



Towards a Conversion of the Prague Dependency Treebank Data to the Uniform Meaning Representation

Markéta Lopatková, Eva Fučíková, Federica Gamba,
Jan Štěpánek, Daniel Zeman and Šárka Zikánová

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

*Supported by the LUSyD project (GAČR, no. 20-16819X) and the LINDAT/CLARIAH-CZ project (MŠMT, no. LM2023062);
partially supported by CUNI (GAUK, project no. 104924, and SVV, project no. 260 698).*



Motivation and Goal

Meaning representation

- intriguing theoretical problem
- its practical implications for applications
 - interlingua for machine translation
 - a basis for knowledge representation and knowledge systems
- a sound and reliable basis for logical inference



LLM dominates the field, BUT

- problems with hallucinating
- tend to fabricate information

Motivation and Goal

Meaning representation

- intriguing theoretical problem
- its practical implications for applications
 - interlingua for machine translation
 - a basis for knowledge representation and knowledge systems
- a sound and reliable basis for logical inference



LLM dominates the field, BUT

- problems with hallucinating
- tend to fabricate information

Goal:

- compare 2 meaning representations
 - with different linguistic traditions, based on different theoretical assumptions, with different focuses
- a substantially deeper understanding of language semantics

Two Meaning Representations at a Glance

PDT-MR

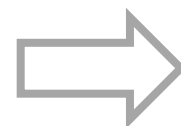
- theory: Functional Generative Description
(esp. Sgall et al, 1967; 1986; 2020)
- data and tools: treebank (esp. Hajič et al., 2020)
Czech (~130k sentences); English (~55k); Latin (~5k)
- dependency-oriented formalism
- covers:
 - deep and surface syntax (argument structure)
 - meaning-relevant morphology (tense, modality)
 - coreference annotation
 - information structure and discourse relations



focus on **meaning as structured**
by **the given language**
more-or-less **directly refers to the text**

UMR

- semantics, abstracting away from syntax
(esp. van Gysel et al, 2018; Bonn et al, 2013)
- typological perspective
- limited data, no supporting infrastructure
6 languages (~ 2k sentences)
- (directed) acyclic graphs
- covers:
 - argument structure
 - multiword expressions, named entities
 - enhanced info on aspect, modality, temporality
 - coreference



broad **sem. interpretation** of the text
for cross-lingual applications

Two Meaning Representations at a Glance

PDT-MR

- theory: Functional Generative Description
(esp. Sgall et al, 1967; 1986; 2020)
- data and tools: treebank (esp. Hajič et al., 2020)
Czech (~130k sentences); English (~55k); Latin (~5k)
- dependency-oriented formalism
- covers:
 - deep and surface syntax (argument structure)
 - meaning-relevant morphology (tense, modality)
 - coreference annotation
 - information structure and discourse relations

