PML: Prague Markup Language

- XML language
- Generic language for representation of tree structures and their linking
- PDT: 4 layers (3 layers of anotation) 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)

PARSEME

ÚFAL

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)

PARSEME

ÚFAL

- **atomic** – a (formated) 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).

PARSEME

ÚFAL

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)

PARSEME

ÚFAL

PML[4]

- **Non-trrebanking applications:**
  - Treex: 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/

PARSEME

ÚFAL

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

PARSEME

ÚFAL

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

PARSEME

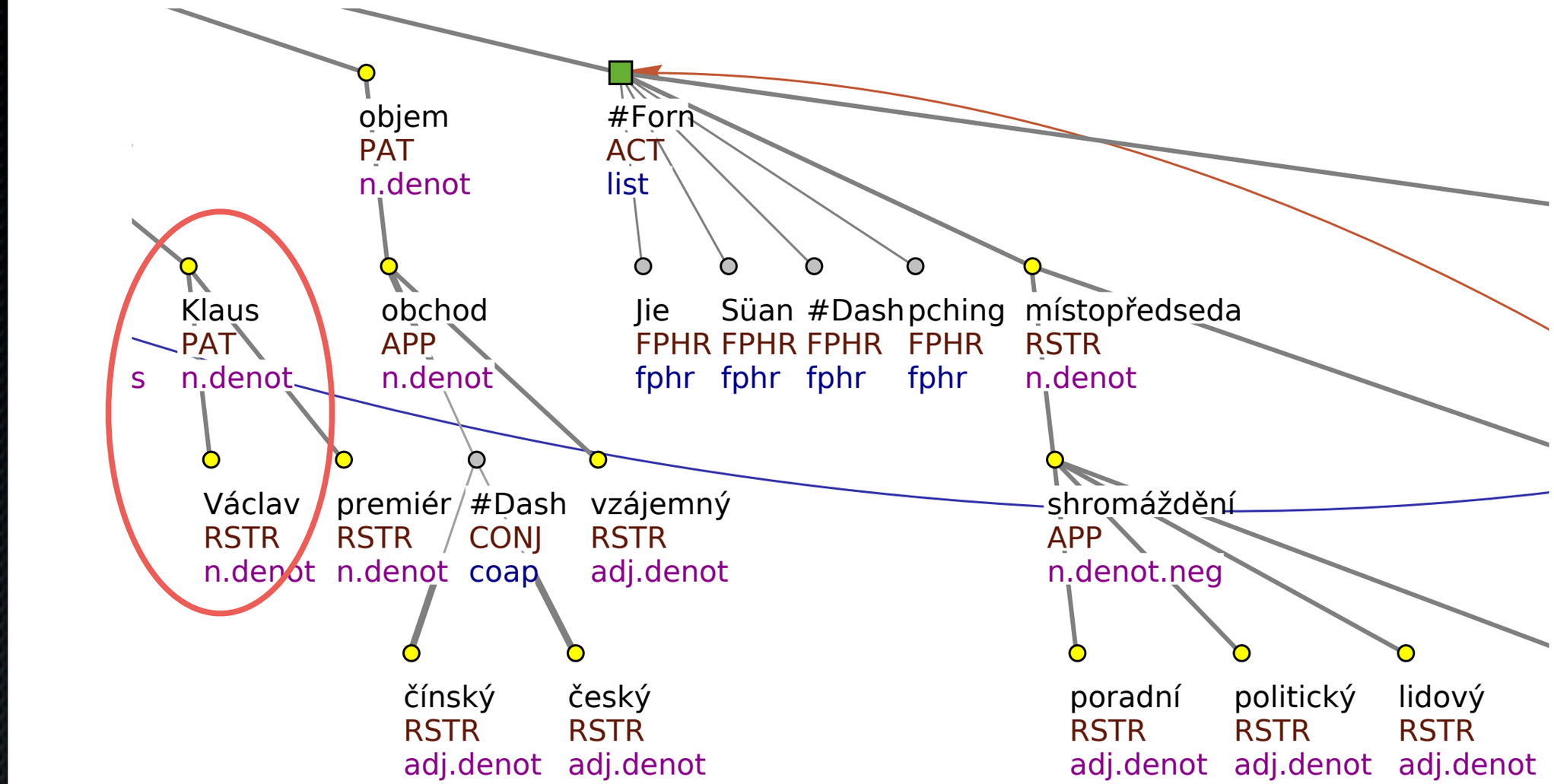
ÚFAL

More PDT 2.0 annotation – list structures

- **List structures:** nodetype= list, a/lex.rf is empty
- **Foreign Phrases**
  - generated node with t_lemma #Forn + flat list of members with a functor FPHR
- **Identifying expressions**
  - generated node with t_lemma #Idph + flat list of members with a functor ID

