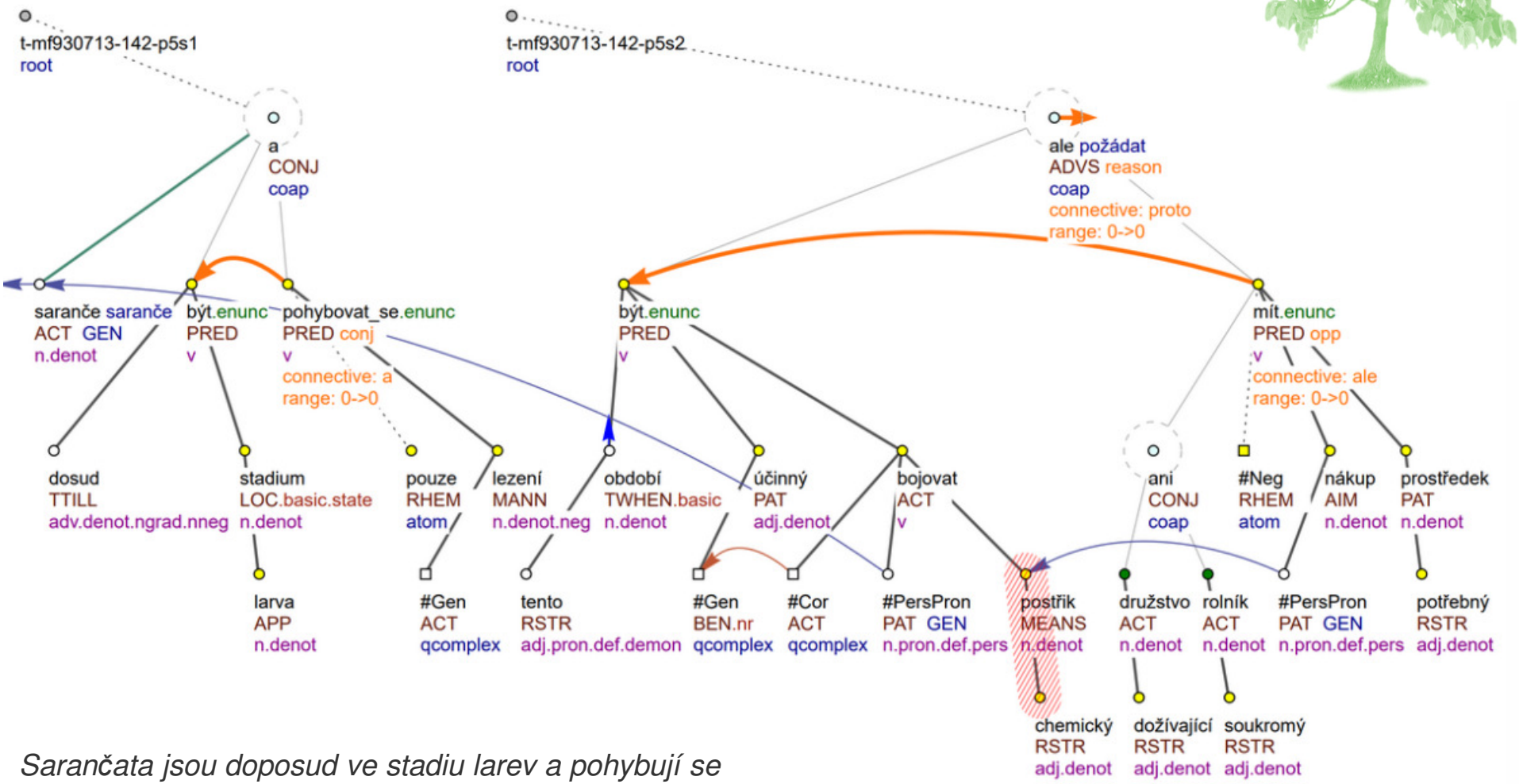
Family of Prague Dependency Treebanks: Introduction

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz

Prague Dependency Treebank (PDT)



Sarančata jsou doposud ve stadiu larev a pohybují se pouze lezením. V tomto období je účinné bojovat proti nim chemickými postřiky, ale dožívající družstva ani soukromí rolníci nemají na jejich nákup potřebné prostředky.

Lit: *Grasshoppers are still in the larvae stadium, crawling only. At this time of the year, it is efficient to fight them using chemicals, but neither the ailing cooperatives nor private farmers can afford them.*

Prague Dependency Treebanks - Czech

application of the FGD theory on the large set of Czech data
→ family of Prague Dependency Treebanks

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

- 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 Treebanks - Czech

application of the FGD theory on the large set of Czech data
→ family of Prague Dependency Treebanks

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

Prague Dependency Treebank – Consolidated (PDT-C) includes:

- Prague Dependency Treebank (PDT 3.5) … written texts
- Prague Czech-English Dependency Treebank (PCEDT 2.0 and PCEDT 2.0 Coref … translated data (Czech part)
- Prague Dependency Treebank of Spoken data (PDTSC 2.0) … spoken data
- PDT-FAUST …"user-generated" texts, unpublished data

https://ufal.mff.cuni.cz/pdt-c

Prague Dependency Treebanks - Czech

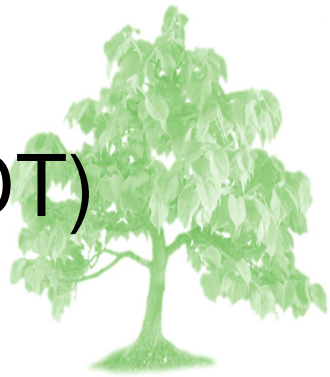
4 layers:

- raw text (word) layer (w-layer)
- morphological layer (m-layer)
- analytical layer (a-layer) = surface syntax
- tectogrammatical layer (t-layer) = deep s.
- audio and speech recognition layer (z-layer)

} ***layers of annotation***

	PDT (written)	PCEDT (translated)	PDTSC (spoken)	PDT-Faust (user-generated)	total
morphological layer (# of m-forms)	1,957,150	1,152,289	742,316	33,836	3,885,591
analytical layer (# of a-nodes)	1,503,741	1,152,289	742,316	33,837	3,432,183
tectogrammatical layer (# of t-nodes)	675,034	932,334	608,472	30,105	2,245,945

"The" Prague Dependency Treebank (PDT)



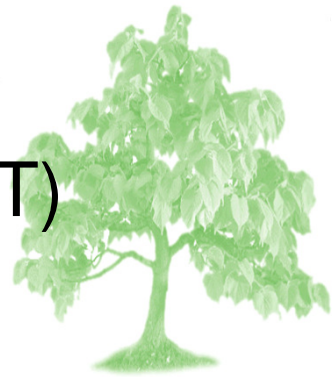
4 layers:

- raw text (word) layer (w-layer)
- morphological layer (m-layer)
- analytical layer (a-layer) = surface syntax
- tectogrammatical layer (t-layer) = deep s.
- ~~audio and speech recognition layer (z-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

“The” Prague Dependency Treebank (PDT)