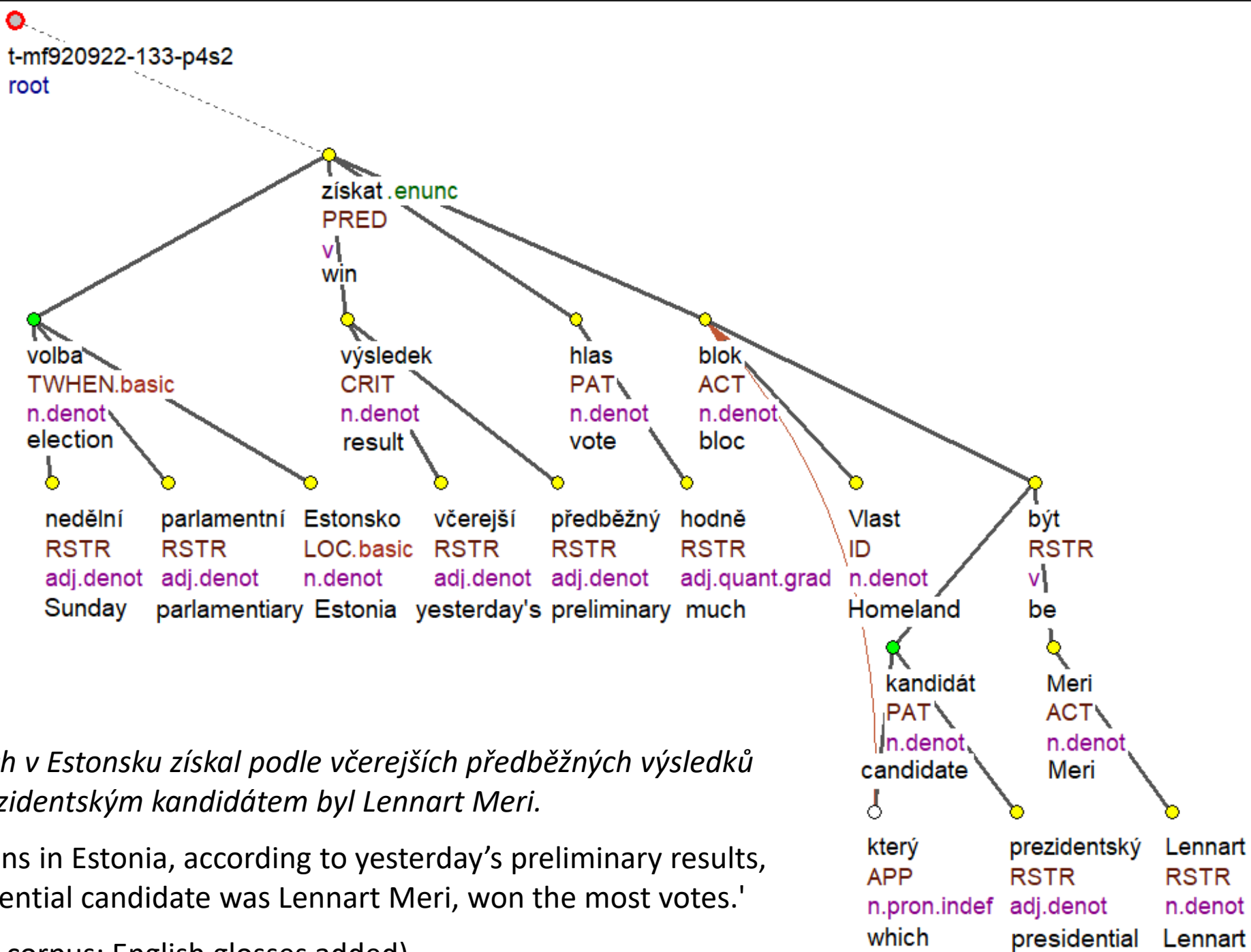

 focus on **meaning as structured**
 by the given language
 more-or-less **directly refers to the text**

UMR

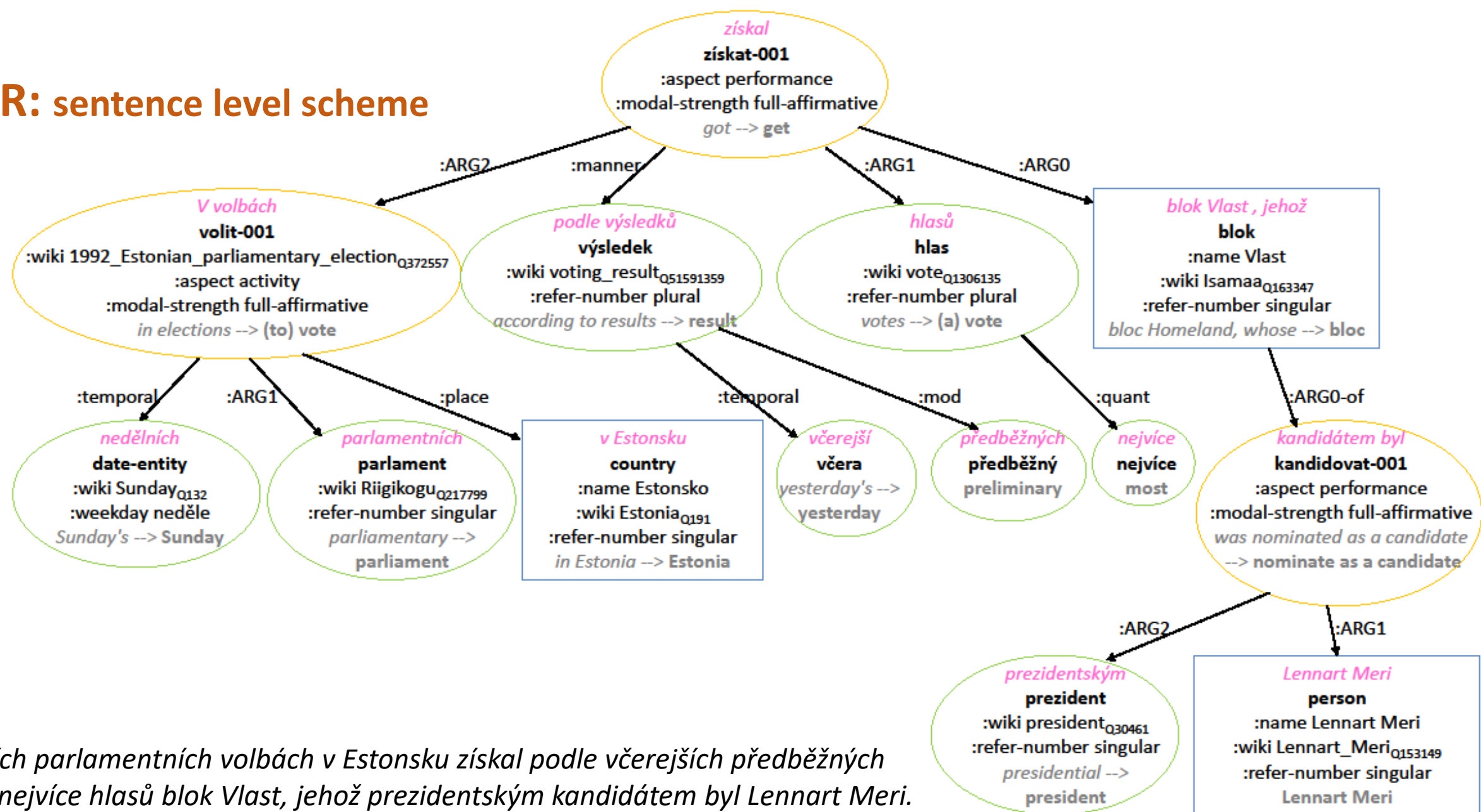
- semantics, abstracting away from syntax
(esp. van Gysel et al, 2018; Bonn et al, 2013)
- typological perspective
- limited data, no supporting infrastructure
6 languages (~ 2k sentences)
- (directed) acyclic graphs
- covers:
 - argument structure
 - multiword expressions, named entities
 - enhanced info on aspect, modality, temporality
 - coreference