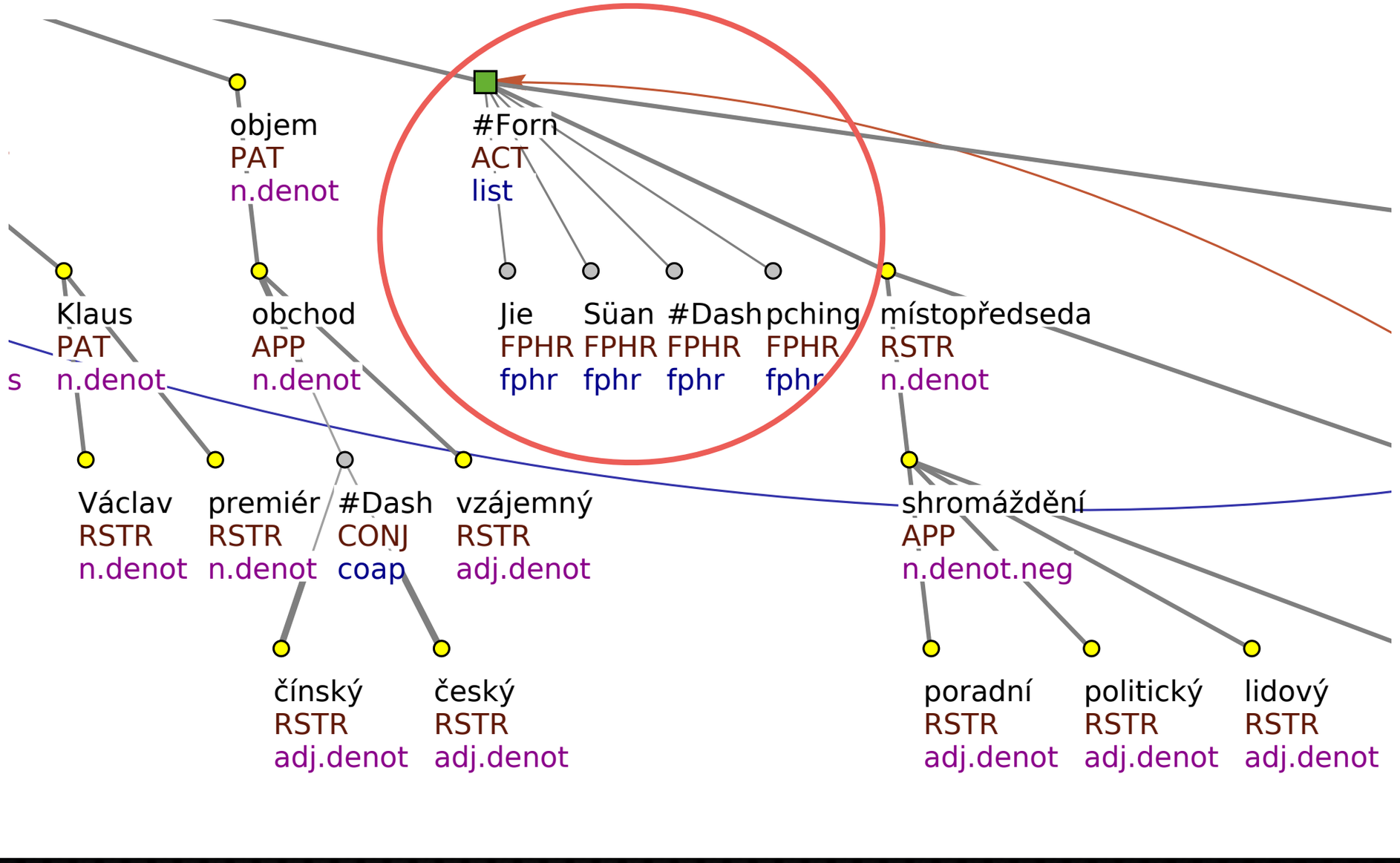
PARSEME

ÚFAL



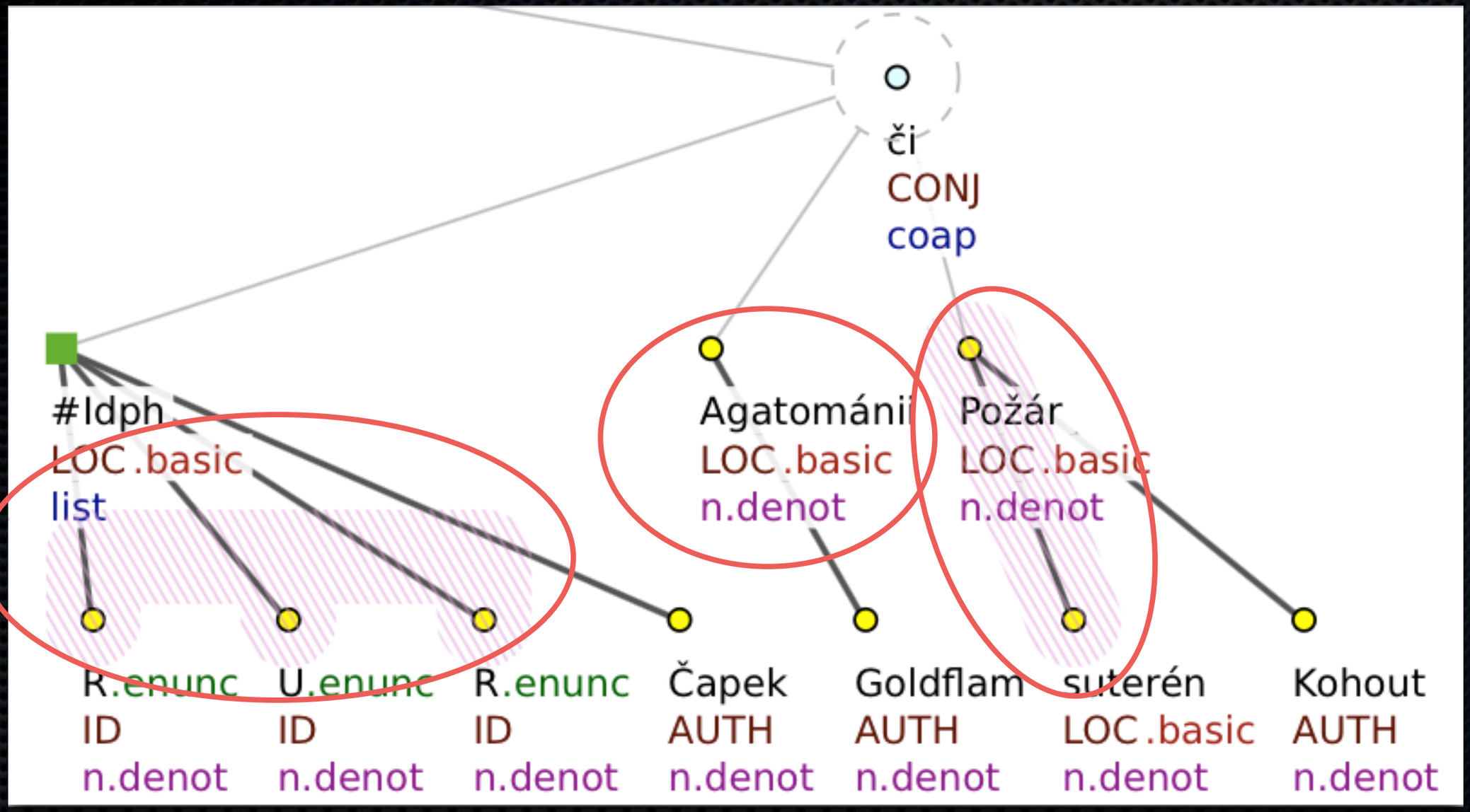
Foreign phrases

Czech vs. Chinese name of a person



Foreign phrases

Czech vs. Chinese name of a person



Identifying phrases

Often proper nouns, unfortunately not always. 3 book titles here, only the first #Idph.

