

The Prague Trees

the basic tools and a use case

Pavel Straňák

Språkbanken, Göteborg, 5 March 2015



PML: Prague Markup Language

- XML language
- Generic language for representation of tree structures and their linking
- PDT: 4 layers (3 layers of annotation) with links in between + valency lexicon (also a list of trees)
- No bias towards dependencies or constituencies
- PML Schema to define the structures (think DTD, XML Schema)

Multiword expressions in the Prague Dependency Treebank – Day 2

Pavel Straňák



PML: data structures

- atomic type (string): format
 - ID
 - PMLREF
 - integer, positiveInteger, date, time, duration, ...
 - any
- enumeration type
- structure (attr: value pairs)
- list (ordered, unordered)
- alternatives
- sequence
- (container)

- **atomic** – a (formatted) string
- **enumerated** type – given set of possible values
- **structure** – set of attribute-value pairs
- **list** – (un)ordered list of units of one type
- **alternative** – similar to unordered list, but with different semantics
- **sequence** – similar to ordered list, but allowing members with diverse types and supporting mixed content).

PML: roles and validation

- roles: TREES, NODE, CHILDNODES, ID, KNIT, ORDER, HIDE
- Cross-reference (e.g. coreference)
- Multi-layered
 - separated files
 - file-id#id
- Validation
 - PML Schema can be validated by a RNG Schema
 - PML Schema can be converted via XSLT to RNG Schema (validation of the data)

PML⁴

- **Non-trrebanking applications:**
 - Treeex: NLP processing framework; originally developed for tree-to-tree MT system.
 - Learner corpus CzeSL: multilayer error annotation with corrections
- **Treebanks in PML:**
 - <https://lindat.mff.cuni.cz/services/pmltq/>

PML Framework

- **Libraries:**

- **Perl:** Treex::PML package (CPAN)
- **Java** libraries built for feat editor

- **Tools:**

- **Tred** (edit trees, search trees, search PML-TQ databases remotely)
 - btred, ntred, jtred (various command line versions)
- **PML Tree Query:** <https://lindat.mff.cuni.cz/services/pmltq/>
- **MEd:** linear annotations, e.g. alignment of parallel data, audio transcription
- **Law:** morphological annotation tool
- **Feat:** layered annotation of learners corpora
- **Capek:** annotation tool tailored to school children and grammar

PML References

- Hana Jirka, Štěpánek Jan: Prague Markup Language Framework. In: Proceedings of the Sixth Linguistic Annotation Workshop, Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-937284-32-9, pp. 12-21, 2012
- Pajas Petr, Štěpánek Jan: A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague DependencyTreebank 2.0. Technical report no. 2005/TR-2005-29, Copyright © ÚFAL MFF UK, ISSN 1214-5521, 37 pp., 2005

Tree Editor (TrEd)

- Build a treebank from the ground up.
- Save your first syntactic trees within 10 minutes:
<http://youtu.be/ZaY2K786TyM>
- Search and visualise any trees
- backends and interfaces (btred, jtred, etc.)
- view modes and stylesheets

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

- Documentation (manual, basic library, macros)
- Perl + Tk
 - Ready-made packages for Windows, Linux (.deb and .rpm) and OS X (10.7 and higher)
- Extensions (formats, annotation environments)

PML Tree Query

- PML-TQ can represent and query any treebank (to the best of our knowledge).
- Currently we have about 30 treebanks converted and available for search:
 - phrase structures
 - dependencies
 - multi-layer
 - original design w. layers of linguistic description
 - combining information from different projects over the same data (WSJ)
- parallel

Treex (::Web)

- <http://ufal.mff.cuni.cz/treex>
- NLP processing framework. Perl APIs. Developed for machine translation (tree-to-tree system TectoMT), but now
- it works as a general processing workflow that includes many standard tools
- (tokenizers, morphological analysers and taggers, parsers, NERs) trained models, etc.
- Web version: Pick a scenario, run it on your data, view and download results.