 broad **sem. interpretation** of the text
 for cross-lingual applications

PDT-MR



UMR: sentence level scheme



V nedělních parlamentních volbách v Estonsku získal podle včerejších předběžných výsledků nejvíce hlasů blok Vlast, jehož prezidentským kandidátem byl Lennart Meri.

'In Sunday's parliamentary elections in Estonia, according to yesterday's preliminary results, the Homeland bloc, whose presidential candidate was Lennart Meri, won the most votes.'

UMR: document level scheme

```
(s5s0 / sentence
  :temporal ((document-creation-time :before s5v3)
    (s5v3 :before s5d)
    (s5d :before s5k)
    (s5d :contained s5z)
    (s5d :contained s5v)
    (s5v :after s5z))
  :modal ((root :modal author)
    (author :full-affirmative s5v)
    (author :full-affirmative s5k)
    (author :full-affirmative s5z))
  :coref ((s3c :same-entity s5c)
    (s3p3 :same-entity s5p)
    (s3v :same-event s5v)))
```

včera 'yesterday'

neděle 'Sunday' (date-entity)

kandidovat-001
'nominate as a candidate'

získat-001 'get'

volit-001 'vote'

V nedělních parlamentních volbách v Estonsku získal podle včerejších předběžných výsledků nejvíce hlasů blok Vlast, jehož prezidentským kandidátem byl Lennart Meri.

'In Sunday's parliamentary elections in Estonia, according to yesterday's preliminary results, the Homeland bloc, whose presidential candidate was Lennart Meri, won the most votes.'

Towards PDT-MR to UMR Conversion

Selected deep syntactic phenomena

I. change of the graph structure

- coreference relation: re-entrancies, inverse roles, listing
- coordination (and re-entrancies)

II. events vs. entities

III. graph labeling:

- valency frames → argument structure
 - verb specific mapping of arguments
 - default mapping of arguments
- default mappings of adjuncts

I. Change of the Graph Structure: Coreference



coreference \approx relation between two or more expressions that refer to the same concept

"words"

"mental concept"
of a real-world
entity/event

- such expressions typically form coreferential chains \rightarrow text coherence

Mary lives in Prague. *She* likes ice-cream. *The girl* decided \emptyset to go for a trip.

←
antecedent

←
anafor

I. Change of the Graph Structure: Coreference



coreference \approx relation between two or more expressions that refer to the same concept

"words"

"mental concept"
of a real-world
entity/event

- such expressions typically form coreferential chains \rightarrow text coherence

Mary lives in Prague. *She* likes ice-cream. *The girl* decided \emptyset to go for a trip.

←
antecedent

anafor

I. Change of the Graph Structure: Coreference



coreference \approx relation between two or more expressions that refer to the same concept

"words"

"mental concept"
of a real-world
entity/event

- such expressions typically form coreferential chains \rightarrow text coherence

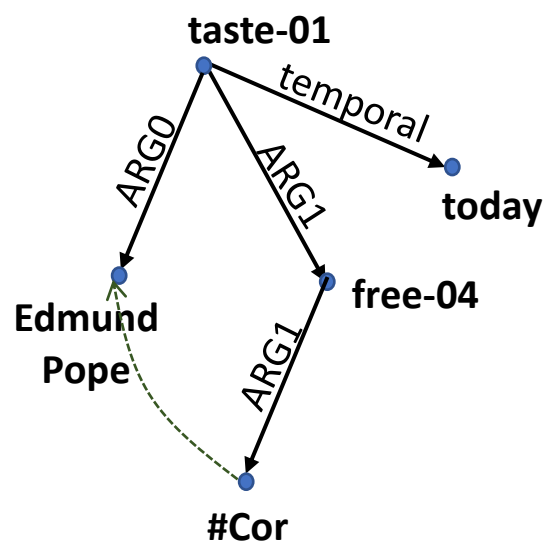
Mary lives in Prague. *She* likes ice-cream. *The girl* decided \emptyset to go for a trip.

$\xleftarrow{\text{antecedent}}$ $\xleftarrow{\text{anafor}}$

- all types: the same representation in PDT-MR:
 - (the node for) the anaphor bears attributes for ID of its antecedent(s), type of relation
- BUT different treatment in UMR

Ia. PDT-MR Coreference → UMR "Re-entrancy"

Coreference of 2 nodes in PDT-MR



Edmund Pope tasted freedom today.