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

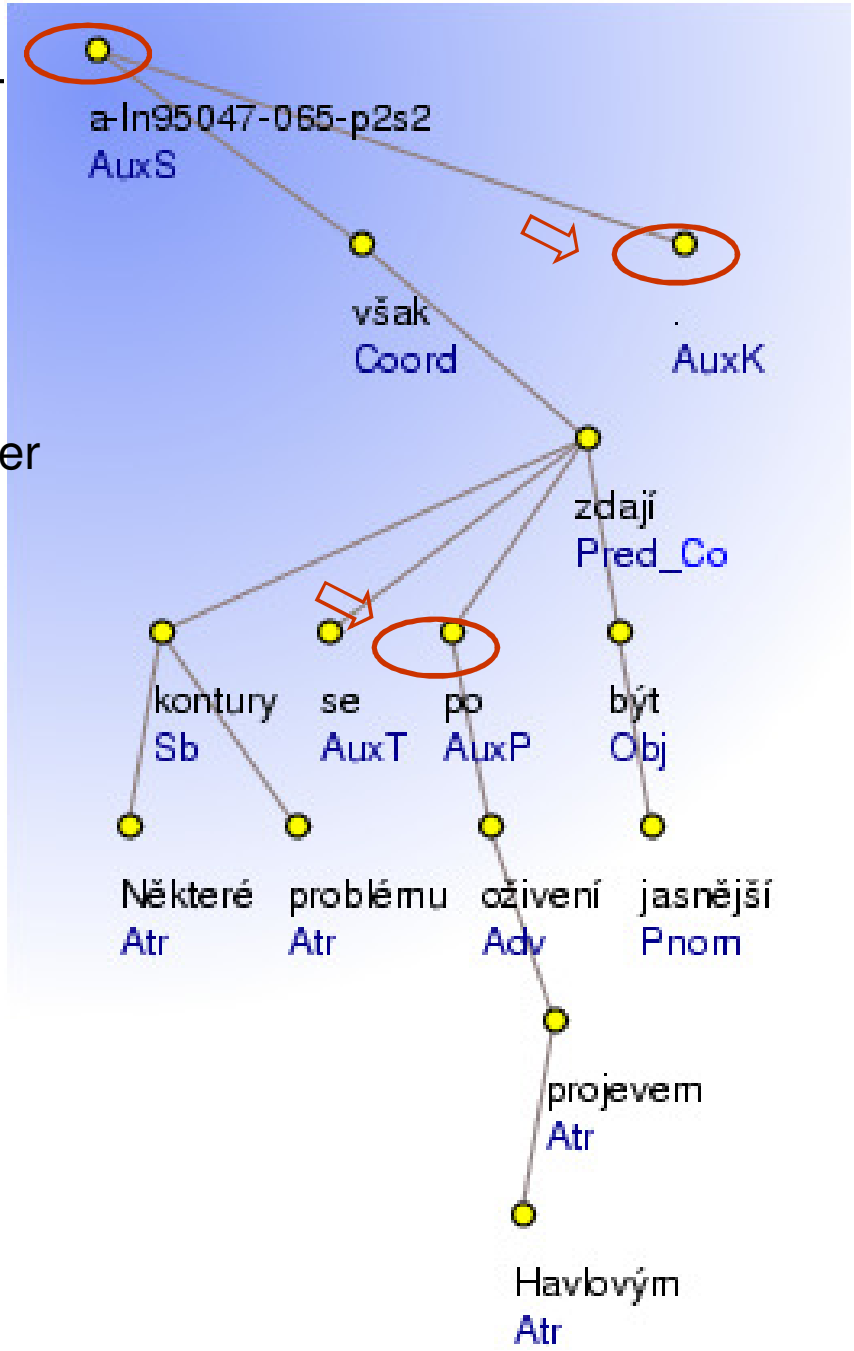
PDT: a-layer

- 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 … Sb, Obj, Atr, …)
  - *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

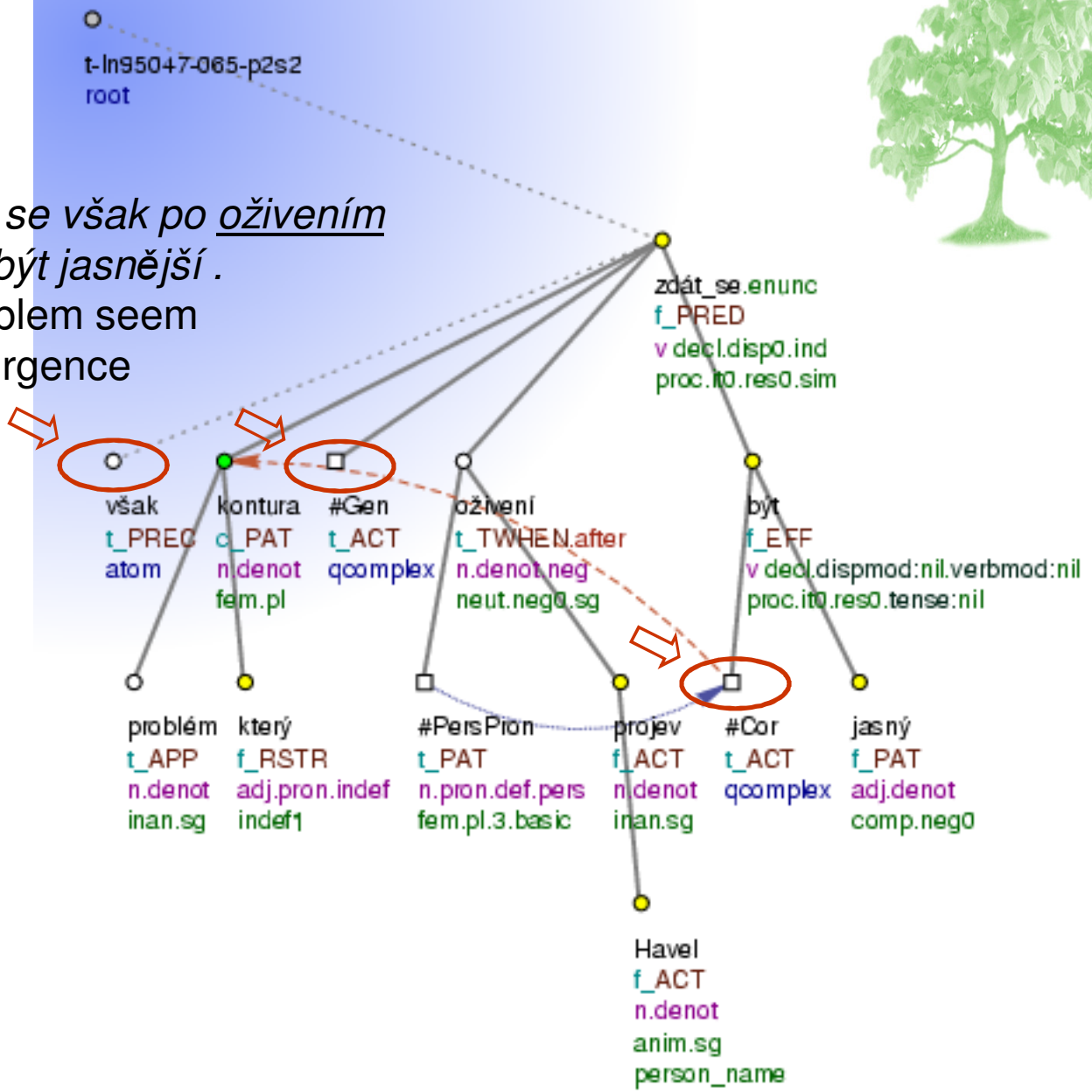
- tectogrammatical *tree structure* ~ dependency tree
  - nodes for content words only
    functional words as attributes of lexical words
    (plus technical root)
  - ellipses as nodes
  - edges ~ relations (dependency, coordination, others)
  - link to a valency lexicon for verbs and (certain types of) nouns
- *topic-focus articulation* (TFA)
  - linear ordering ~ deep word order
  - contextually bounded and unbounded nodes
- *coreference*
- *discourse annotation* added to PDT 3.0 (2016):
  - extended textual coreference (incl. 1st and 2nd person), bridging anaphora
  - discourse relations (explicit connectives)

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.]