#Forn and ID

- Foreign expressions are not always governed by #Idph and have ID functor

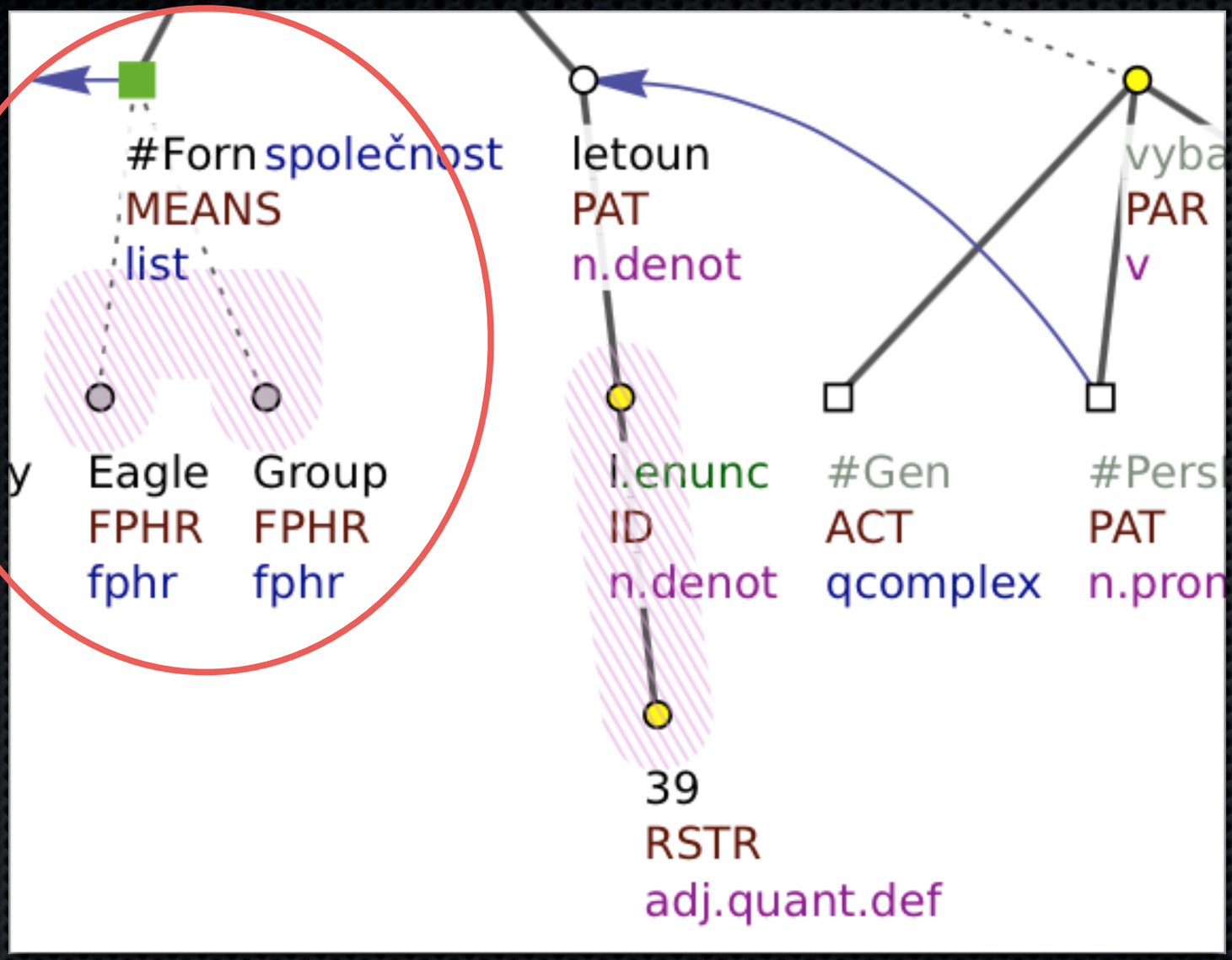
```
t-node [ t_lemma = '#Forn', functor = 'ID' ]
>>count()
660

t-node [ t_lemma = '#Forn', functor != 'ID' ]
>>count()
834
```

- Does this mean that non-IDs are not proper nouns?

PARSEME

ÚFAL




```
t-node [ t_lemma = ‘#Forn', functor != 'ID' ]
```

Some foreign multi-word proper nouns have the functor ID, some have other deep syntactic functions.

What about all those bubbles?

They seem to mark MWEs more consistently

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

PARSEME

ÚFAL

Tectogram. structure of a MWE

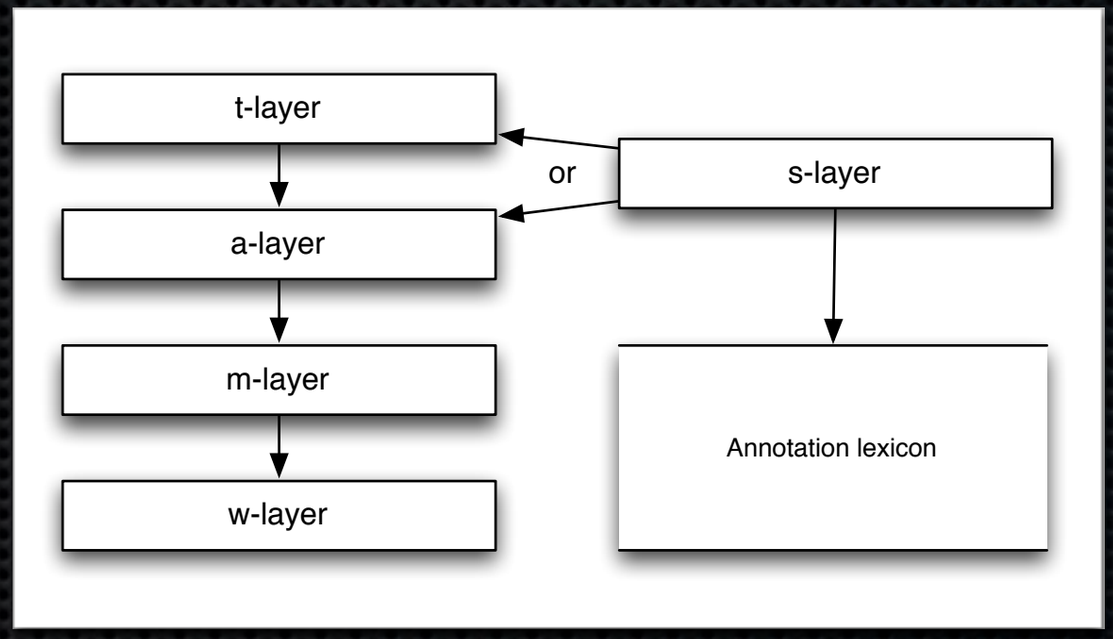
- **hypothesis**: dependency structure + deep word order: each MWE should only have 1 t-structure, that should always be contiguous
- means of effective automatic identification (given t-trees)
- 771 SemLex entries have more than one t-structure in data:
- systematic deficiencies of PDT 2.0 t-lemmata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