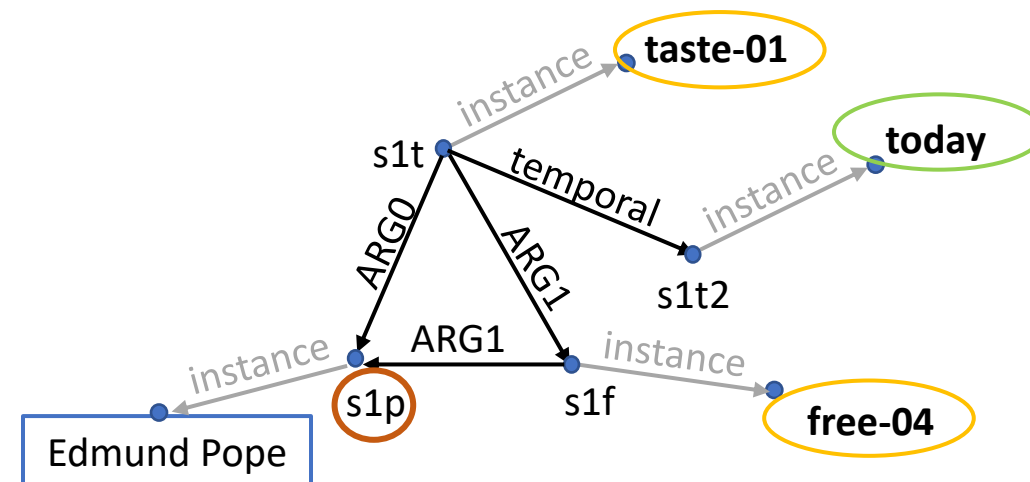
(taken from the released UMR data, simplified;
also used as an example sentence in the UMR 0.9 Specification)

Ia. PDT-MR Coreference → UMR "Re-entrancy"



Concept of re-entrancy in UMR

```
(s1t / taste-01
  :ARG0 (s1p / person :wiki "Edmund_Pope"
    :name (s1n / name
      :op1 "Edmund"
      :op2 "Pope")))
:ARG1 (s1f / free-04
  :ARG1 (s1p)
  :temporal (s1t2 / today))
```

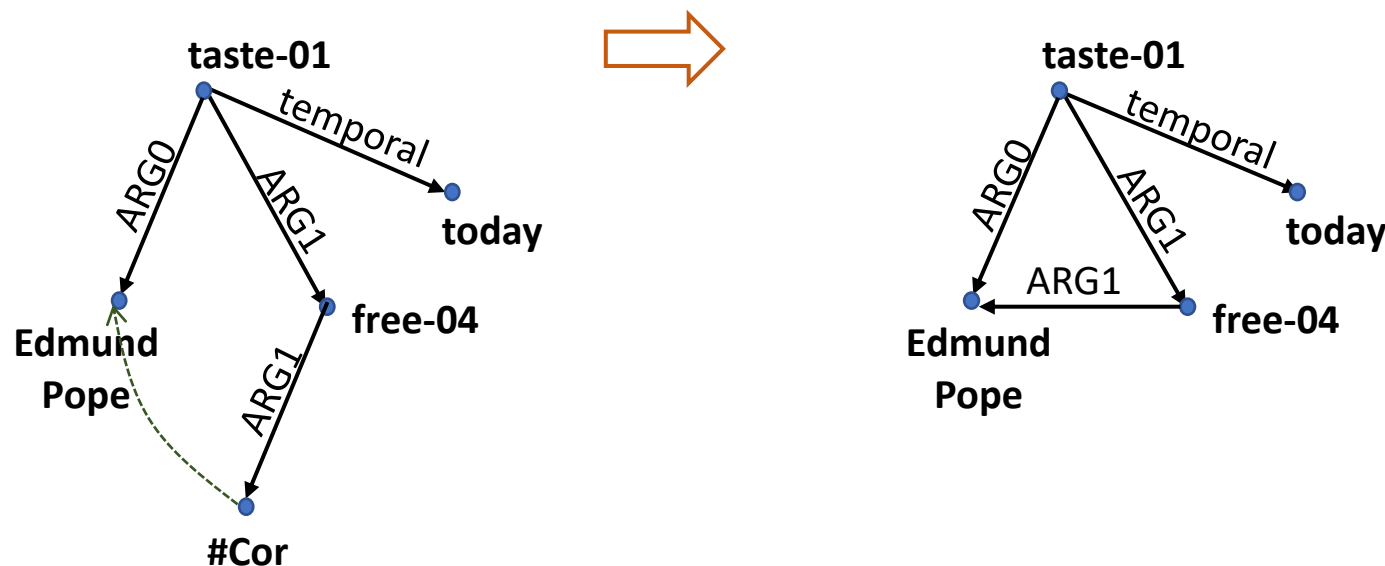


Edmund Pope tasted freedom today.