MWE annotation in PDT ≥ 2.5

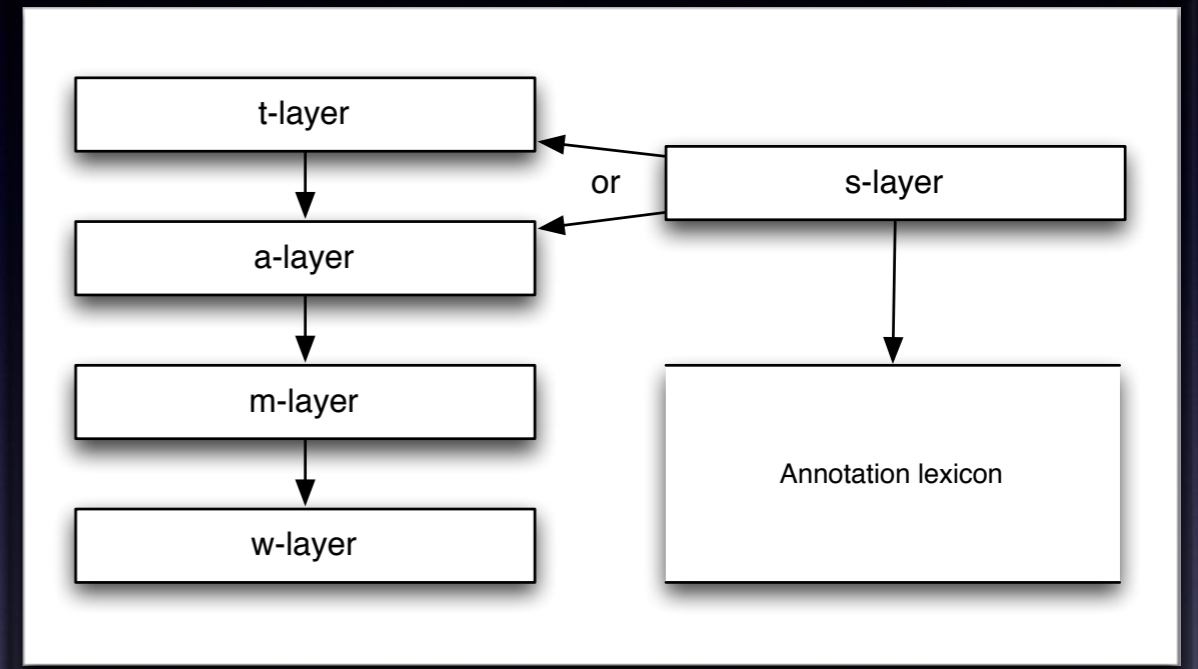
- New annotation of “any MWEs”, looking at plain text
- additional annotation over t-layer (s-files), merged back
- Storing the lexicon: SemLex
 - Pre-annotating known MWEs using trees
- NEs also annotated

Tectogram. structure of a MWE

- **hypothesis**: dependency structure + deep word order: each MWE should only have 1 t-structure, that should always be contiguous
- means of effective automatic identification (given t-trees)
- 771 SemLex entries have more than one t-structure in data:
- systematic deficiencies of PDT 2.0 t-lemmata
- occasional errors in t-layer or our annotations

PDT 2.0 + s-data

- PDT 2.0 data; format: PML
- addition of s-layer (“sense”)
 - not a deeper layer, refers to a-nodes or t-nodes
 - a list of pairs (lexicon.ref, [t/a]-node.ref.list)



Prezident má za týden jmenovat **ústavní soudce**
 Parlament doporučil třináct kandidátů
 Praha (li, ben) -

Prezident Havel by měl **15. července na Pražském hradě** jmenovat třináct soudců **Ústavního soudu**. Řekl nám to včera prezidentův mluvčí krátce poté, co parlament vyslovil souhlas se třinácti ze čtrnácti kandidátů na soudce, které Havel navrhl. **Ústavní soud**, který bude **soudním orgánem** ochrany ústavnosti, má zahájit činnost v Brně zřejmě v září. Podle ústavy se skládá z 15 soudců, jmenovaných na deset let. Soud se může ujmout činnosti složením slibu dvanáctého soudce. Funkce **ústavního soudce** je neslučitelná s členstvím **v politických stranách**. Základní plat soudců bude **25 tisíc** korun.