PARSEME

ÚFAL

PDT 2.0 + s-data

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



PARSEME

ÚFAL

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 třináct 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.

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.

Typical texts with MWEs

colours specify multiword lexemes vs. types of named entities

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.

(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) (Ústavního|t-mf930709-001-p3s1Bw14) (soudu|t-mf930709-001-p3s1Bw15). (Řekl|t-mf930709-001-p3s2w1)

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 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 evčík (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 Vladim 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-p3s4w
(skládá|t-mf930709-001-p3s4w4) (z|t-mf930709-001-p3s4w7) (15|t-mf930709-001-p3s4w6) (soudců|t-mf930709-001-p3s4w7),
(jmenovaných|t-mf930709-001-p3s4w9) (na|t-mf930709-001-p3s4w12) (deset|t-mf930709-001-p3s4w11)

LexemAnn - mf930709_001

Soubor Úpravy Debug

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

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.

markup (show, remove, NEs)

Obecné: Ukázat | Odstranit

Pojmenované entity: Jméno | Instituce | Místo | Objekt | Adresa | Čas | Biblio | Foreign | X

Automatická anotace

SemLex Editor

Označkovat | Nové heslo | Ulož heslo | A=a | Hledat | P | N

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

Příklad: Synonyma:

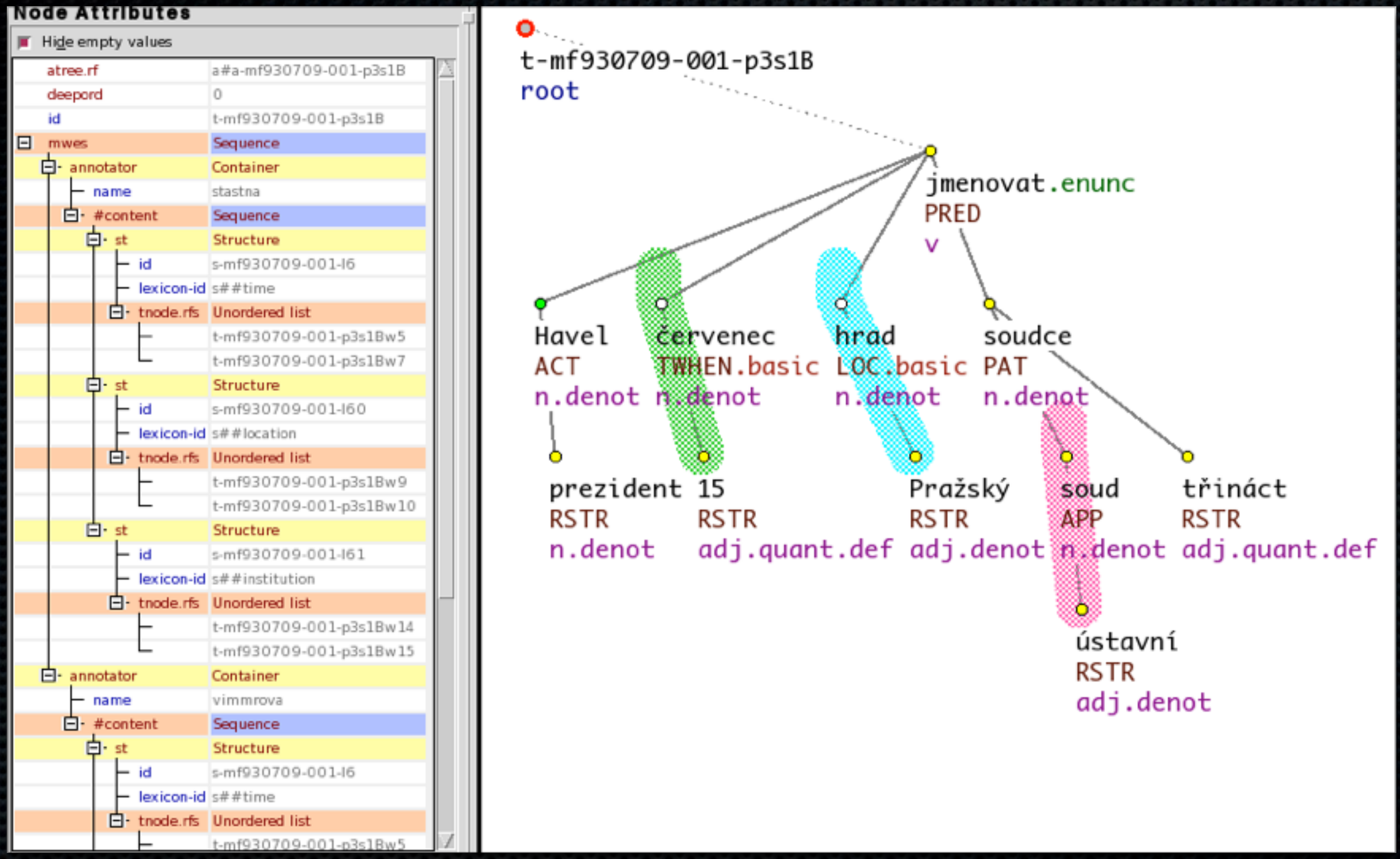
Glosa: Změněno: 090607125630 merger

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

PARSEME

ÚFAL



MWEs in a tree, attributes listed

Statistics[1]