(taken from the released UMR data, simplified;
also used as an example sentence in the UMR 0.9 Specification)

Ia. PDT-MR Coreference → UMR "Re-entrancy"

Conversion: Merging 2 nodes in PDT



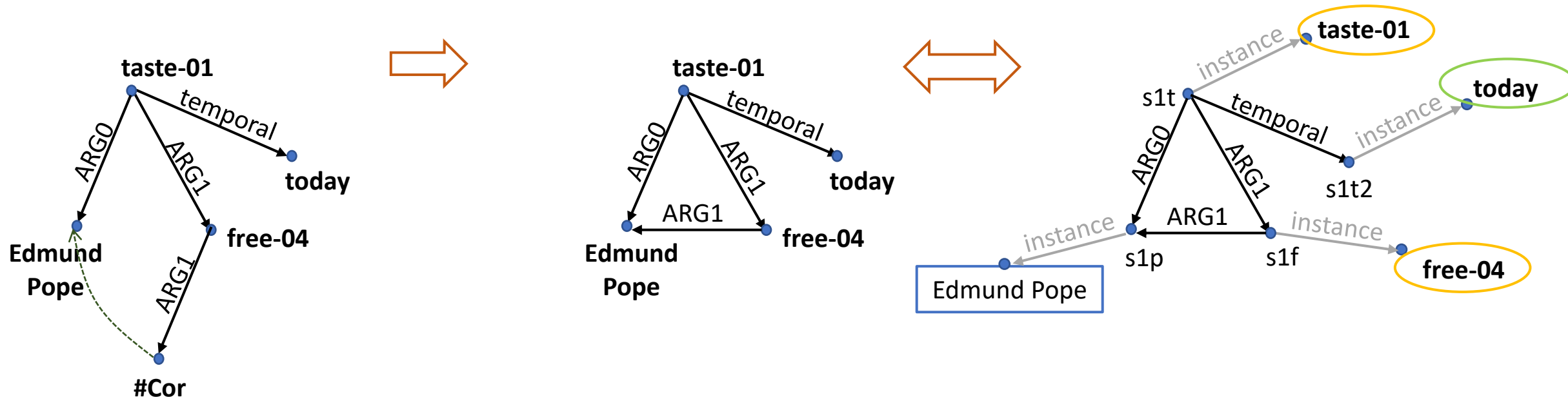
Edmund Pope tasted freedom today.

(taken from the released UMR data, simplified;
also used as an example sentence in the UMR 0.9 Specification)

Ia. PDT-MR Coreference → UMR "Re-entrancy"



Conversion: Merging 2 nodes in PDT

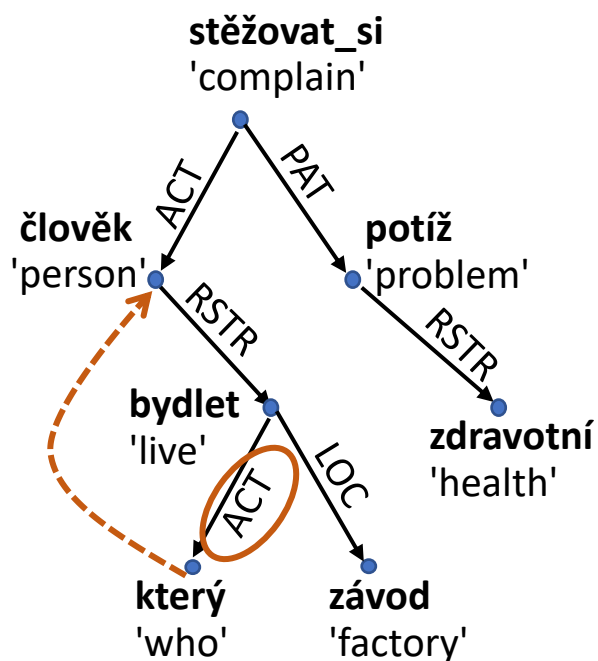


Edmund Pope tasted freedom today.

(taken from the released UMR data, simplified;
also used as an example sentence in the UMR 0.9 Specification)

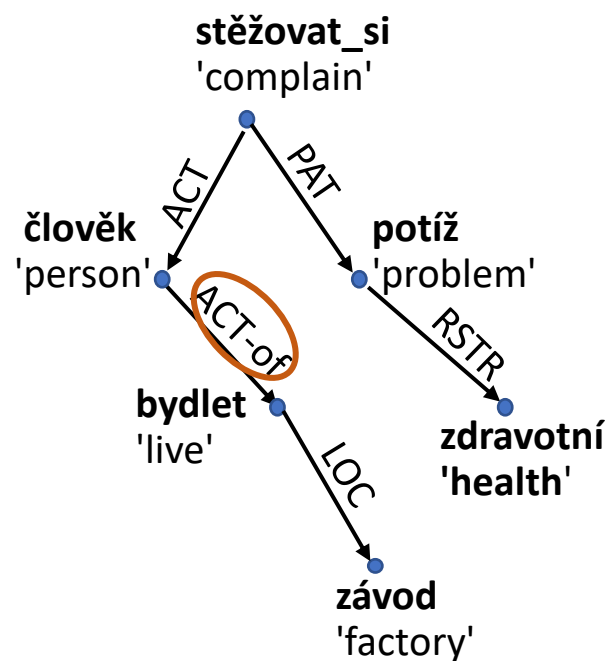
Ib. PDT-MR Coreference → UMR Inverse role