Soudci by se mělo stát pět bývalých členů **Ústavního soudu ČSFR** **Zdeněk Kessler** (dříve poslanec FS za ODS), **Vlastimil Ševčík** (předtím poslanec FS za OH), **Antonín Procházka** (dříve poslanec ČNR za KDS), **Vojen Güttler** a **Pavel Mates** a čtyři vysokoškolští učitelé práva - **Vladimír Klokočka**, **Vojtěch Cepi**, **Vladimír Čermák** a Slovák **Pavol Holländer**. Tři další kandidáti, **Iva Brožová**, **Miloš Holeček** a **Vladimír Jurka**, jsou soudci **krajského a okresního soudu**. Třináctým kandidátem je komerční právník **Vladimír Paul**. Parlament neschválil kandidaturu docentky **Ireny Pelikánové**, členky ODA, pro niž hlasovala jen část poslanců **vládních stran**.

Značky
 Obecné **Pojmenované entity** **Automatická anotace**

SemLex **A=a**

ID: 0000028437 Source: CWN2a POS: **Základní tvar: soudní orgán** **Lematizovaný tvar: soudní orgán**

Příklad: **Synonyma:**

Typical texts with MWEs

colours specify multiword lexemes vs. types of named entities

Festival **Starý zákon v umění** zahájí v září Job **Petra Ebena**

Varhanní kompozice Petra Ebena Job zahájí 3. září v **Lichtenštejnském paláci** v Praze mezinárodní festival nazvaný **Starý zákon v umění**. Rozsáhlá akce soustředí výstavy, **divadelní a baletní představení** i koncerty, které se uskuteční především v pražském Rudolfinu, **u sv. Jakuba a v kostele sv. šimona a Judy**. **Kromě České republiky** se zatím k účasti přihlásily Norsko, Německo, Itálie, Rakousko, švédsko, Izrael a Rusko.

Z hudebních kolektivů se v Praze představí například **Norští sólisté, Jeruzalémský symfonický orchestr a Komorní orchestr z Halle**. Pravděpodobně vystoupí i **Česká filharmonie, Symfonický orchestr Českého rozhlasu, brněnská Státní filharmonie, Talichův komorní orchestr a soubory Archi Boemi a Virtuosi di Praga**. Do programu festivalu přispěje **Státní opera Praha, Divadlo na Vinohradech a balet Národního divadla z Brna**. V rámci festivalu se bude konat reprezentativní izraelská archeologická výstava nazvaná **City of David (Město Davidovo)**, výstavy **výtvarného umění** i expozice studentských prací pražské **Akademie výtvarných umění**. Další akce připraví pražské **židovské muzeum, Památník národního písemnictví, Národní galerie a Národní muzeum**.

Heslo zobrazeno.

Značky

Obecné

Ukázat Odstranit

Pojmenované entity

Jméno Instituce Místo Objekt Adresa Čas Biblio Foreign X

Automatická anotace

SemLex

Označkovat Nové heslo Ulož heslo

A=a Hledat P N

ID: 0000003772 Source: Eurovoc POS: N Základní tvar: výtvarná umění Lematizovaný tvar: výtvarný umění

Typical texts with MWEs

colours specify multiword lexemes vs. types of named entities

Praha (li, ben) -

Prezident Havel by měl 15. července na Pražském hradě jmenovat třináct soudců Ústavního soudu. Řekl nám to včera prezidentův mluvčí krátce poté, co parlament vyslovil souhlas se třinácti ze čtrnácti kandidátů na soudce, které Havel navrhl. Ústavní soud, který bude soudním orgánem ochrany ústavnosti, má zahájit činnost v Brně zřejmě v září. Podle ústavy se skládá z 15 soudců, jmenovaných na deset let. Soud se může ujmout činnosti složením slibu dvanáctého soudce. Funkce ústavního soudce je neslučitelná s členstvím v politických stranách. Základní plat soudců bude 25 tisíc korun.