- 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%

PARSEME

ÚFAL

Statistics[2]

```
# List the MWE types
t-root $r := [ ];
  >> for $r.mwes/type give
$1,count() sort by $2 desc
```

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

PARSEME

ÚFAL

Semlex

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

```
!!perl/hash:SemLex_heslo
BASIC_FORM: deficit státního rozpočtu
CREATED: '110914162730'
EXAMPLE: ~
GLOSS: ~
ID: '0000032344'
LEMMATIZED: deficit státní rozpočet
MODIFIED: ~
MODIFIER: bejcek
MORPHO_TAGS: ''
ORIGID: ~
PDT25_FREQ: 2
POS: 'N'
SOURCE: stastna
SYNONYMS: []

TREE_STRUCT:
  -
    - deficit
    - ~
  -
    - rozpočet
    - 0
  -
    - státní
    - 1
```

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

PARSEME

ÚFAL

Problem:
t-lemma not generic enough

- All of these variations share the basic lexical meaning
- They should be specified by attributes explicating the relationship to the basic lemma (meaning) – **someday**
  - Diminutives: dům, domek, domeček (1st, 2nd degree)
  - Gender opposites: ředitelka (female director)
  - Lemma variants: občanský zákoník (citizen law codex)
- 771 of 8816 Semlex entries with >1 tree-structure

PARSEME

ÚFAL

More Problems

t-layer

- empty nodes
  - especially in coordinations (red [wine] and white wine)

Semlex

- Too simple tree structures
  - Aux representation needed for MWE semantics
  - “na zdraví” – [on health] meaning “cheers”
    - “na” isn’t there in the t-layer

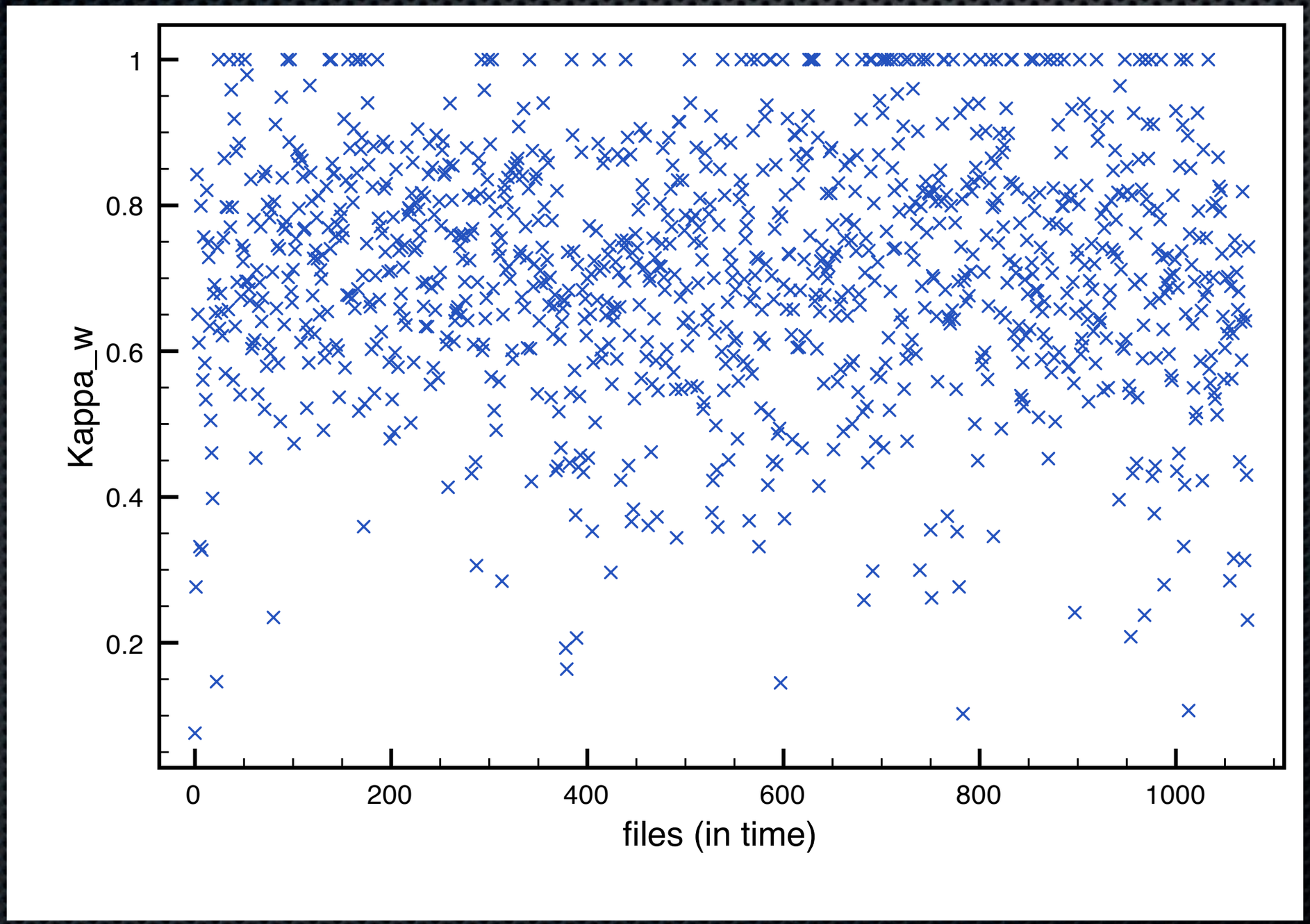
Inter-annotator agreement

- Modified Weighted Kappa (Cohen's)
- Weight function with values indicating "relative information" of a tag
- Calculated for each pair of annotators (for which we have parallel data)
- Over all parallel data, per batch of files, and per file
- Final pairwise Kappa around 0.7