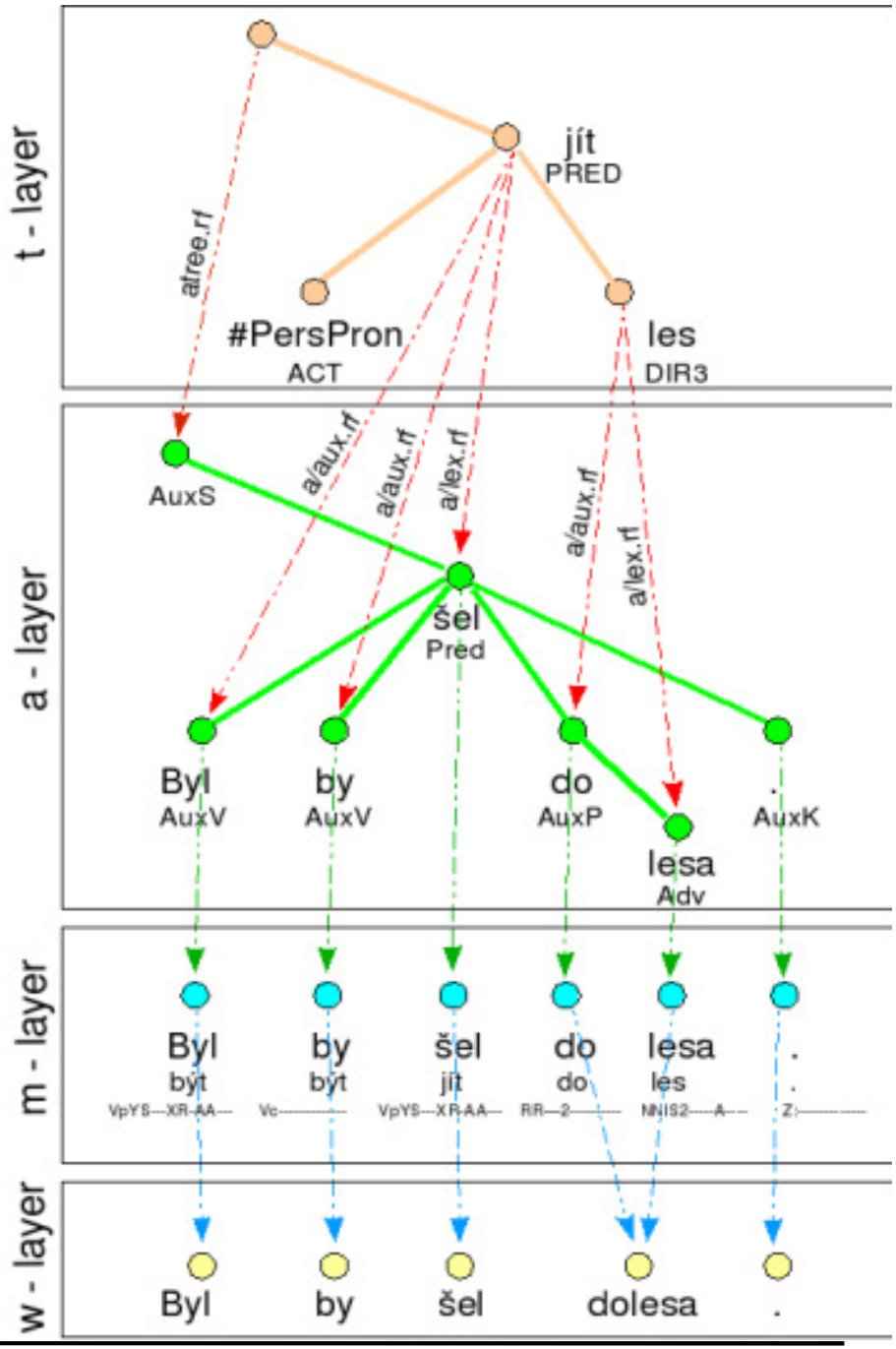


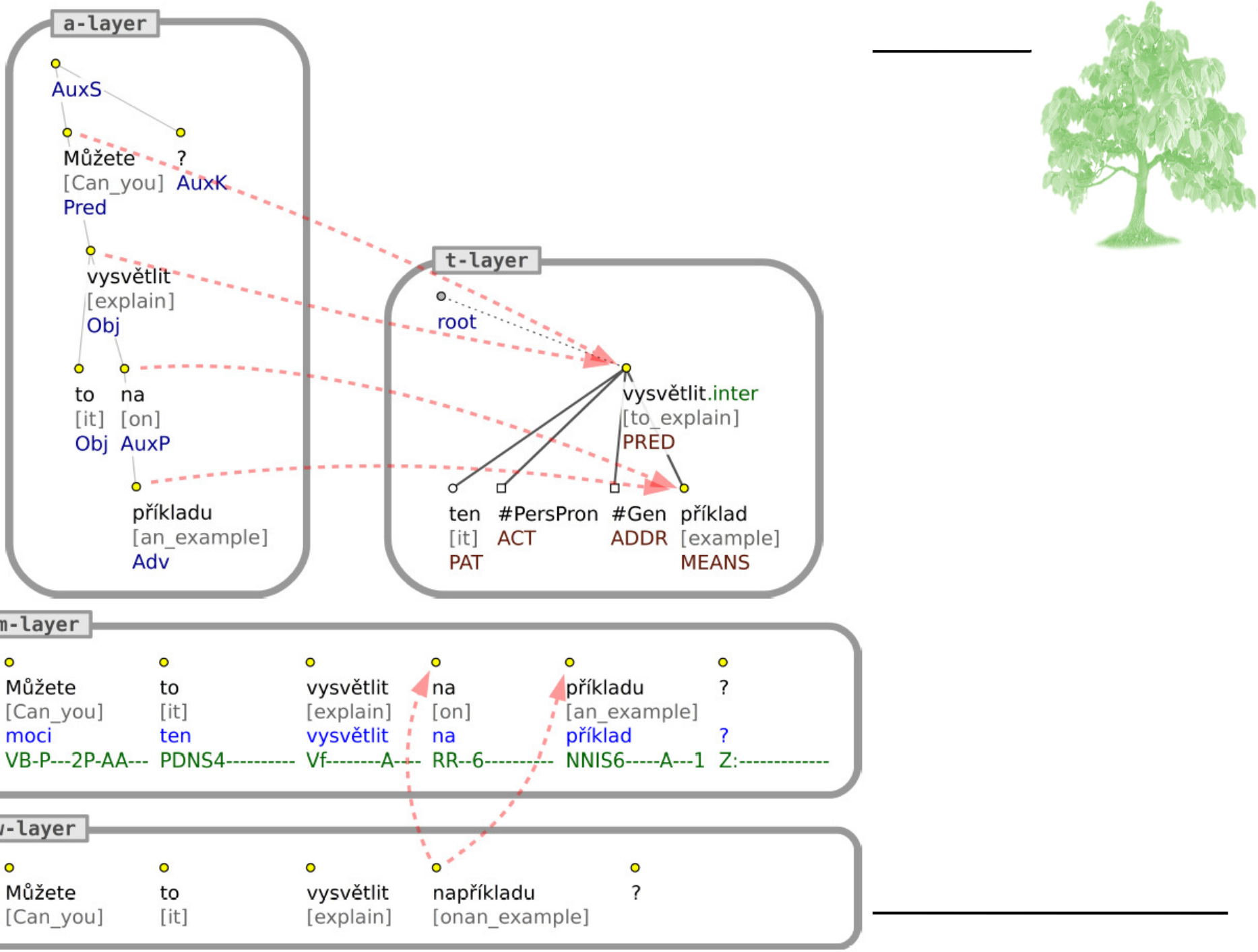
Linking the layers

- 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*

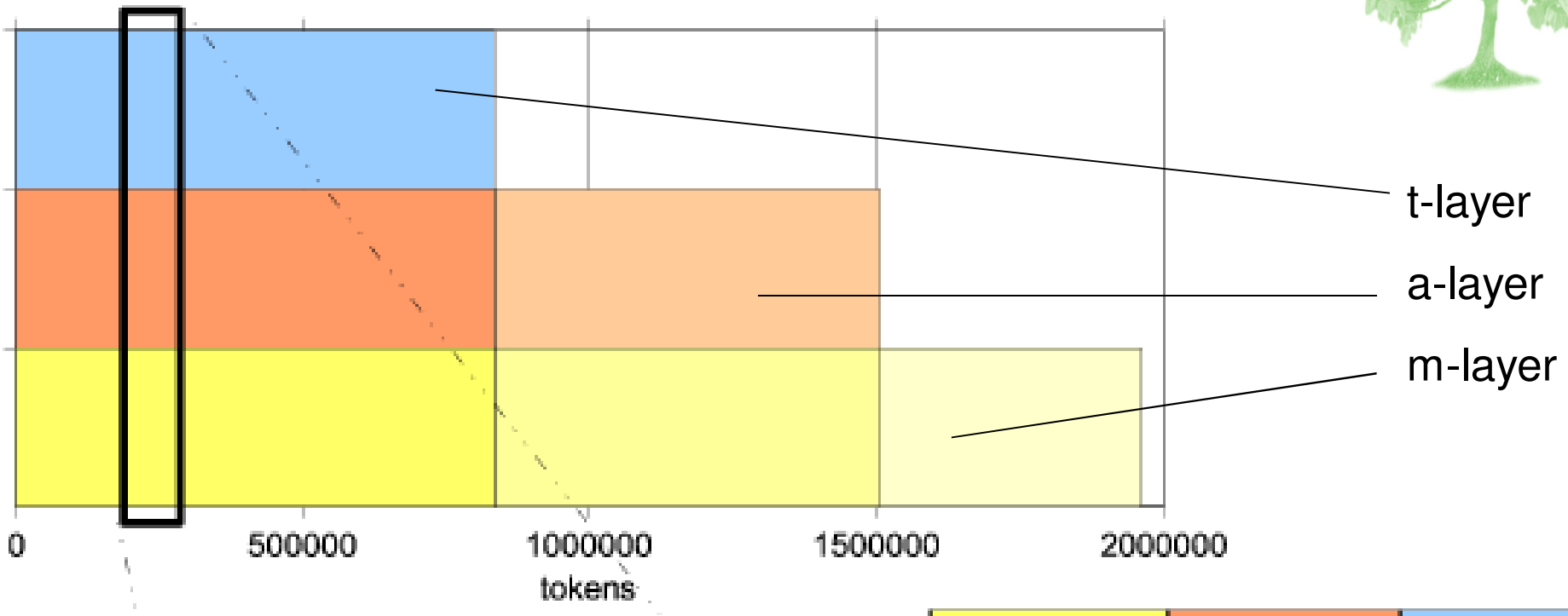
(He) would - have - gone - to - (the) forest - .

(He) would - have - gone - to (the) forest - .



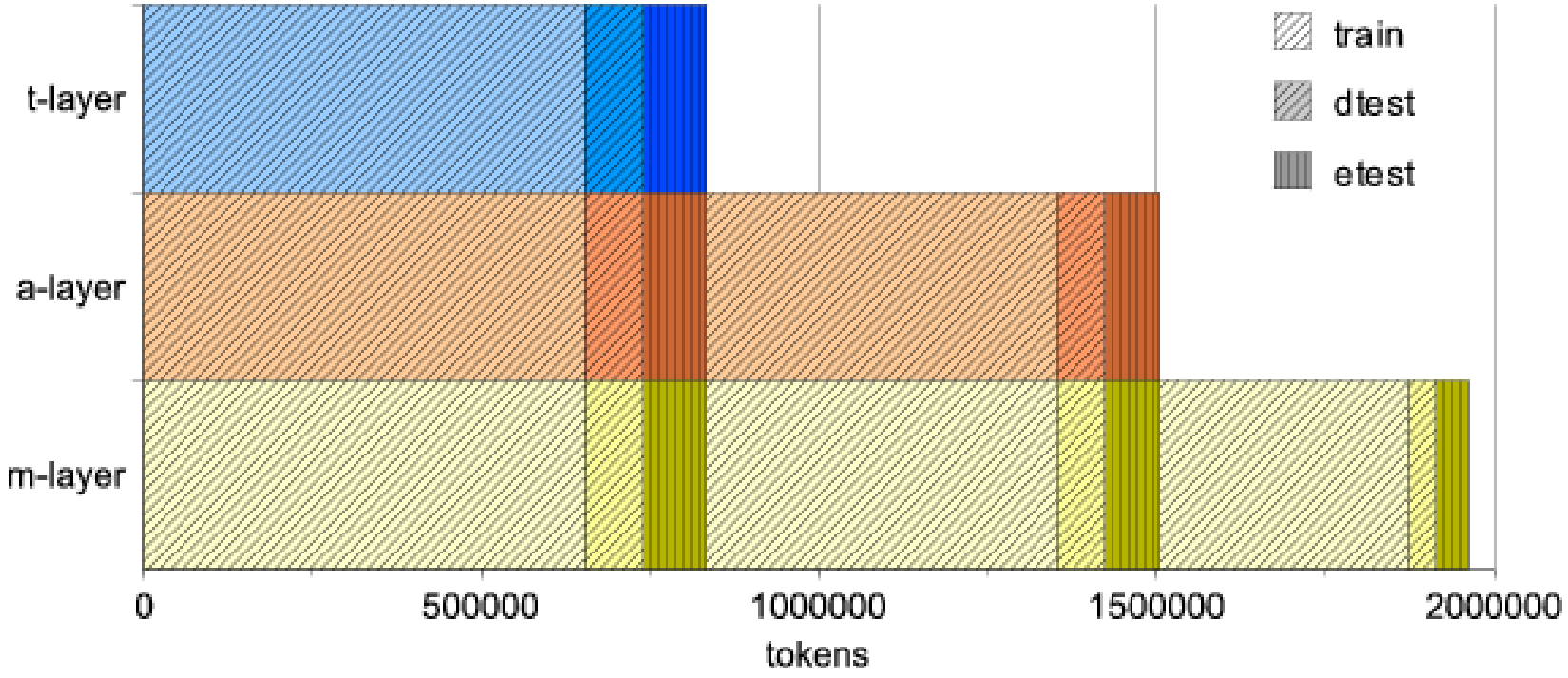

a-layer
AuxS
Můžete
[Can_you]
Pred
?
AuxK
vysvětlit
[explain]
Obj
to
[it]
Obj
na
[on]
AuxP
příkladu
[an_example]
Adv
t-layer
root
vysvětlit.inter
[to_explain]
PRED
ten
[it]
PAT
#PersPron
ACT
#Gen
ADDR
příklad
[example]
MEANS
m-layer
Můžete
[Can_you]
moci
VB-P---2P-AA---
to
[it]
ten
PDNS4----------
vysvětlit
[explain]
vysvětlit
Vf--------A----
na
[on]
na
RR--6----------
příkladu
[an_example]
příklad
NNIS6-----A---1
?
?
Z:----------
w-layer
Můžete
[Can_you]
to
[it]
vysvětlit
[explain]
napříkladu
[onan_example]
?