Soudci by se mělo stát pět bývalých členů Ústavního soudu ČSFR Zdeněk Kessler (dříve poslanec FS za ODS), Vlastimil Ševčík (předtím poslanec FS za OH), Antonín Procházka (dříve poslanec ČNR za KDS), Vojen Güttler a Pavel Mates a čtyři vysokoškolské učitelé práva - Vladimír Klokočka, Vojtěch Cepl, Vladimír Čermák a Slovák Pavol Holländer. Tři další kandidáti, Iva Brožová, Miloš Holeček a Vladimír Jurka, jsou soudci krajského a okresního soudu. Třináctým kandidátem je komerční právník Vladimír Paul. Parlament neschválil kandidaturu docentky Ireny Pelikánové, členky ODA, pro niž hlasovala jen část poslanců vládních stran.

(Prezident|t-mf930709-001-p1s1w1) (má|t-mf930709-001-p1s1w5) (za|t-mf930709-001-p1s1w4) (týden|t-mf930709-001-p1s1w4) (jmenovat|t-mf930709-001-p1s1w5) (ústavní|t-mf930709-001-p1s1w6) (soudce|t-mf930709-001-p1s1w7)

(Parlament|t-mf930709-001-p2s1w1) (doporučil|t-mf930709-001-p2s1w2) (třináct|t-mf930709-001-p2s1w3) (kandidátů|t-mf930709-001-p2s1w4)

(Praha|t-mf930709-001-p3s1Aw1) ((li|t-mf930709-001-p3s1Aw3)(,|t-mf930709-001-p3s1Aw4) (ben|t-mf930709-001-p3s1Aw5)) -

(Prezident|t-mf930709-001-p3s1Bw1) (Havel|t-mf930709-001-p3s1Bw2) (by|t-mf930709-001-p3s1Bw11) (měl|t-mf930709-001-p3s1Bw11) (15|t-mf930709-001-p3s1Bw5). (července|t-mf930709-001-p3s1Bw7) (na|t-mf930709-001-p3s1Bw10) (Pražském|t-mf930709-001-p3s1Bw9) (hradě|t-mf930709-001-p3s1Bw10) (jmenovat|t-mf930709-001-p3s1Bw11) (třináct|t-mf930709-001-p3s1Bw12) (soudců|t-mf930709-001-p3s1Bw13)

Parlament doporučil třináct kandidátů

Praha (li, ben) -

Prezident Havel by měl 15. července na Pražském hradě jmenovat třináct soudců Ústavního soudu. Řekl nám to včera prezidentův mluvčí krátce poté, co parlament vyslovil souhlas se třinácti ze čtrnácti kandidátů na soudce, které Havel navrhl. Ústavní soud, který bude soudním orgánem ochrany ústavnosti, má zahájit činnost v Brně zřejmě v září. Podle ústavy se skládá z 15 soudců, jmenovaných na deset let. Soud se může ujmout činnosti složením slibu dvanáctého soudce. Funkce ústavního soudce je neslučitelná s členstvím v politických stranách. Základní plat soudců bude 25 tisíc korun.

Soudci by se mělo stát pět bývalých členů Ústavního soudu ČSFR Zdeněk Kessler (dříve poslanec FS za ODS), Vlastimil Ševčík (předtím poslanec FS za OH), Antonín Procházka (dříve poslanec ČNR za KDS), Vojen Güttler a Pavel Mates a čtyři vysokoškolské učitelé práva - Vladimír Klokočka, Vojtěch Cepel, Vladimír Cermák a Slovák Pavol Holländer. Tři další kandidáti, Iva Brožová, Miloš Holeček a Vladimír Jurka, jsou soudci krajského a okresního soudu. Třináctým kandidátem je komerční právník Vladimír Paul. Parlament neschválil kandidaturu docentky Ireny Pelikánové, členky ODA, pro niž hlasovala jen část poslanců vládních stran.