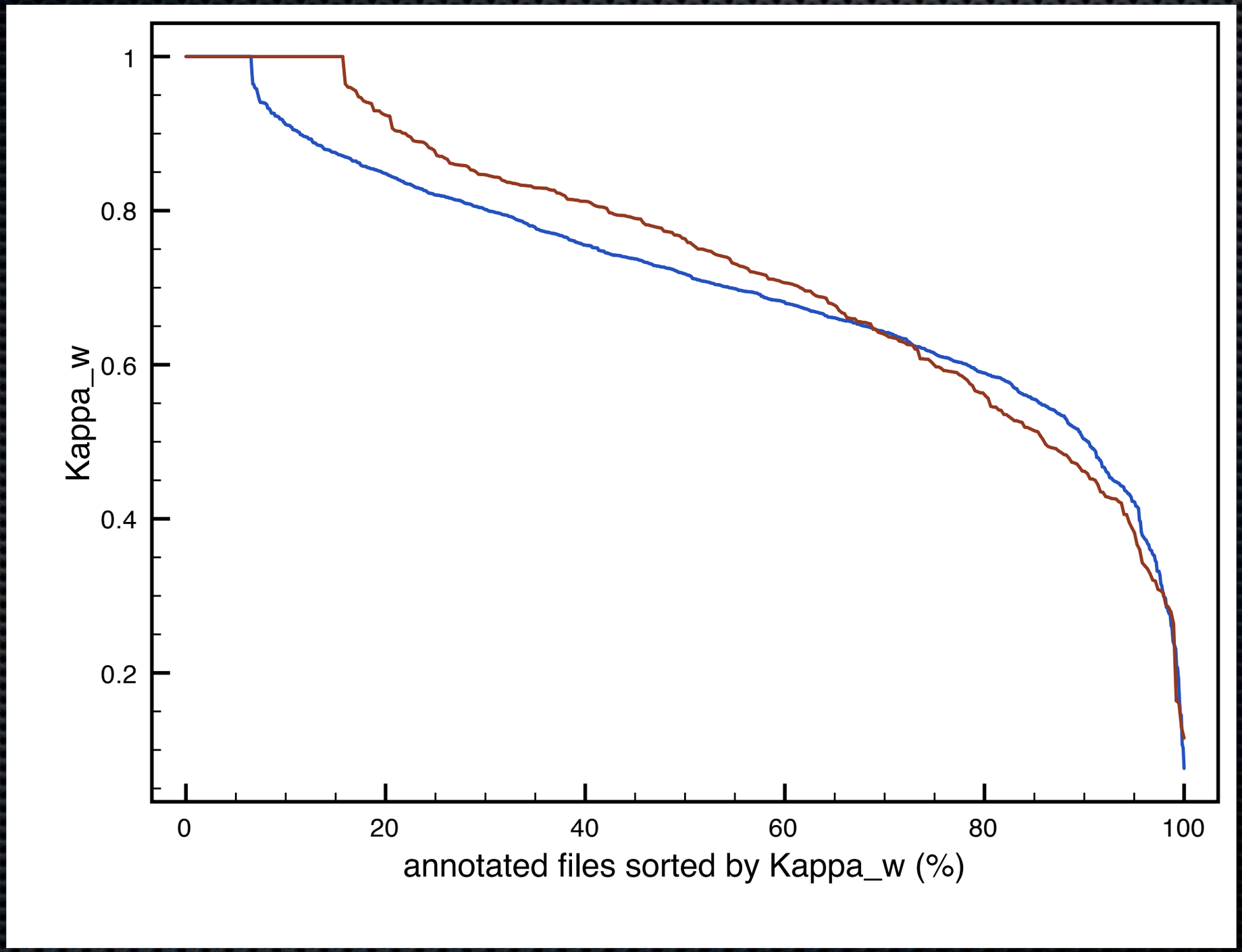
PARSEME

ÚFAL

Pairwise Kappa per file



Pairwise Kappa per file

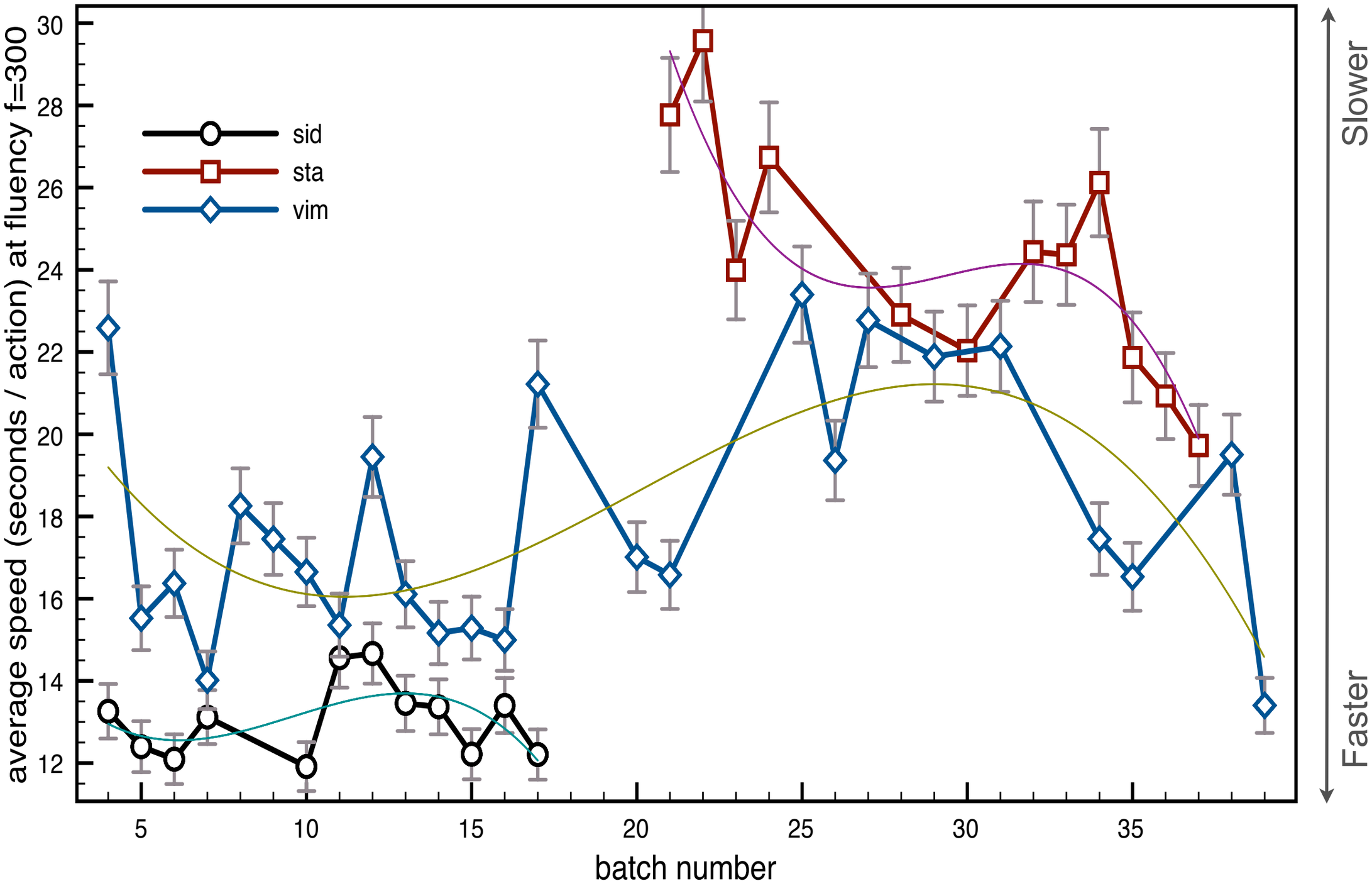


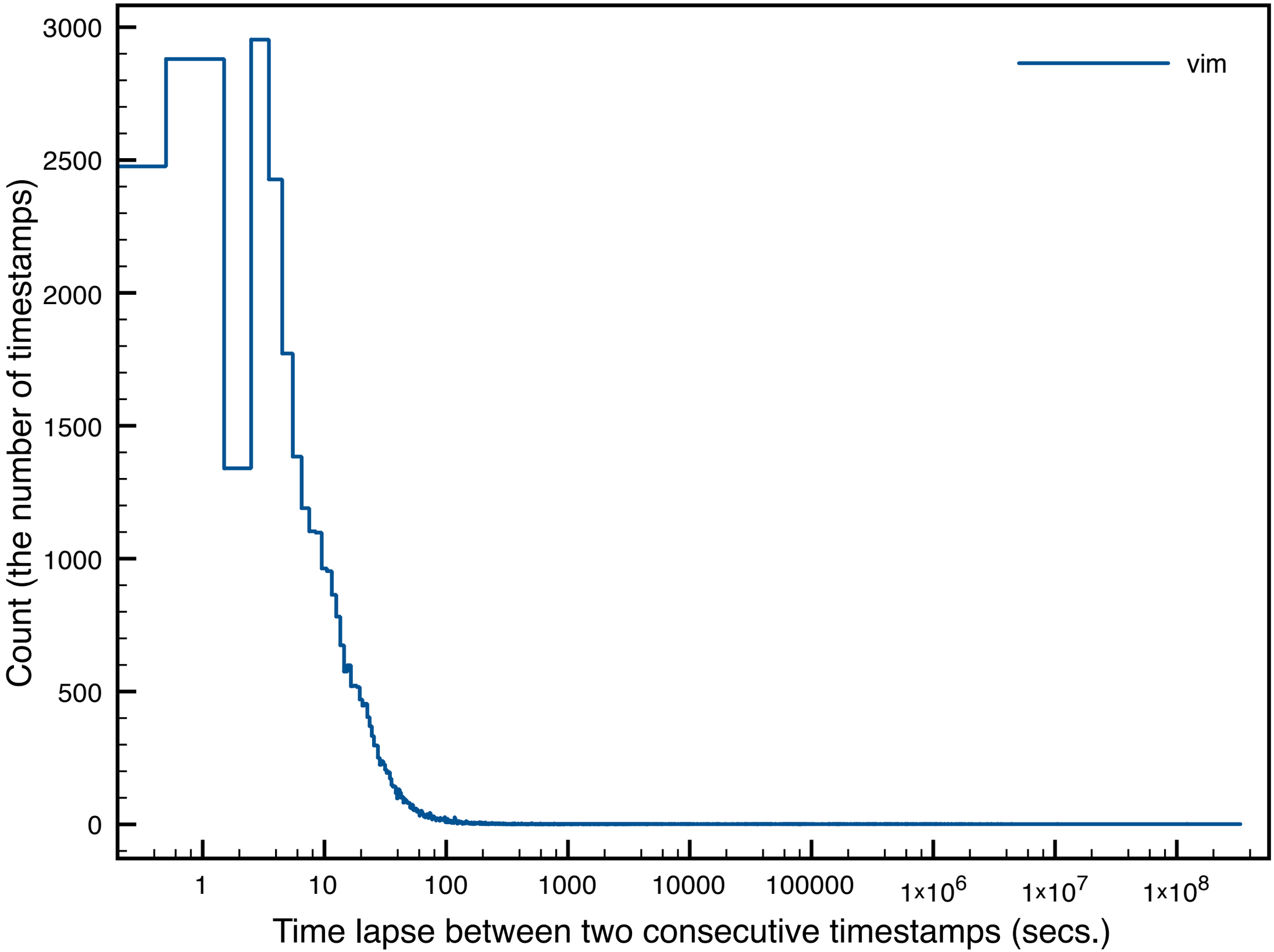
Speed analysis

- Log files for both undo/redo and analysis of annotators' workflow
- Each action is time stamped
- A preliminary analysis of just the timestamps
- Used also to establish annotators' wages

PARSEME

ÚFAL


sid
sta
vim
average speed (seconds / action) at fluency f=300
30
28
26
24
22
20
18
16
14
12
5
10
15
20
25
30
35
batch number
Slower
Faster


vim
3000
2500
2000
1500
1000
500
0
Count (the number of timestamps)
1
10
100
1000
10000
100000
1x10^6
1x10^7
1x10^8
Time lapse between two consecutive timestamps (secs.)

