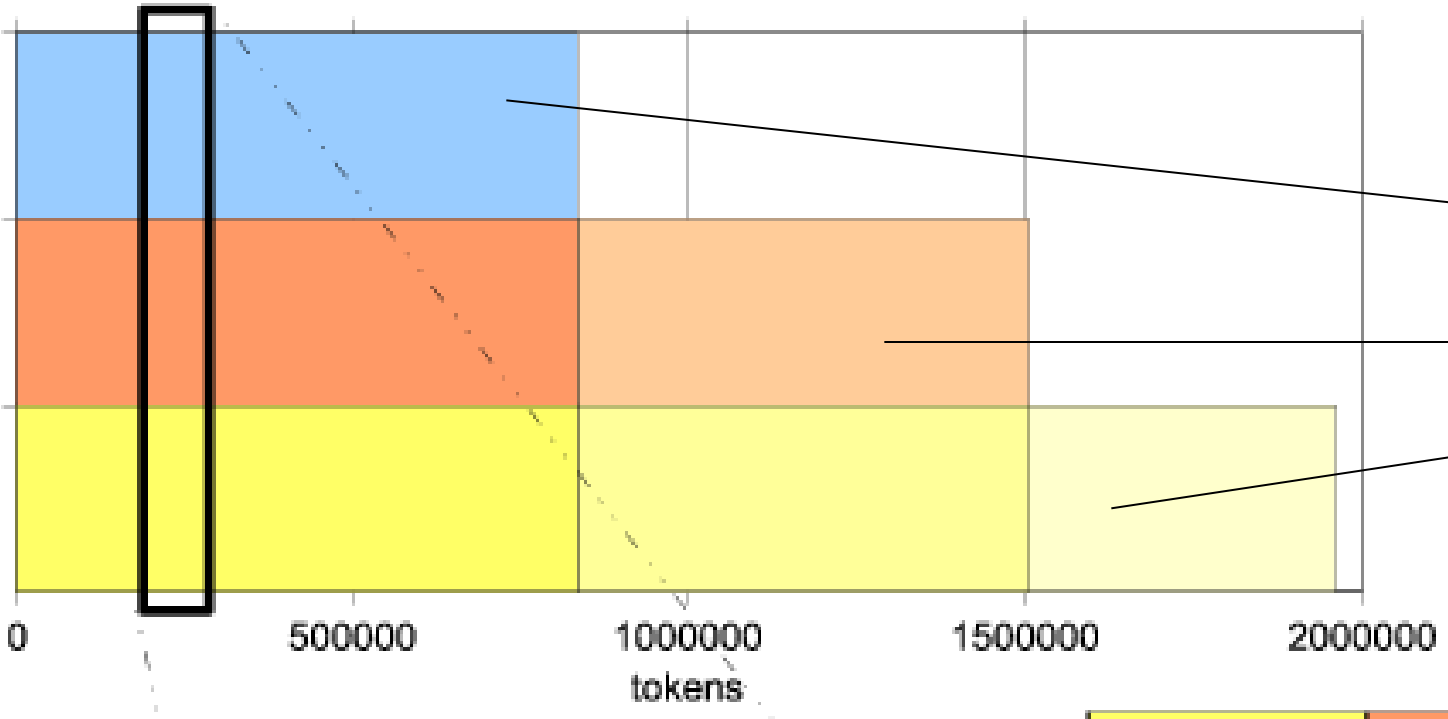


Linking the layers

- 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***

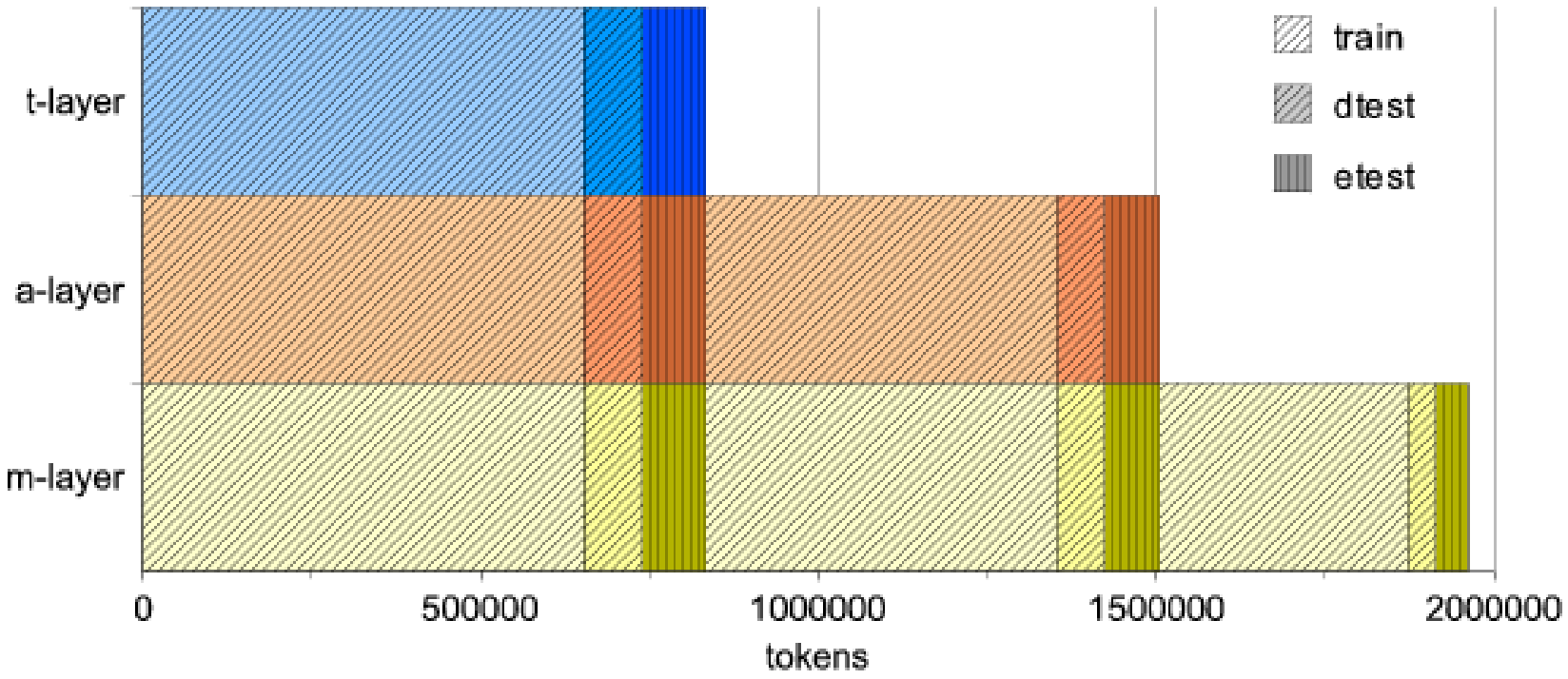


PDT: Division of the data to layers

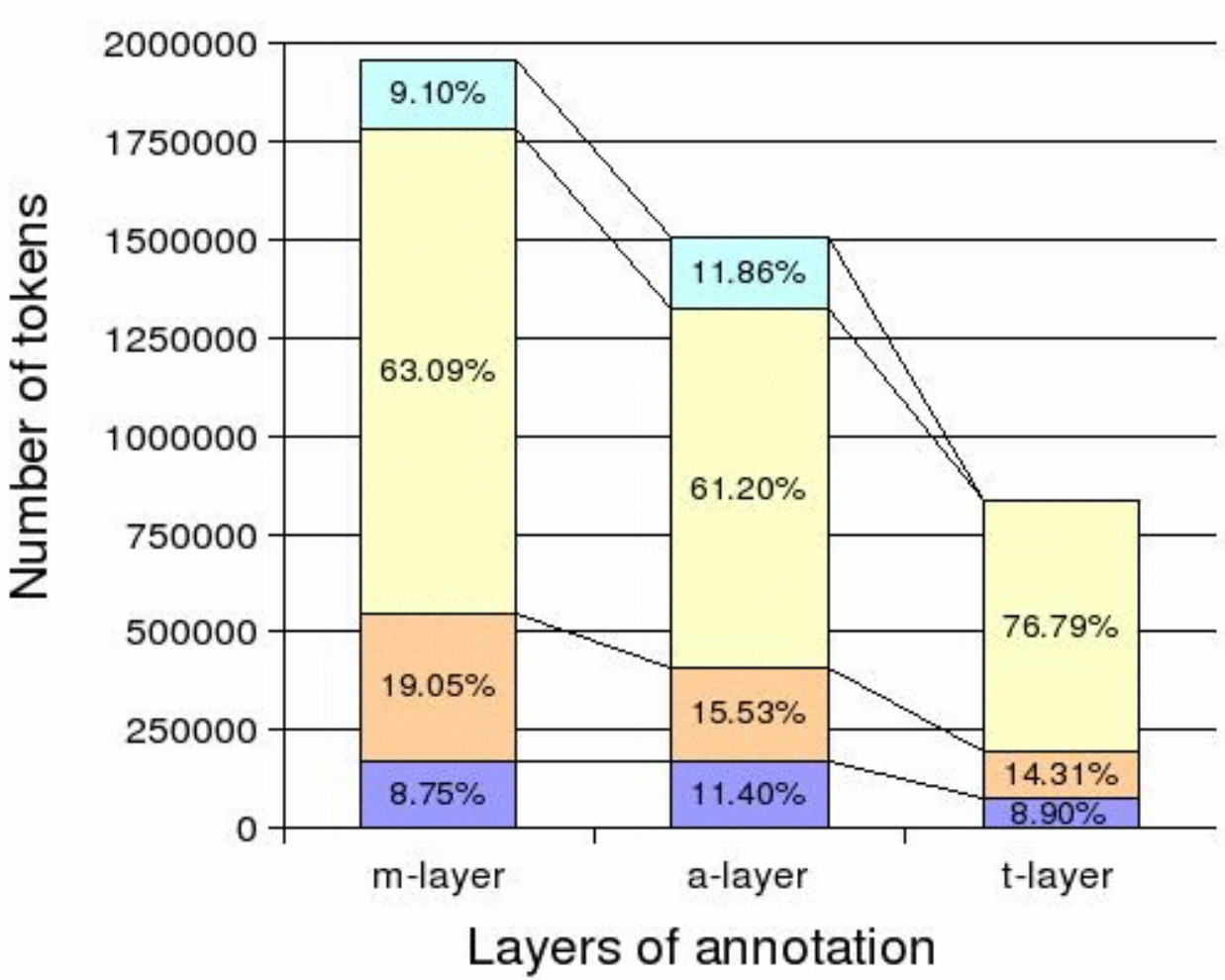


ln94206_1.m.gz + ln94206_1.w.gz	ln94206_1.a.gz	ln94206_1.t.gz
ln94206_2.m.gz + ln94206_2.w.gz	ln94206_2.a.gz	ln94206_2.t.gz
ln94206_3.m.gz + ln94206_3.w.gz	ln94206_3.a.gz	ln94206_3.t.gz

PDT: Division of the data into training and test sets



PDT: Number of tokens from the particular sources



Other treebanks: Prague dependency family

Czech – written

- Prague Discourse Treebank http://ufal.mff.cuni.cz/pdit2.0

  1.0 (2001); 2.0 (2006); added to PDT 3.0 (2016) … 50 k sentences as a new layer of language description:
  - extended textual coreference (incl. 1st and 2nd person), bridging anafora
  - discourse relations (explicit connectives)
- Czech Academic Corpus 2.0 (2008) http://ufal.mff.cuni.cz/cac
  - morphological annotation (652 k tokens, 32 k sentences)
  - analytical annotation (493 k tokens, 25 k sentences)
  - both written and spoken language; manually annotated
- Czech Legal Text Treebank

  http://ufal.mff.cuni.cz/czech-legal-text-treebank
  - morphological and analytical annotation; manually annotated (1,121 sent.)
  - the layer of accounting entities, and the layer of semantic entity relations

Other treebanks: Prague dependency family

Czech – spoken:

- Prague Dependency Treebank of Spoken Czech

  http://ufal.mff.cuni.cz/pdtsc2.0

  74 k sentences ≈ over 100 hours (742 k tokens)
  spontaneous dialogs
  several layers:
  - audio recordings
    - testimonies from the MALACH project
      https://malach.umiacs.umd.edu/ http://ufal.mff.cuni.cz/malach/en
    - Companions project
  - automatic and manual transcripts
  - manually reconstructed text
    with morphological annotation
  - analytical layer
  - deep syntax … manual annotation

  } PDT-like

Other treebanks: Prague dependency family

- Prague Czech-English Dependency Treebank 2.0
  - Penn Treebank data (English)
  - translated by profesinal translators
  - 49 k parallel sentences, 1:1 sentence-aligned
    http://ufal.mff.cuni.cz/pcedt2.0/

Czech part:

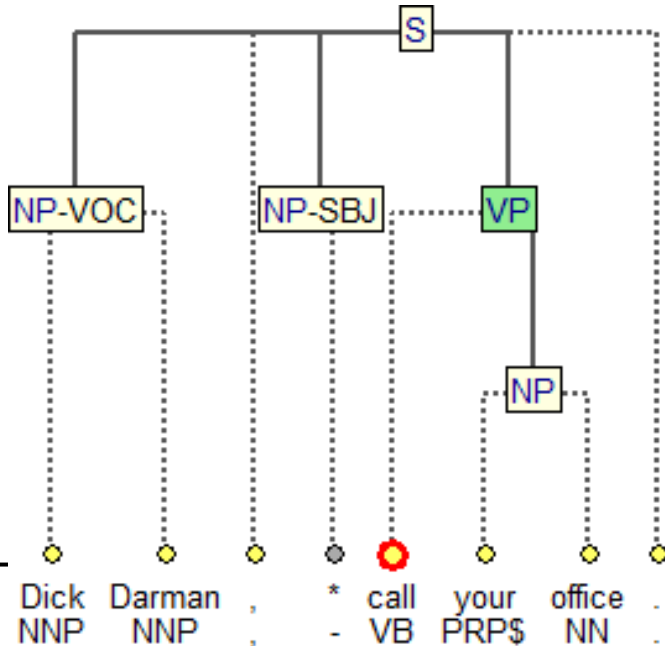
- w-layer
- m-layer
- a-layer
- t-layer … manual

English part:

- original Penn Treebank annotation p-layer

→ w-layer
m-layer
a-layer

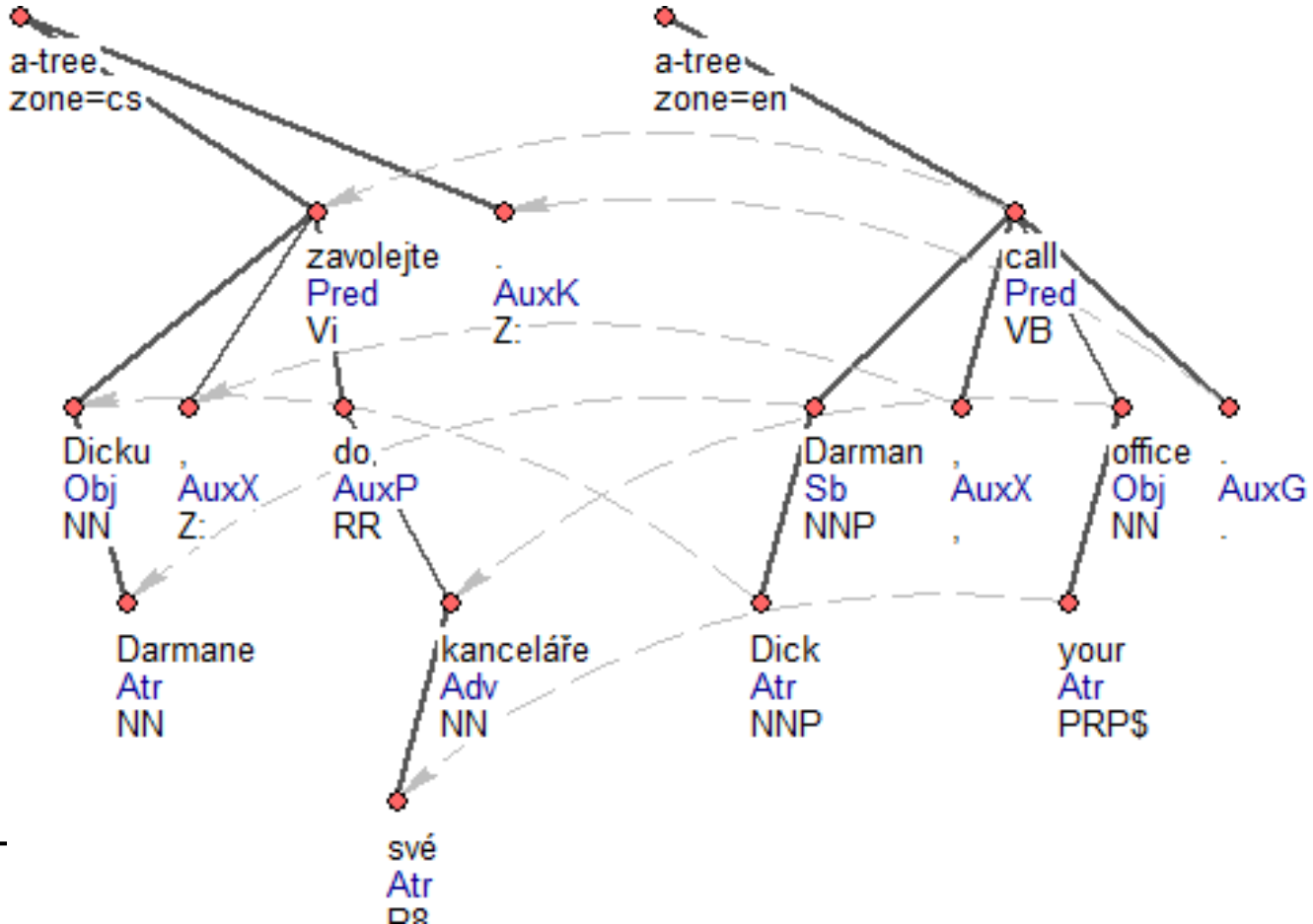
- t-layer … manual



Other treebanks: Prague dependency family

- Prague Czech-English Dependency Treebank 2.0
  - automatically aligned on the node-level