(na|t-mf930709-001-p3s1Bw10) (Pražském|t-mf930709-001-p3s1Bw9) (hradě|t-mf930709-001-p3s1Bw10)
(jmenovat|t-mf930709-001-p3s1Bw11) (třináct|t-mf930709-001-p3s1Bw12) (soudců|t-mf930709-001-p3s1Bw13)
(Ústavního|t-mf930709-001-p3s1Bw14) (soudu|t-mf930709-001-p3s1Bw15). (Řekl|t-mf930709-001-p3s2w1)
(nám|t-mf930709-001-p3s2w2) (to|t-mf930709-001-p3s2w3) (včera|t-mf930709-001-p3s2w4)
(prezidentův|t-mf930709-001-p3s2w5) (mluvčí|t-mf930709-001-p3s2w6) (krátce|t-mf930709-001-p3s2w7)
(poté|t-mf930709-001-p3s2w12), (co|t-mf930709-001-p3s2w12) (parlament|t-mf930709-001-p3s2w11)
(vyslovil|t-mf930709-001-p3s2w12) (souhlas|t-mf930709-001-p3s2w13) (se|t-mf930709-001-p3s2w15)
(třinácti|t-mf930709-001-p3s2w15) (ze|t-mf930709-001-p3s2w18) (čtrnácti|t-mf930709-001-p3s2w17)
(kandidátů|t-mf930709-001-p3s2w18) (na|t-mf930709-001-p3s2w20) (soudce|t-mf930709-001-p3s2w20),
(které|t-mf930709-001-p3s2w22) (Havel|t-mf930709-001-p3s2w23) (navrhl|t-mf930709-001-p3s2w24).
(Ústavní|t-mf930709-001-p3s3w1) (soud|t-mf930709-001-p3s3w2), (který|t-mf930709-001-p3s3w4)
(bude|t-mf930709-001-p3s3w5) (soudním|t-mf930709-001-p3s3w6) (orgánem|t-mf930709-001-p3s3w7)
(ochrany|t-mf930709-001-p3s3w8) (ústavnosti|t-mf930709-001-p3s3w9), (má|t-mf930709-001-p3s3w12)
(zahájit|t-mf930709-001-p3s3w12) (činnost|t-mf930709-001-p3s3w13) (v|t-mf930709-001-p3s3w15)
(Brně|t-mf930709-001-p3s3w15) (zřejmě|t-mf930709-001-p3s3w16) (v|t-mf930709-001-p3s3w18)
(září|t-mf930709-001-p3s3w18). (Podle|t-mf930709-001-p3s4w2) (ústavy|t-mf930709-001-p3s4w2) (se|t-mf930709-001-p3s4w4)
(skládá|t-mf930709-001-p3s4w4) (z|t-mf930709-001-p3s4w7) (15|t-mf930709-001-p3s4w6) (soudců|t-mf930709-001-p3s4w7)

Soubor Úpravy Debug

Prezident má za týden jmenovat **ústavní soudce**

Parlament doporučil třináct kandidátů

Praha (li, ben) -

Prezident Havel by měl **15. července na Pražském hradě** jmenovat třináct soudců **Ústavního soudu**. Řekl nám to včera prezidentův mluvčí krátce poté, co parlament vyslovil souhlas se třinácti ze čtrnácti kandidátů na soudce, které Havel navrhl. **Ústavní soud**, který bude **soudním orgánem** ochrany ústavnosti, má zahájit činnost v Brně zřejmě v září. Podle ústavy se skládá z 15 soudců, jmenovaných na deset let. Soud se může ujmout činnosti složením slibu dvanáctého soudce. Funkce **ústavního soudce** je neslučitelná s členstvím **v politických stranách**. Základní plat soudců bude **25 tisíc** korun.

Soudci by se mělo stát pět bývalých členů **Ústavního soudu ČSFR** **Zdeněk Kessler** (dříve poslanec FS za ODS), **Vlastimil Ševčík** (předtím poslanec FS za OH), **Antonín Procházka** (dříve poslanec ČNR za KDS), **Vojen Güttler** a **Pavel Mates** a čtyři vysokoškolští učitelé práva - **Vladimír Klokočka**, **Vojtěch Cepl**, **Vladimír Čermák** a Slovák **Pavol Holländer**. Tři další kandidáti, **Iva Brožová**, **Miloš Holeček** a **Vladimír Jurka**, jsou soudci **krajského a okresního soudu**. Třináctým kandidátem je komerční právník **Vladimír Paul**. Parlament neschválil kandidaturu docentky **Ireny Pelikánové**, členky ODA, pro niž hlasovala jen část poslanců **vládních stran**.

text generated from t-layer
+ MWEs (t-nodes)