All data and tools are public, except SemLex:

http://ufal.mff.cuni.cz/lexemann/mwe/

Gold standard version is included in PDT since v2.5

PARSEME

ÚFAL

MWE annotation vs. DPHRs

```
# DPHRs agree with MWEs

t-node $n3 :=
 [ functor = "DPHR",
    same-tree-as t-root
    [ member mwes
        [ tnode.rfs $n3 ]
] ];

>>count()
```

335 occurrences

být	na_stejné_lodi
mít	na_mysli
přijít	na_řadu
být	k_dispozici
pád	tím
stát	co_stůj

PARSEME

ÚFAL

MWE annotation vs. DPHRs[2]

```
# DPHRs DO NOT agree with
MWEs

t-node $n :=
[ functor = "DPHR",
    0x same-tree-as t-root
    [ member mwes
        [ tnode.rfs $n3 ]
] ];

>>count()
```

800 occurrences 😞

dávat	váhu
dát	za_pravdu
mít	k_dispozici
hodit	jiskru
spadnout	z_nebe
dostat	přes_prsty

PARSEME

ÚFAL

MWE annotation vs. DPHRs[3]

- 335 agreements vs. 800 disagreements
- Agreements and disagreements look rather similar
- Clearly work to be done

PARSEME

ÚFAL

MWE annotation vs. `is_name_of_person`

`is_name_of_person` is an attribute: heuristic used in PDT 2.0 to identify names of persons

- MWE annotators also annotated names, as it is a common type of named entities
- In future we would like to have both structure of the Person NEs (as in NE taggers: title+first+last+title) and grounding (e.g. links to wikipedia + extracted attributes in a dictionary)

PARSEME

ÚFAL

Person²

error of the heuristics

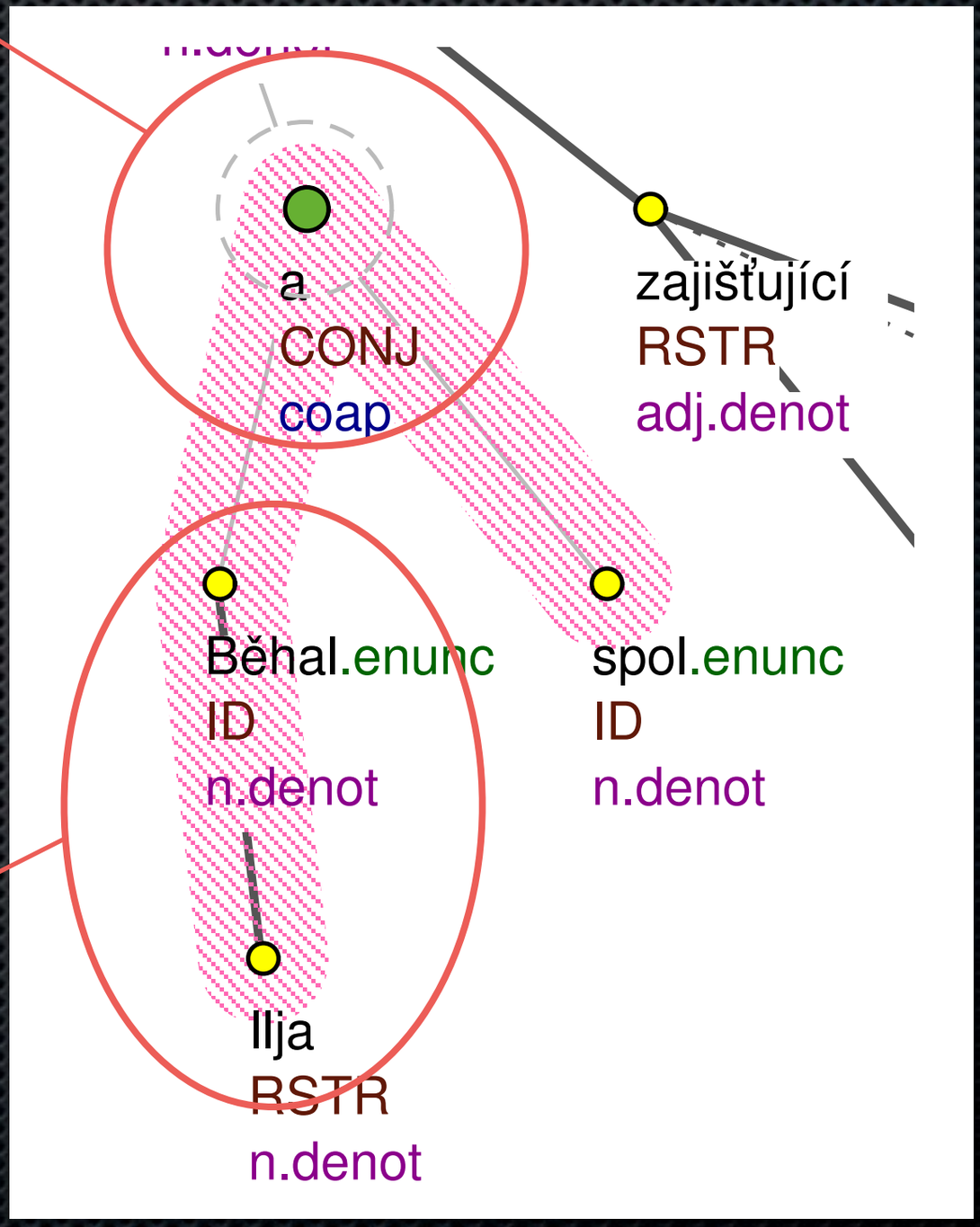
```
# is_name_of_person not
annotated as MWE: PERSON

t-node $n :=
[ is_name_of_person = "1",
    0x same-tree-as t-root
    [ member mwes
[ type = “person", tnode.rfs $n
] ] ];

>>count()
```

9352

(**398** the other way, e.g. Kaláb->doktor)



No embedded entities (yet)

“Ilja Běhal and co.”

MWEs in PDT Summary

- **DPHR** – should be marked as MWEs, but they are not
  - To be done
- **is_name_of_person**
  - different instructions: also single-words, no titles
  - sometimes errors
- **NEs** – add to Semlex, add structures, add grounding

PARSEME

ÚFAL

MWEs in PDT Summary[2]

- **Lexemes**
  - t_lemmas need to be more generic (e.g. no gender)
  - tree structures in Semlex need to capture more:
    - auxiliaries
    - morphological and other surface limitations

PARSEME

ÚFAL

MWEs in PDT Summary[3]

- Quite some work remains to be done, but:
- We still have a large treebank with deep annotation of a very large number of multiword expressions
- 43955 sentences, 43280 include MWE annotation
- Semlex: a lexicon of over 8 000 lexemes, most of them newly identified (not present in existing lexicons)

PARSEME

ÚFAL