PDT: Division of the data to layers

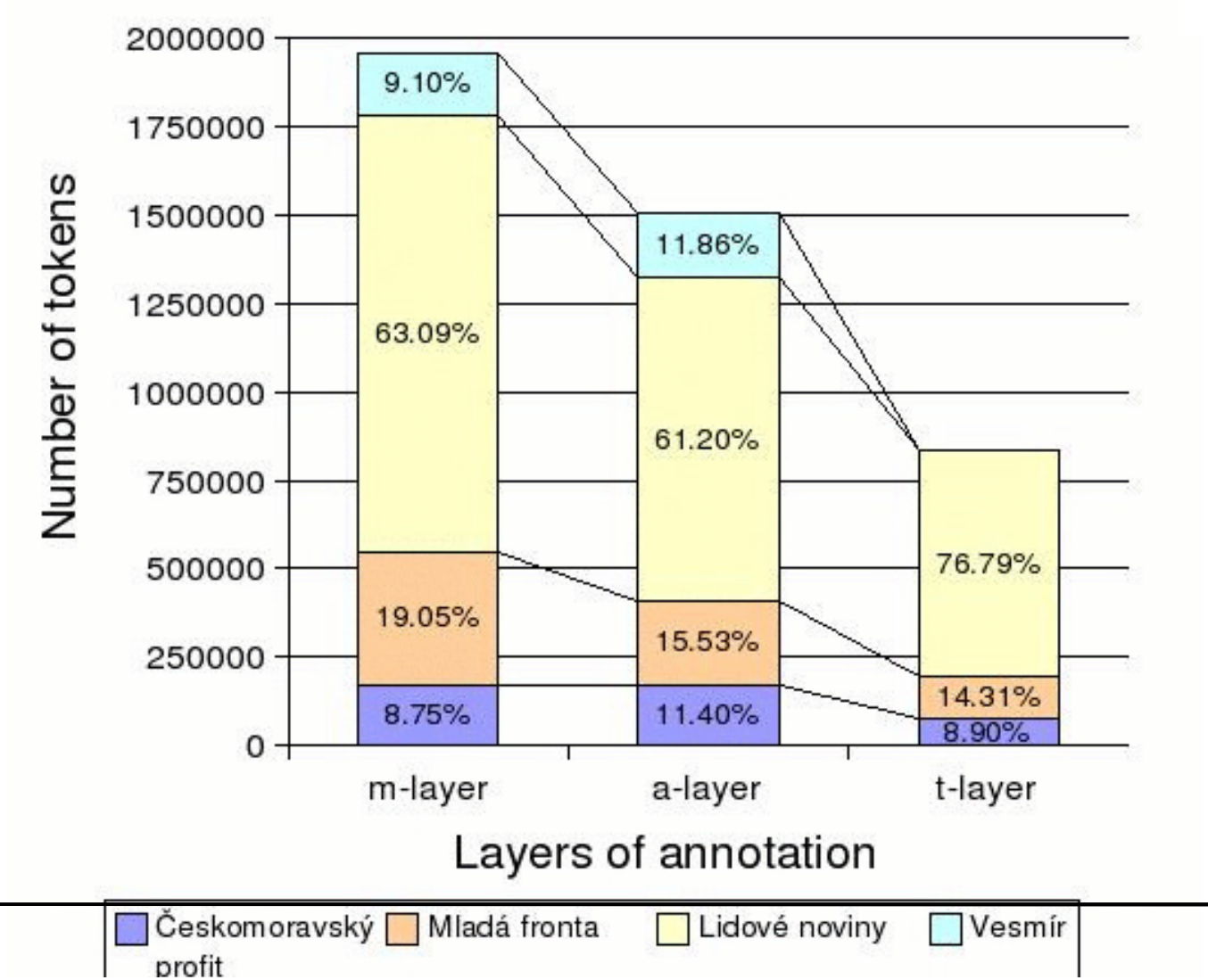


ln94206_1.m.gz + ln94206_1.w.gz	ln94206_1.a.gz	ln94206_1.t.gz
ln94206_2.m.gz + ln94206_2.w.gz	ln94206_2.a.gz	ln94206_2.t.gz
ln94206_3.m.gz + ln94206_3.w.gz	ln94206_3.a.gz	ln94206_3.t.gz

PDT: Division of the data into training and test sets



PDT: Number of tokens from the particular sources



PDTSC: spoken Czech

- 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 (manual) morphological annotation
  - analytical layer ... automatically
  - deep syntax … manually annotated

  } PDT-like

PCEDT: translated Czech

- Prague Czech-English Dependency Treebank 2.0
  - Penn Treebank data (English)
  - translated by professional 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

PCEDT: translated Czech

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

Czech part:

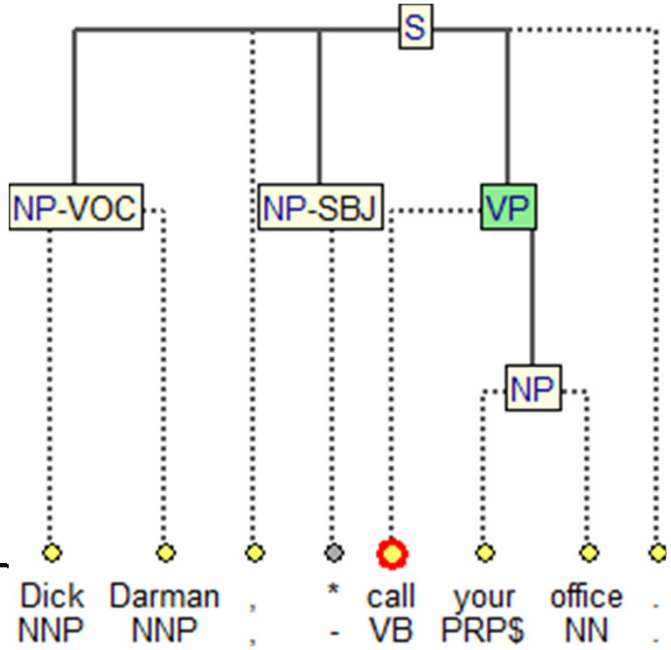
- w-layer
- m-layer
- a-layer
- t-layer … manual

English part:

- original Penn Treebank annotation p-layer

→ w-layer
m-layer
a-layer

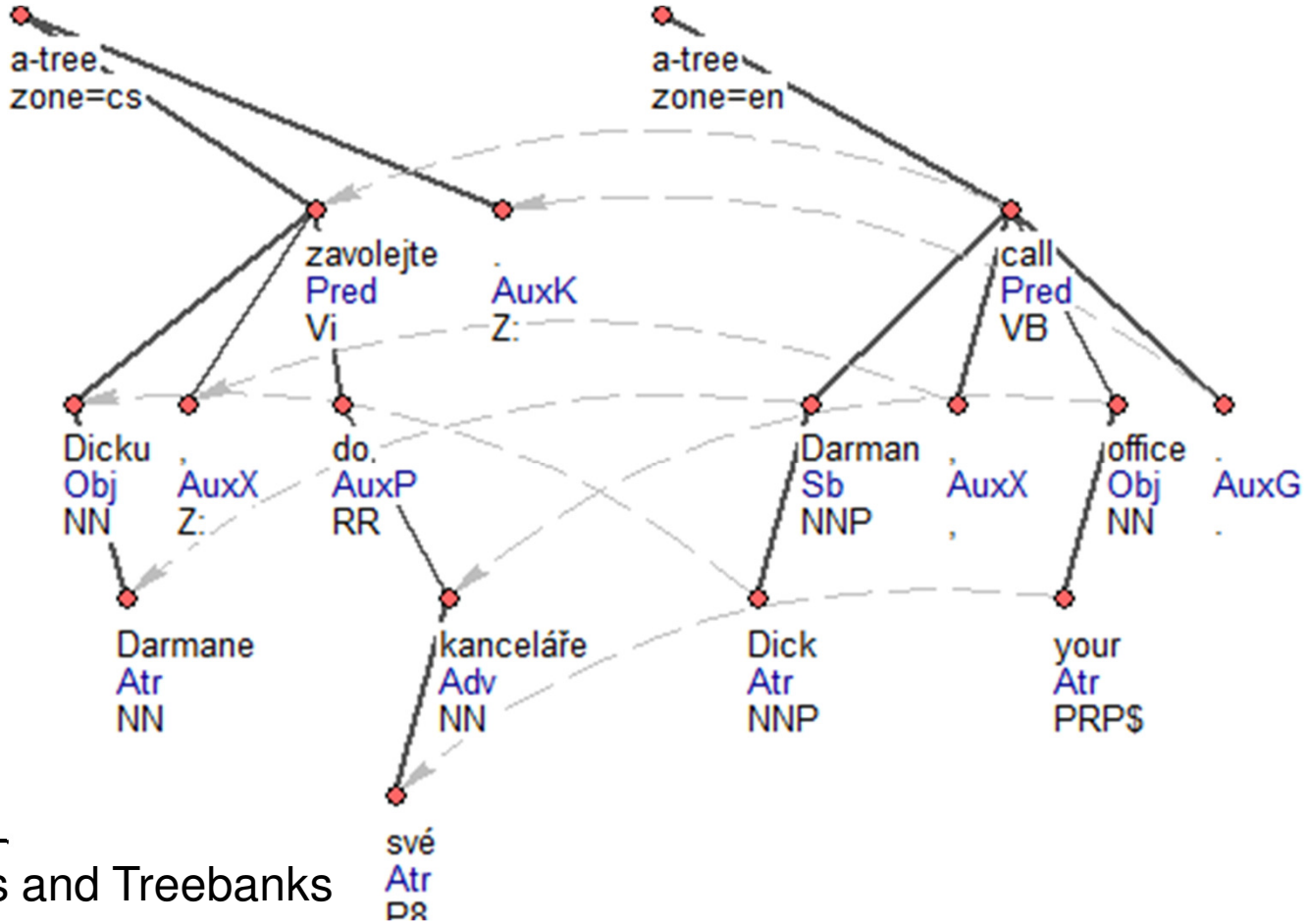
- t-layer … manual



Prague Dependency Family: English

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

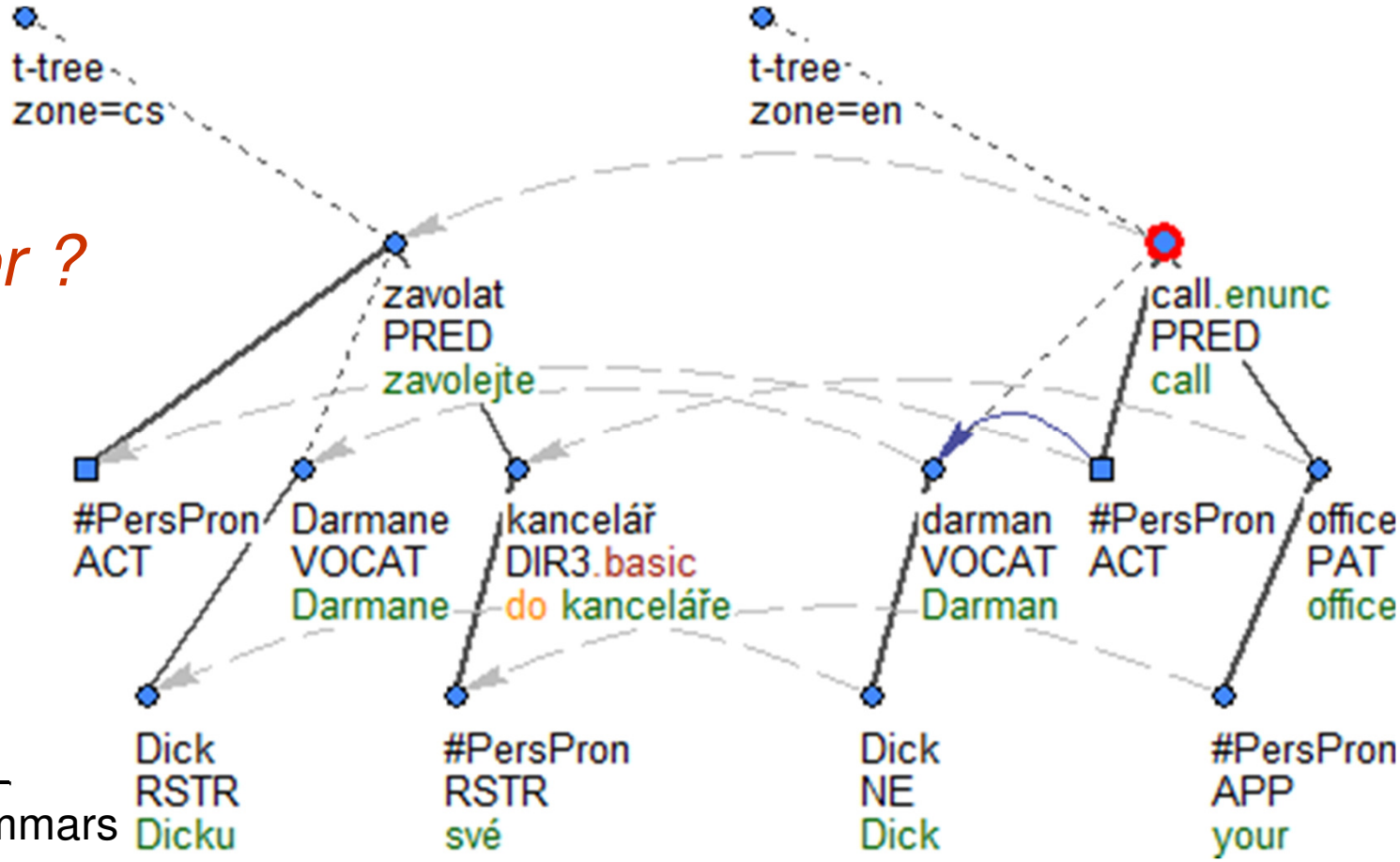
Which layer ?