*

Značky

Obecné

Ukázat

Odstranit

Pojmenované entity

Jméno

Instituce

Místo

Objekt

Adresa

Čas

Bibli

Foreign

X

 Automatická anotace

markup (show, remove, NEs)

SemLex

Označkovat

Nové heslo

Ulož heslo

SemLex Editor

 A=a

Hledat

P

N

ID: 0000028437 Source: CWN2a POS: N

Základní tvar: soudní orgán

Lematizovaný tvar: soudní orgán

Příklad:

Synonyma:

Glosa:



Změněno:

090607125630

merger

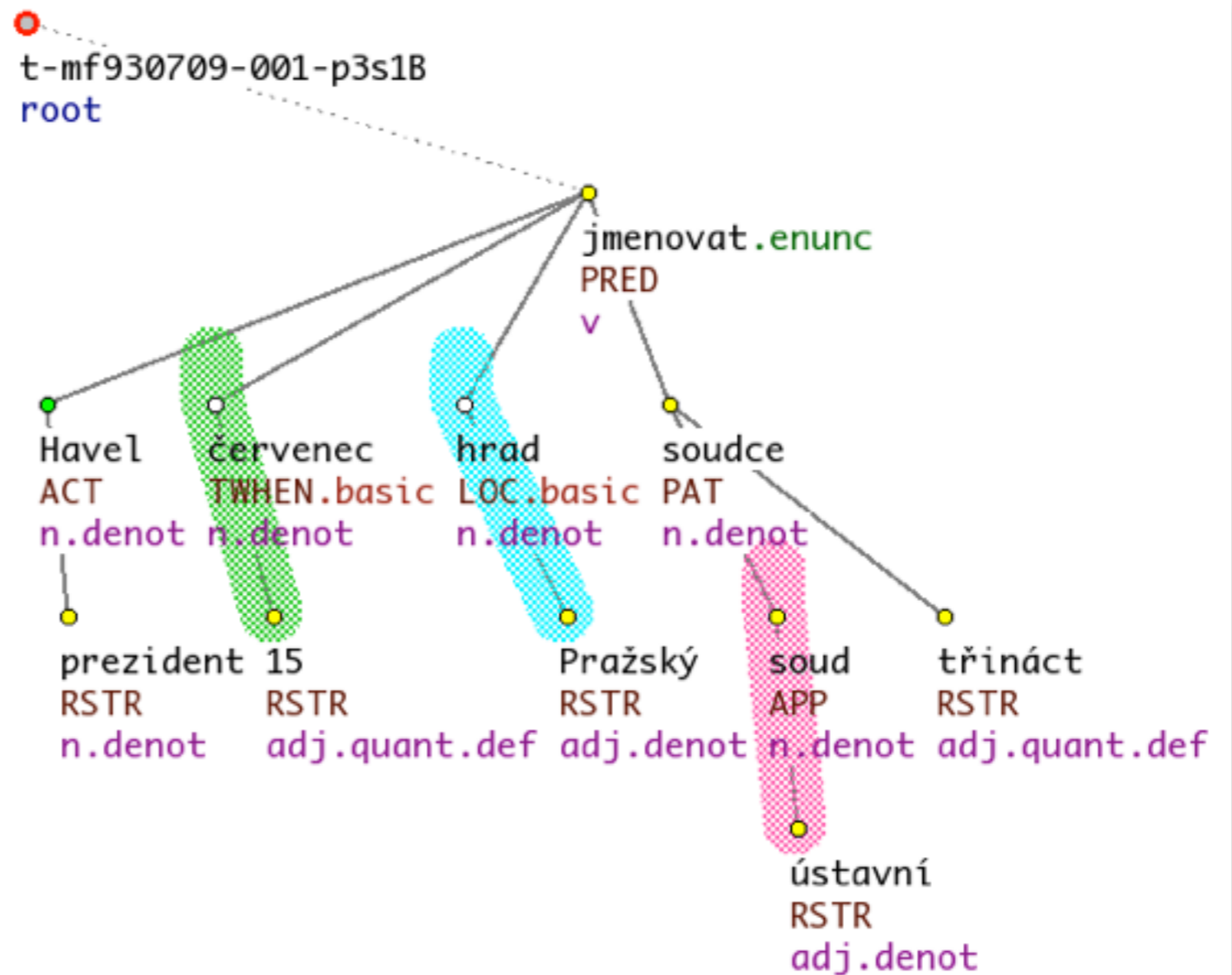
Annotation method

1. Pre-annotate: rule-based (limited MWEs, external)
2. `on_load(t_file)` pre-annotate MWEs from SemLex, that already have a t-tree
3. Identify NEs and MWEs
 1. from Semlex
 2. new – add lexemes (or frequent NEs) to Semlex
4. each MWE from 3 is again pre-annotated in the file via its t-tree

Node Attributes

Hide empty values

atree.rf	a#a-mf930709-001-p3s1B
deepord	0
id	t-mf930709-001-p3s1B
mwes	Sequence
annotator	Container
name	stastna
#content	Sequence
st	Structure
id	s-mf930709-001-16
lexicon-id	s##time
tnode.rfs	Unordered list
	t-mf930709-001-p3s1Bw5
	t-mf930709-001-p3s1Bw7
st	Structure
id	s-mf930709-001-160
lexicon-id	s##location
tnode.rfs	Unordered list
	t-mf930709-001-p3s1Bw9
	t-mf930709-001-p3s1Bw10
st	Structure
id	s-mf930709-001-161
lexicon-id	s##institution
tnode.rfs	Unordered list
	t-mf930709-001-p3s1Bw14
	t-mf930709-001-p3s1Bw15
annotator	Container
name	vimmrova
#content	Sequence
st	Structure
id	s-mf930709-001-16
lexicon-id	s##time
tnode.rfs	Unordered list
	t-mf930709-001-p3s1Bw5



MWES in a tree, attributes listed
colours specify multiword lexemes vs. types of named
entities

Statistics¹

- Whole t-layer of PDT: 675 000 t-nodes
- 16,3% of content words take part in MWEs
- 63% of the data annotated in parallel