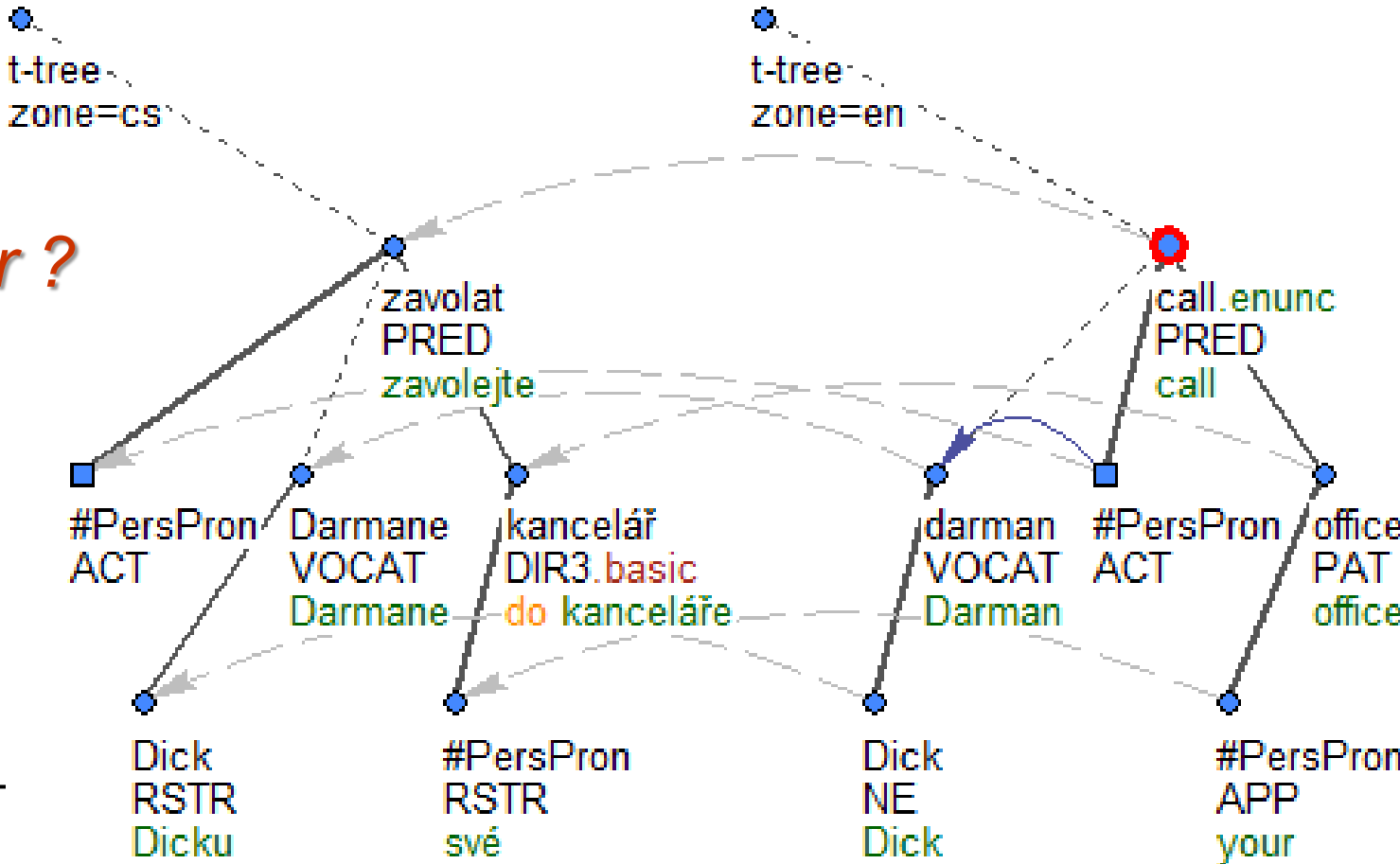
Which layer ?



Other treebanks: Prague dependency family

- Prague Czech-English Dependency Treebank 2.0
  - automatically aligned on the node-level

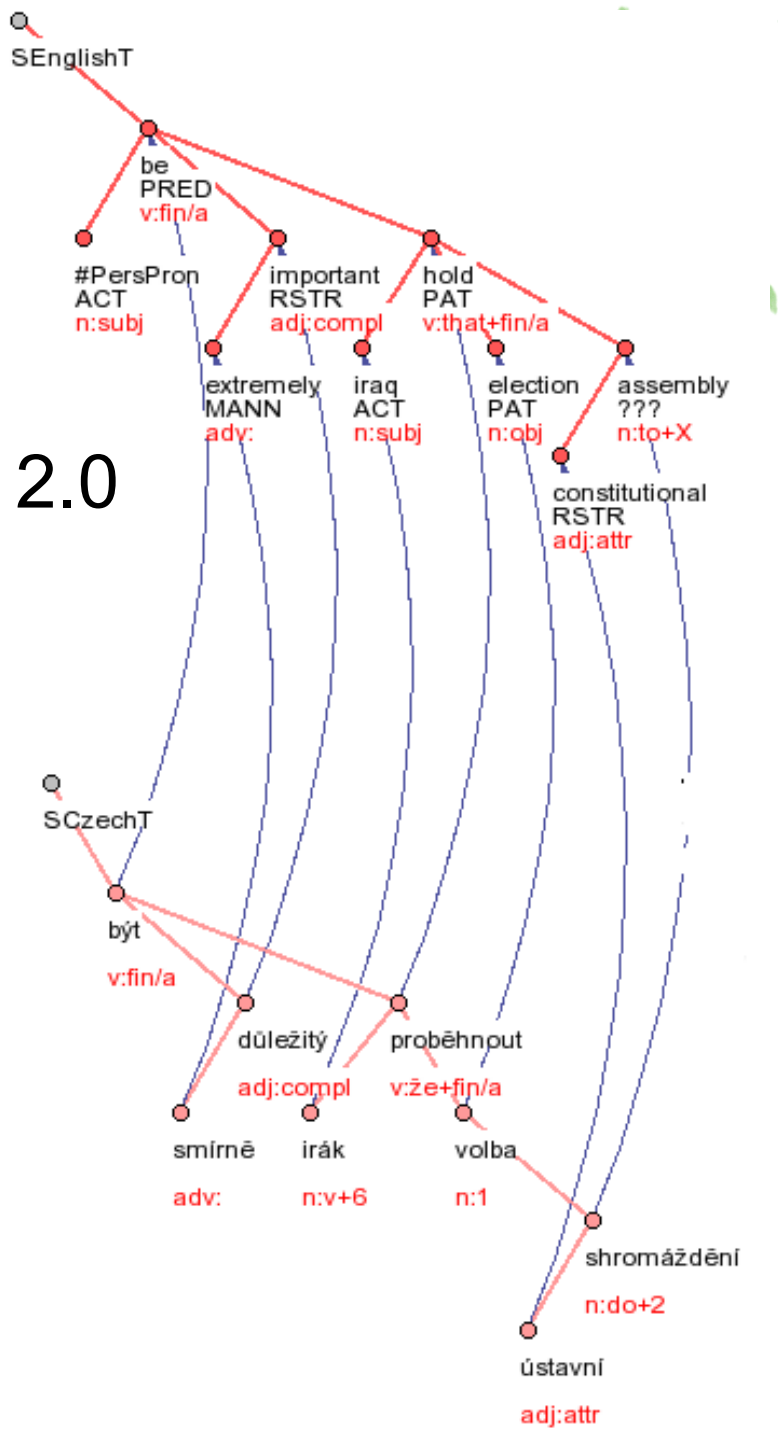
Which layer ?