Prague Dependency Family: English

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

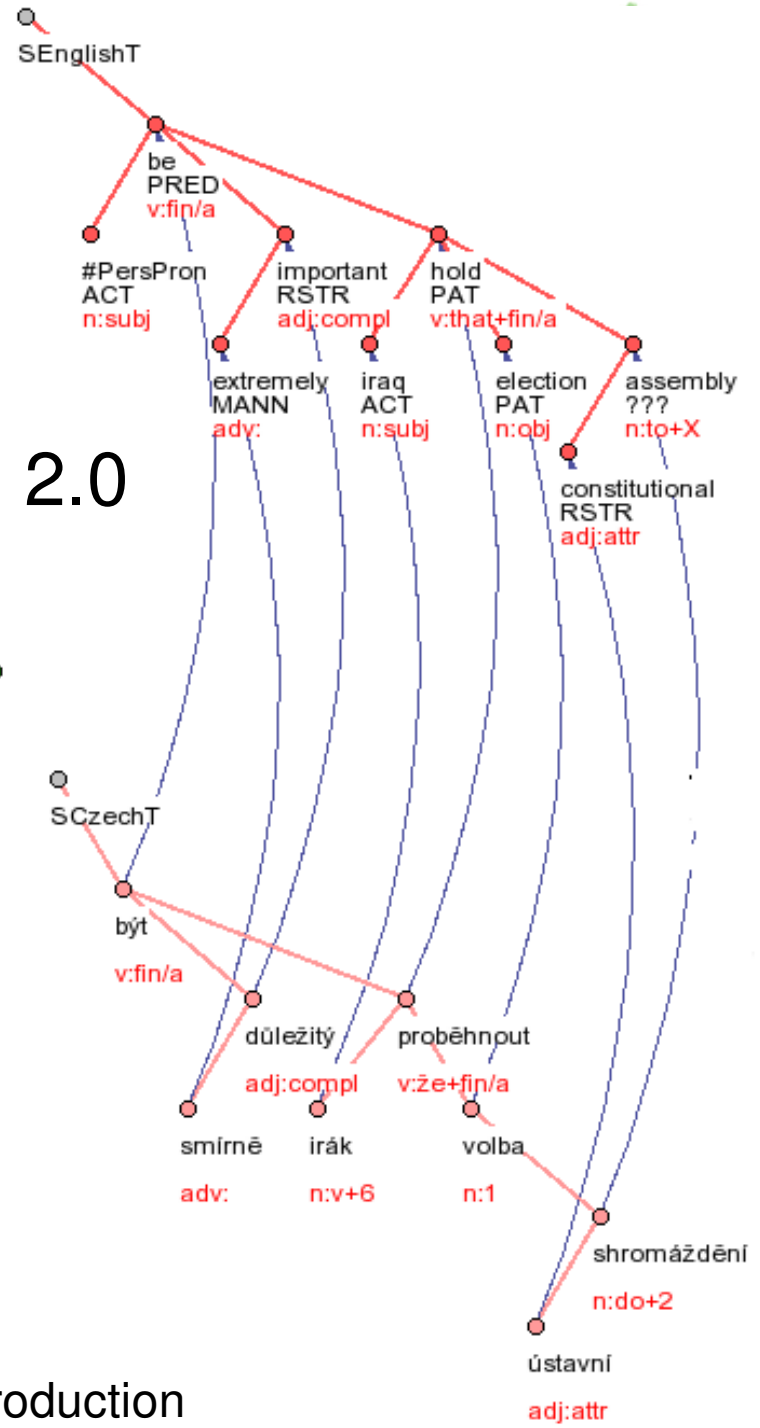
Which layer ?



Prague Dependency Family: English

- 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/



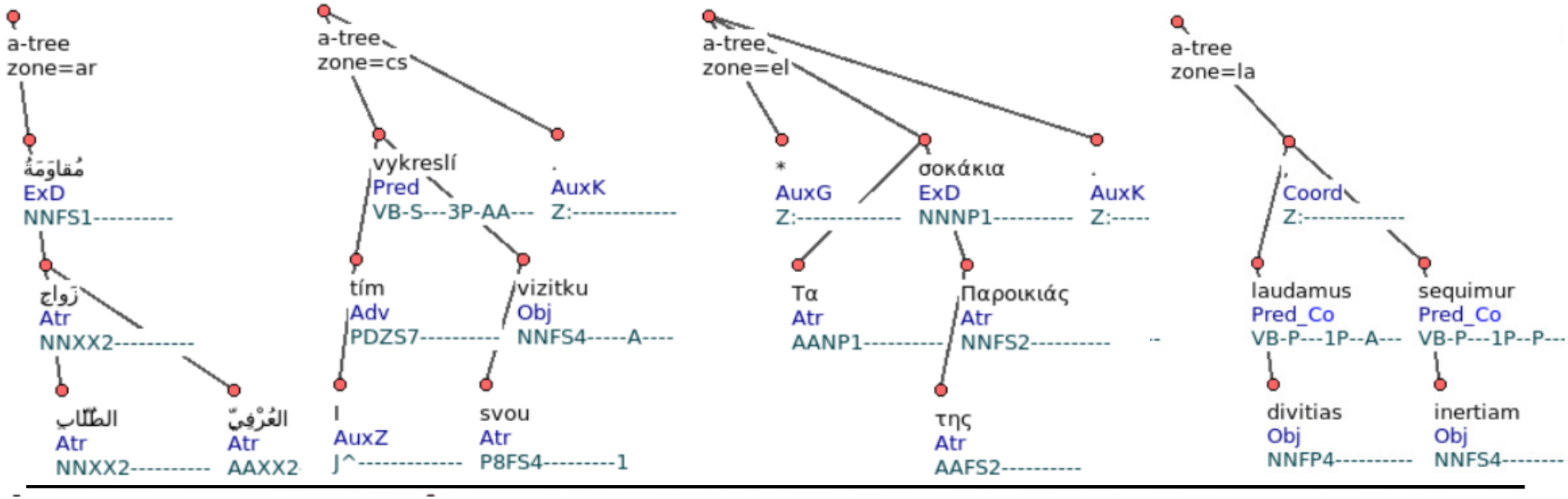
It is extremely important that Iraq held elections to a constitutional assembly.

HamleDT: HArmonized Multi-LanguagE Dependency Treebank

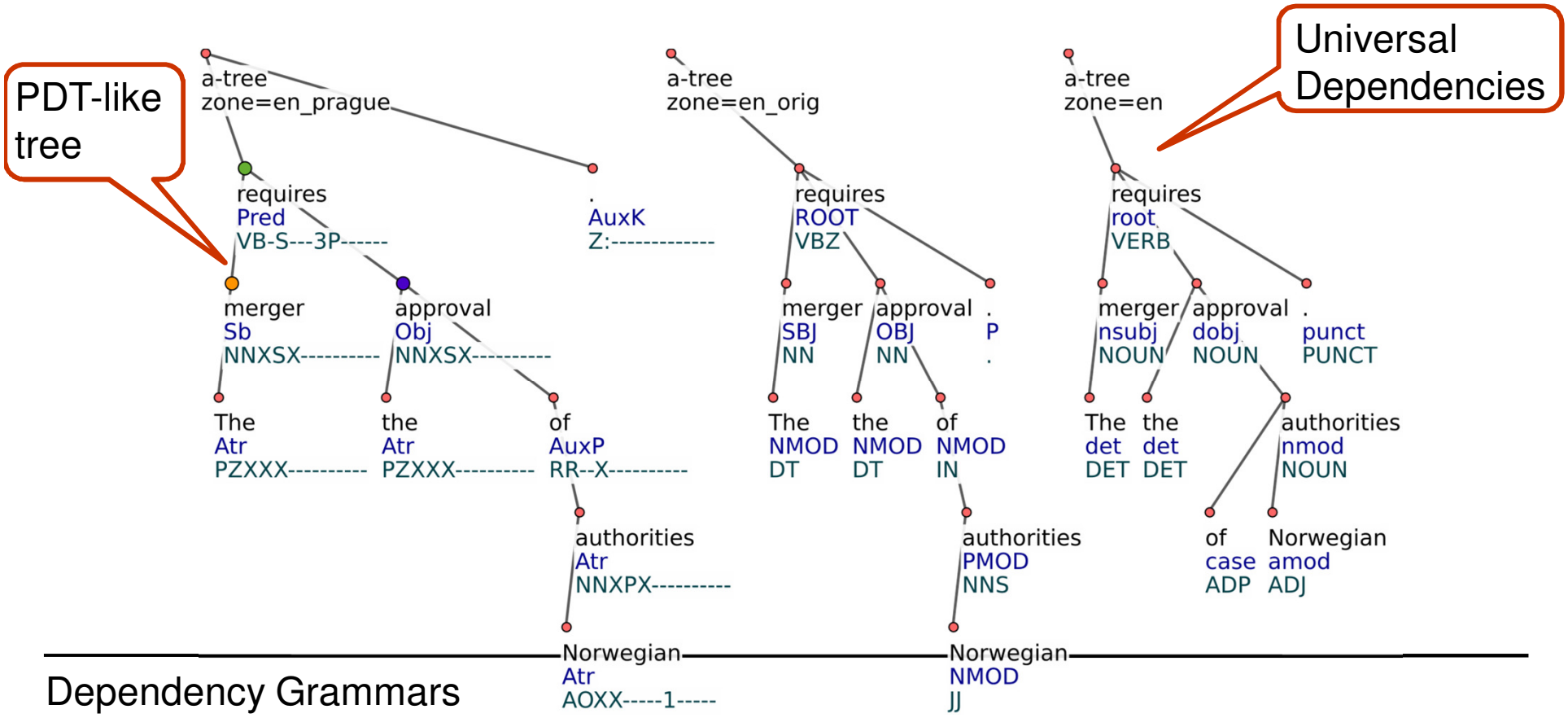
- HamleDT ~ a compilation of existing dependency treebanks (or dependency conversions of other treebanks), transformed so that they all conform to the same annotation style … 2012