Coreference of 2 nodes in PDT-MR



Merging 2 nodes in PDT

Inverse role (= inverse relation) in UMR

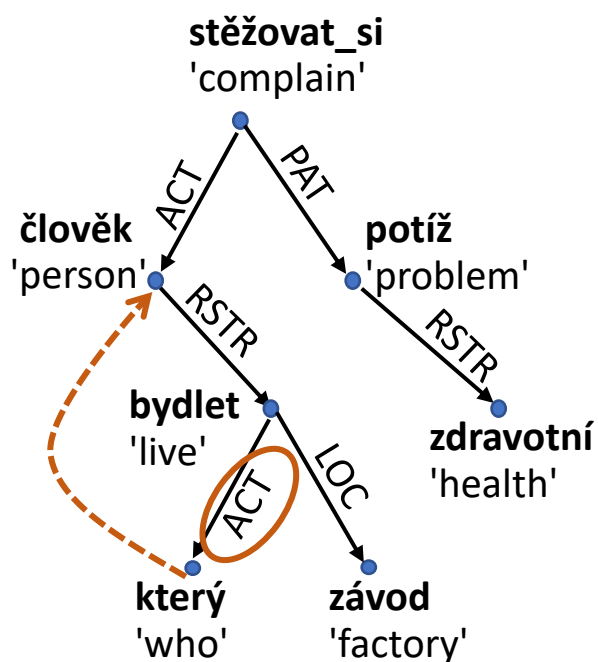


Lidé, kteří bydlí v blízkosti závodu, si stěžují na zdravotní potíže

‘People who live near the factory have been complaining of health problems’.

Ib. PDT-MR Coreference → UMR Inverse role

Coreference of 2 nodes in PDT-MR

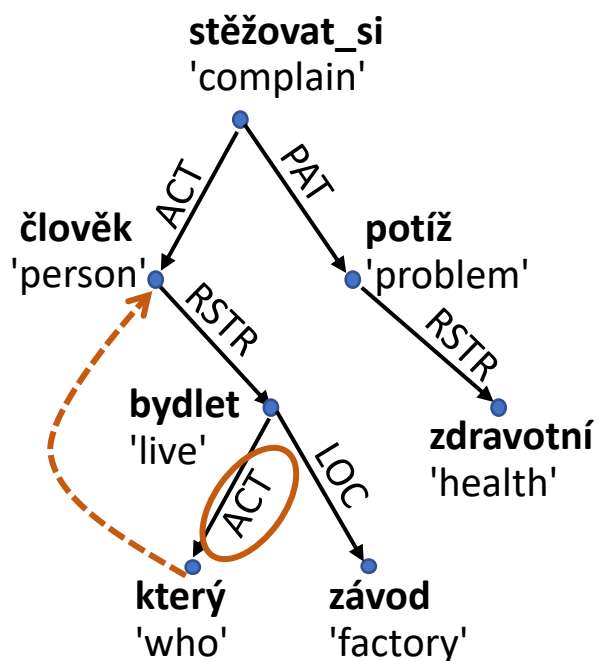


Lidé, kteří bydlí v blízkosti závodu, si stěžují na zdravotní potíže

‘People who live near the factory have been complaining of health problems’.

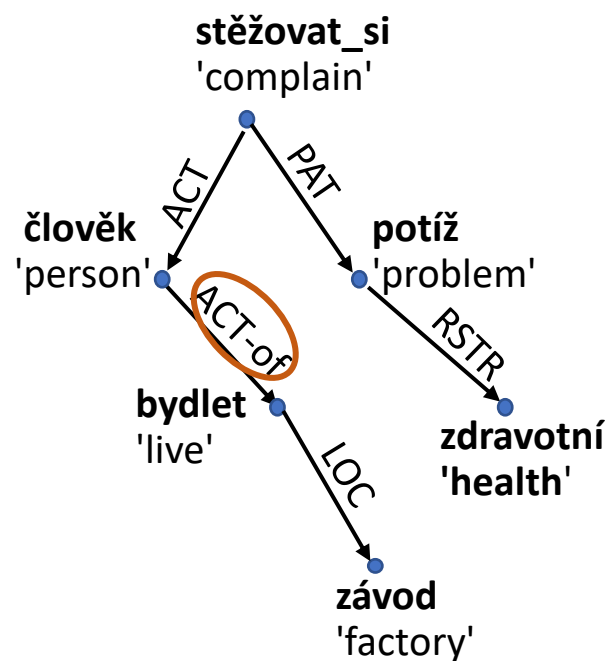
Ib. PDT-MR Coreference → UMR Inverse role

Coreference of 2 nodes in PDT-MR



Merging 2 nodes in PDT

Inverse role (= inverse relation) in UMR



Lidé, kteří bydlí v blízkosti závodu, si stěžují na zdravotní potíže

‘People who live near the factory have been complaining of health problems’.