- 8,816 MWEs (types), 5,352 of those identified by annotators, added to SemLex

parallel annot.	1	2	3	PDT	2+3/PDT	* /PDT
t-files	1,288	1,412	465	3,165	59%	100%
t-nodes	248,448	343,834	82,683	674,965	63%	100%

Statistics²

```
# List the MWE types
```

```
t-root $r := [ ];
```

```
>> for $r.mwes/type give $1  
desc
```

lexeme	20078
person	6927
institution	4940
number	3629
object	2883
time	2644
location	1876
address	136
foreign	132
biblio	35

Semlex

- All MWEs in PDT (t-layer)
- basic (quotation) forms
- lemmatized forms
- dependency structures
 - t-trees, i.e. no prepositions, for instance

```
SemLex.yml (~/.lxmn/data) - VIM
- obchodní
- 1
perl/hash:SemLex_heslo
IC_FORM: deficit státního rozpočtu
DATED: '110914162730'
EMPLE: ~
ISS: ~
'0000032344'
IMATIZED: deficit státní rozpočet
IFIED: ~
IFIER: bejcek
PHO_TAGS: ''
GID: ~
25_FREQ: 2
: 'N'
ORCE: stastna
ONYMS: []

E_STRUCT:
- deficit
- ~

- rozpočet
- 0

- státní
- 1
perl/hash:SemLex_heslo
IC_FORM: přitažený za vlasy
DATED: '110914170030'

203970,29
```

Tectogrammatical structure of a MWE

- **Our hypothesis:** dependency structure + deep word order: each MWE should only have 1 t-structure, that should always be contiguous
 - means of effective automatic identification (given t-trees)
- 771 SemLex entries have more than one t-structure in data:
 - systematic deficiencies of PDT 2.0 t-lemmata
 - occasional errors in t-layer or our annotations

LINDAT/CLARIN services

- Store all corpora and treebanks safely. Make them available and directly citable
- Make all corpora available in Kontext (corpus query) and PML-TQ (tree query)
- Make automatic tagging and parsing easily available for enhancement of the corpora
 - Morphodita, NameTag, Treex Web
- Again, store and make available the enhanced corpora