  36 languages, 42 treebanks in HamleDT 3.0 (2015)

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



HamleDT: HArmonized Multi-LanguagE Dependency Treebank

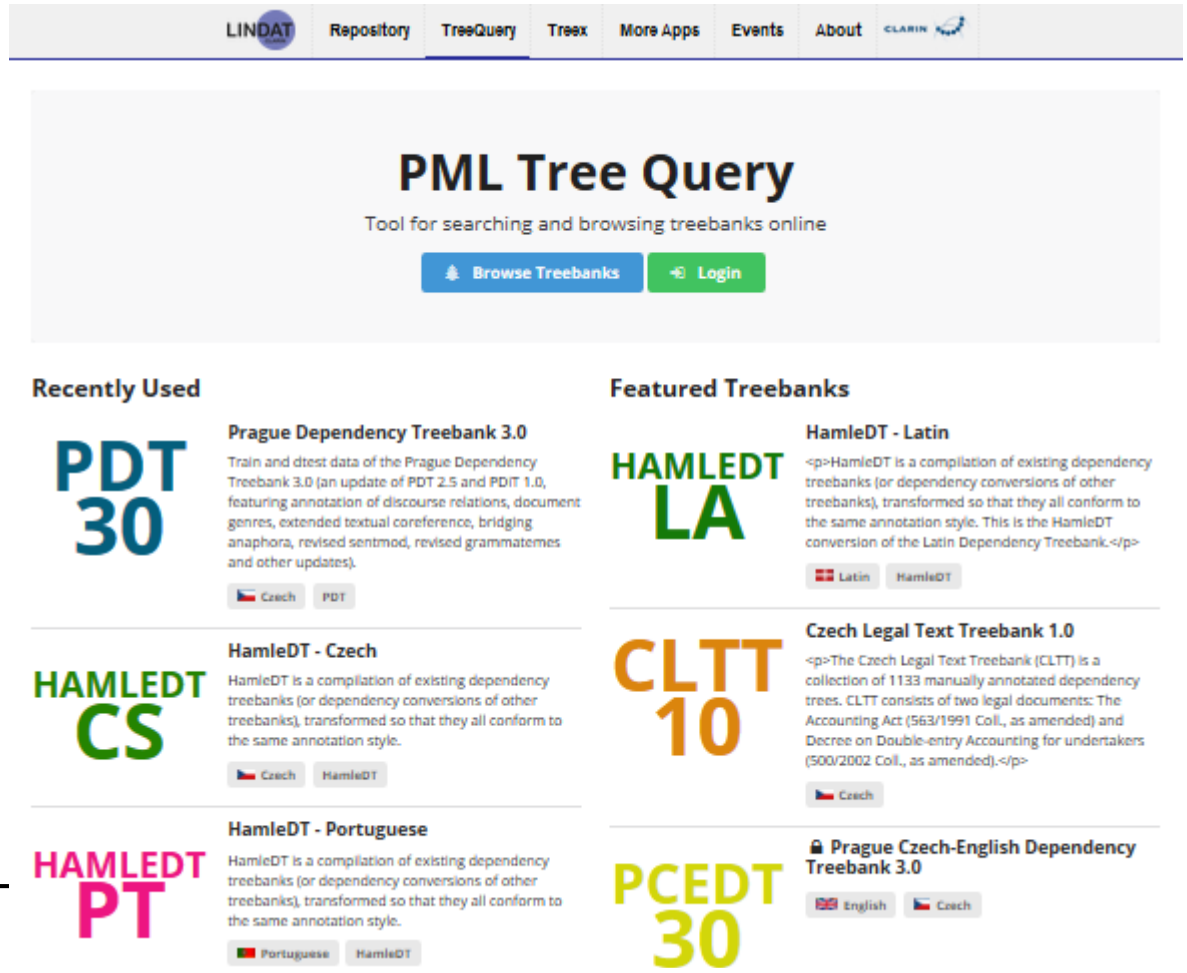


How to access / obtain dependency treebanks

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

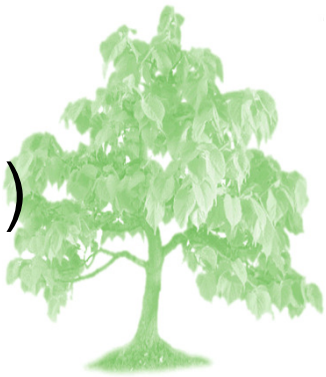
FGD	PDT
• tectogrammar/deep syntax	• t-layer (tectogrammatical l.)
• surface syntax	• a-layer (analytical l.)
• morphematics	• m-layer (morphological l.)
• morphonology	• w-layer (word layer)
• phonology	

structural layers

reasons

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

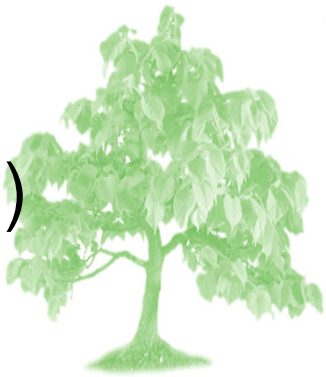
Differences between FGD and PDT (cont.)



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

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

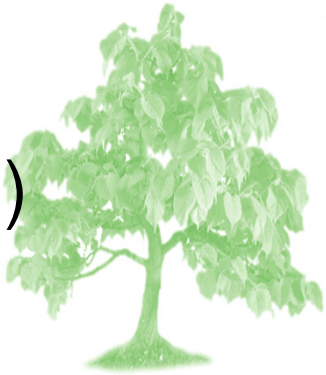
Differences between FGD and PDT (cont.)



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)

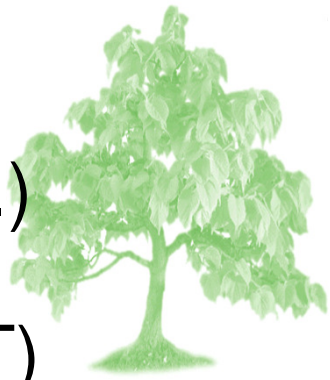
⇨ edges for non-dependency relations
(other than coordination/apposition)

- 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 2.0 guide http://ufal.mff.cuni.cz/pdt2.0/
- PDT-C documentation https://ufal.mff.cuni.cz/pdt-c
- Hajič, J., Hajičová, E., Mírovský, J., Panevová, J.: Linguistically Annotated Corpus as an Invaluable Resource for Advancements in Linguistic Research: A Case Study. *The Prague Bulletin of Mathematical Linguistics*, No. 106, ISSN 0032-6585, pp. 69-124, 2016
- Hajičová, E., Panevová, J., Sgall, P. (2002) *Úvod do teoretické a počítačové lingvistiky*, sv. I. Karolinum, Praha.