Other treebanks: Prague dependency family

- Czech-English Parallel Corpus 2.0
  (~ 180 M parallel sentences)
  http://ufal.mff.cuni.cz/czeng/
  - collected automatically
  - annotated automatically
  - European laws, subtitles, technical documentation, electronic books, newspapers, …

  used in Machine Translation Task
  (WMT conference 2008-2020
  http://www.statmt.org/



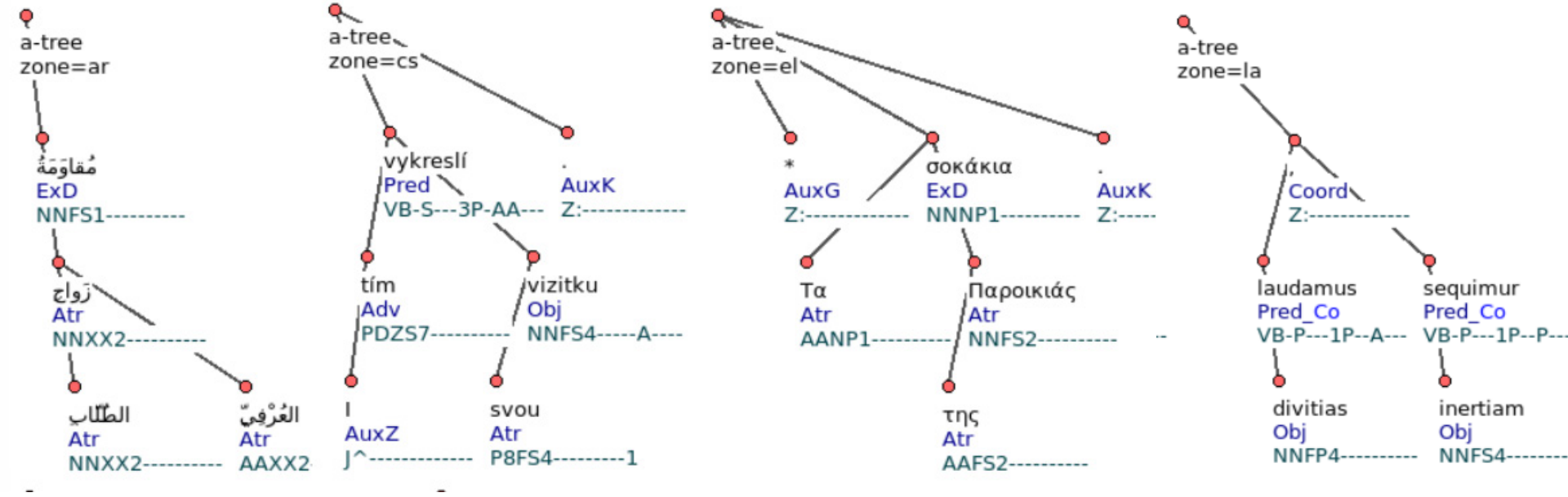
It is extremely important that Iraq held elections to a constitutional assembly.