Ic. PDT-MR Coreference → UMR Pairing

Inter-sentence coreference relation

PDT-MR

- the node for) the anafor bears attributes for
 - ID of its antecedent(s)
 - type of relation
 - type of reference (specific vs. generic)

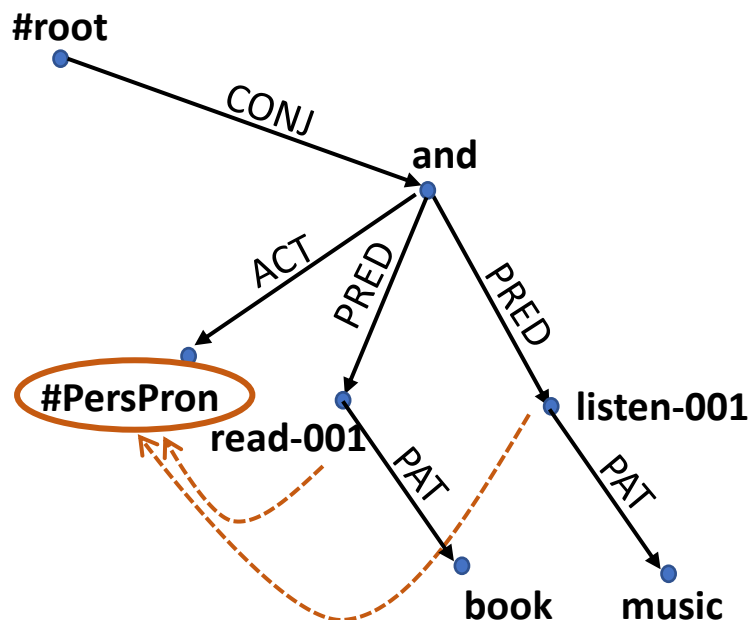
UMR

- lists pairs of coreferring concpets
 - ✓ ID of both concepts
 - event or entity ... entities ✓
 - identity or subset ... identity ✓

Id. Coordination

PDT-MR

- special node for coordinating expression
- coordinated expressions as children
- allows for common arguments/adjuncts

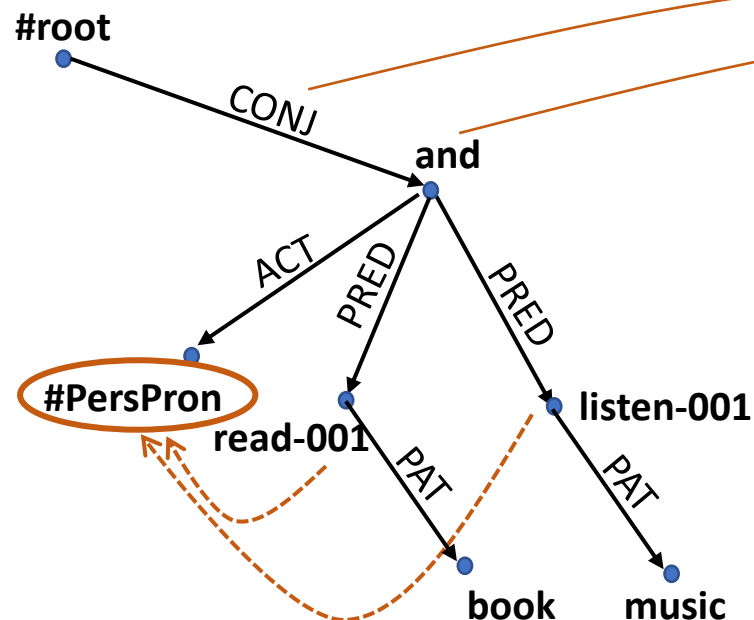


I read a book and listened to music. /
I read a book while listening to music. /
I read a book while I listened to music.

Id. Coordination

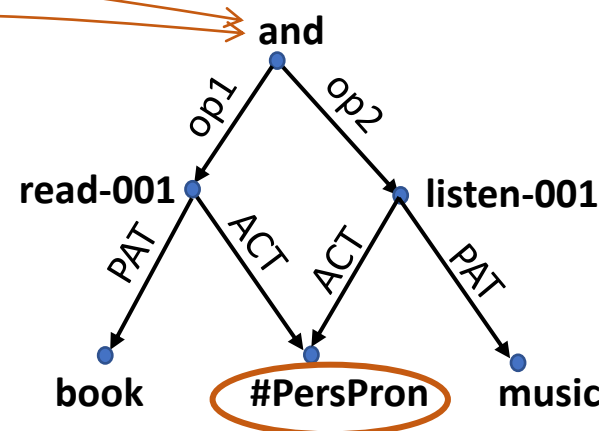
PDT-MR

- special node for coordinating expression
- coordinated expressions as children
- allows for common arguments/adjuncts



UMR

- special keyword for "discourse" relation
- coordinated expressions as children
- allows for common arguments/adjuncts



I read a book and listened to music. /
I read a book while listening to music. /
I read a book while I listened to music.