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 … 2012

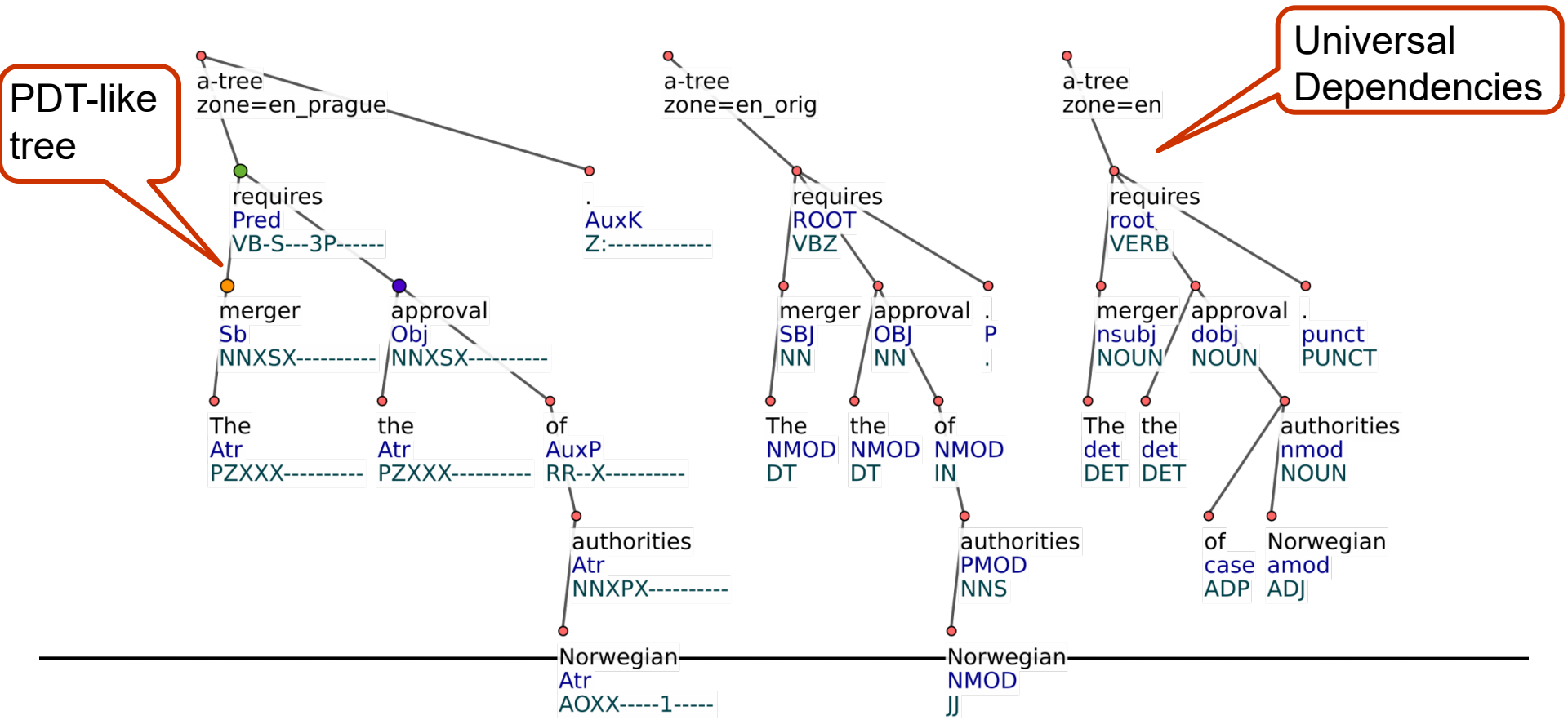
  36 languages, 42 treebanks in HamleDT 3.0 (2015)

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



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

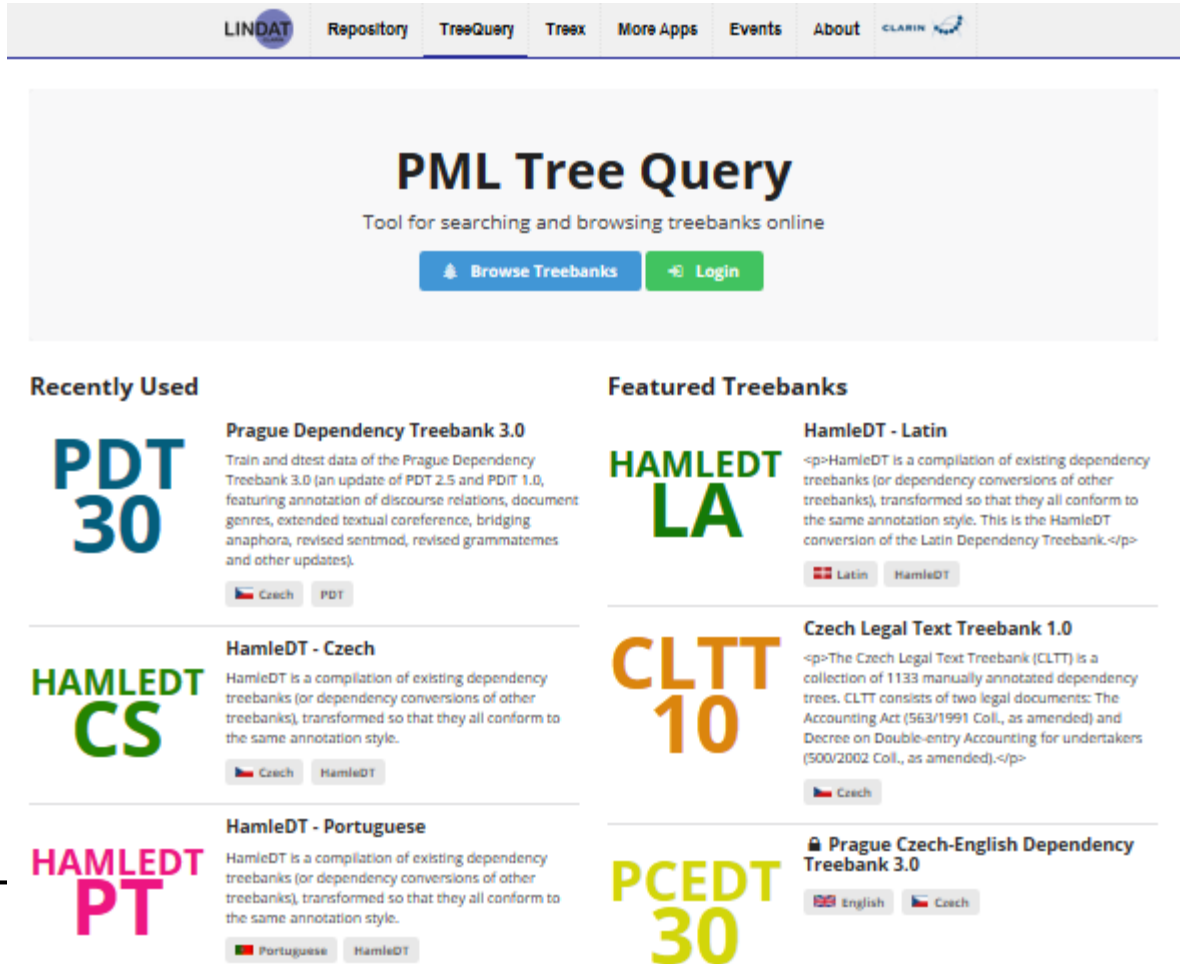


How to access / obtain dependency treebanks

as a web service … LINDAT/CLARIAH-CZ Repository

http://lindat.mff.cuni.cz/services/pmltq/#!/home

PML-TQ search tool



How to access / obtain dependency treebanks

- **as a web service**

  https://lindat.mff.cuni.cz/services/pmltq/#!/home

  LINDAT/CLARIAH-CZ Repository

  PML-TQ search tool

  more stable, quick

- **via Tred instalation**

  PML-TQ search tool

  graphical interface for creating queries (practical lectures)

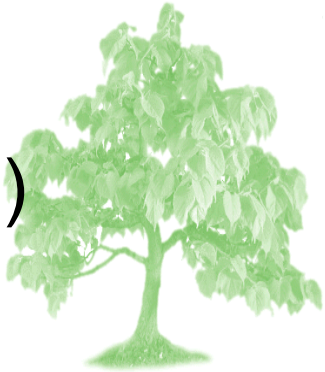
Differences between FGD and PDT

FGD	PDT
• tectogrammar/deep syntax	• t-layer (tectogrammatical l.)
• surface syntax	• a-layer (analytical l.)
• morphematics	• m-layer (morphological l.)
	structural layers
• morphonology	
• phonology	• w-layer (word layer)

reasons

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

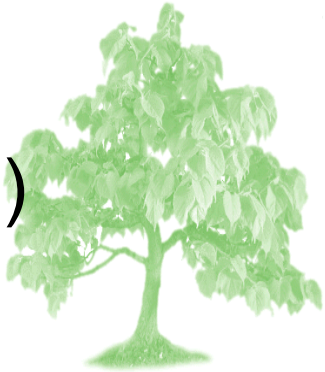
Differences between FGD and PDT (cont.)



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

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

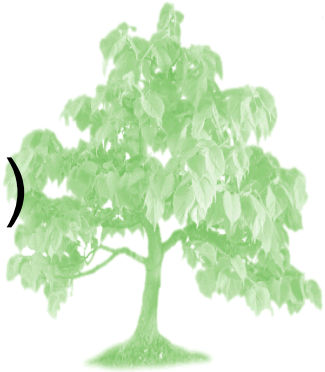
Differences between FGD and PDT (cont.)



structural layers

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

Differences between FGD and PDT (cont.)



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)

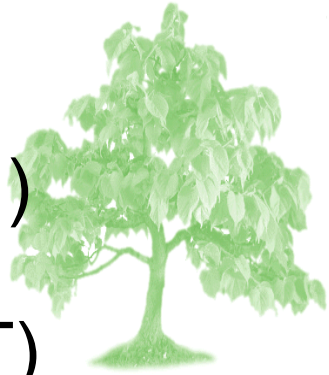
⇨ edges for non-dependency relations
(other than coordination/apposition)

- 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.)



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 (3rd person)
  - grammatical control, complement

References

- 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.
- Hajičová, E., Panevová, J., Sgall, P. (2002) *Úvod do teoretické a počítačové lingvistiky*, sv. I. Karolinum, Praha.