II. Events vs. Entities

UMR

- conceptual distinction:
 - entities (objects) *man, cat*
 - states (properties) *tall, (to) love*
 - events (processes) *cry, storm, elections*
- no clear definition,
missing criteria for distinguishing
- skewed towards English (e.g., statives)
- big impact on annotation
 - modal, temporal, aspectual for events



- **fuzzy boundary** between entities and events
- big space for **different interpretations**
- intuitive decisions



II. Events vs. Entities

PDT-MR

- only small degree of abstraction in PDT
 - e.g., *matčin* 'mother's' → *matka* 'mother' + possessive
 - "normalization", e.g., *jehož* → *který* 'who'
- lack of information
 - even for most systematic changes
 - e.g., *bojování* 'fighting' → *bojovat* '(to fight)'

conversion:

first steps using additional resources

UMR

- conceptual distinction:
 - entities (objects) *man, cat*
 - states (properties) *tall, (to) love*
 - events (processes) *cry, storm, elections*
- no clear definition, missing criteria for distinguishing
- skewed towards English (e.g., statives)
- big impact on annotation
 - modal, temporal, aspectual for events

- **fuzzy boundary** between entities and events
- big space for **different interpretations**
- intuitive decisions



III. Graph labeling

PDT-MR

arguments:

- PDT-Vallex valency lexicon (Hajič et al., 2003)
 - verbs, nouns (adjectives)
 - elaborated valency theory
 - 5 "arguments": ACT, PAT, ADDR, ORIG, EFF

UMR

arguments:

- PropBank lexicon (Palmer et al 2005, Pradhan et al., 2022)
 - verbs, nouns (adjectives)
 - coarse-grained semantic roles
 - ARG0, ARG1, ... ARG5, ARGM

III. Graph labeling

PDT-MR

arguments:

- PDT-Vallex valency lexicon (Hajič et al., 2003)
 - verbs, nouns (adjectives)
 - elaborated valency theory
 - 5 "arguments": ACT, PAT, ADDR, ORIG, EFF

UMR

arguments:

- PropBank lexicon (Palmer et al 2005, Pradhan et al., 2022)
 - verbs, nouns (adjectives)
 - coarse-grained semantic roles
 - ARG0, ARG1, ... ARG5, ARGM



partial mapping verb-specific mapping

~ 43% of PDT-Vallex labels (out of 42,116) (Hajič et al, 2024)

default mapping for the rest verb senses

most frequent argument mappings from the previous

III. Graph labeling

PDT-MR

arguments:

- PDT-Vallex valency lexicon (Hajič et al., 2003)
 - verbs, nouns (adjectives)
 - elaborated valency theory
 - 5 "arguments": ACT, PAT, ADDR, ORIG, EFF

UMR

arguments:

- PropBank lexicon (Palmer et al 2005, Pradhan et al., 2022)
 - verbs, nouns (adjectives)
 - coarse-grained semantic roles
 - ARG0, ARG1, ... ARG5, ARGM



partial mapping verb-specific mapping

~ 43% of PDT-Vallex labels (out of 42,116) (Hajič et al, 2024)

default mapping for the rest verb senses

most frequent argument mappings from the previous

adjuncts:



default mapping based on their semantics
further refined where necessary



What We Have Learned

PDT-MR

- **theory:**

meaning as structured by the particular language

- How different for various language?
- consequence: too close to the text

UMR

- **theory:**

meaning representation as a basis for logical inference

- (BUT not much investigated so far)
- Should be language independent (theory)!
- consequence: broad interpretation

What We Have Learned

PDT-MR

- **theory:**

meaning as structured by the particular language

→ How different for various language?

→ consequence: too close to the text

- **data annotation:**

refined criteria how to annotate

many "running text" examples

stress on consistency of annotation

(→ consequences for ML)

UMR

- **theory:**

meaning representation as a basis for logical inference

(BUT not much investigated so far)

→ Should be language independent (theory)!

→ consequence: broad interpretation

- **data annotation:**

vague description

small number of examples (to illustrate the theory)

interest in the annotator's understanding

(→ consequences for logical inference ?)

What We Have Learned

PDT-MR

- **theory:**

meaning as structured by the particular language

→ How different for various language?

→ consequence: too close to the text

- **data annotation:**

refined criteria how to annotate

many "running text" examples

stress on consistency of annotation

(→ consequences for ML)

- **"technology":**

massive consistency checking

well-defined data format

formal validation

many tools (editing, visualization)

UMR

- **theory:**

meaning representation as a basis for logical inference

(BUT not much investigated so far)

→ Should be language independent (theory)!

→ consequence: broad interpretation

- **data annotation:**

vague description

small number of examples (to illustrate the theory)

interest in the annotator's understanding

(→ consequences for logical inference ?)

- **"technology":**

NO consistency checking

NO formal specification

NO data validation

NO usable tools

Future Work

- Refining the conversion of illustrated phenomena
 - focus on abstract predicates and rolesets (language-independent predicates)
 - nouns/adjectives to predicative verbs
- PDT-MR grammemes to UMR attributes
 - tense, modality, gender, animateness, negation, degree, aspect (not in UMR for the time being), ...
- Named Entities, their anchoring in Wikidata
- Structured data – addresses, sport scores, weather forecast, tables,
(whatever appears in texts)
- Czech/Latin evaluation data !!!



Thank you for your attention!

Questions?



Supported by the LUSyD project (GAČR, no. 20-16819X) and the LINDAT/CLARIAH-CZ project (MŠMT, no. LM2023062); partially supported by CUNI (GAUK, project no. 104924, and SVV, project no. 260 698).

