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

(based on ITAT 2024 paper, slightly adapted for UMR meeting at Brandeis Univ. by J. Hajič)

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).

Two Meaning Representations at a Glance

PDT-TR

- theory: Functional Generative Description
  (esp. Sgall et al, 1967; 1986; Hajičová, 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

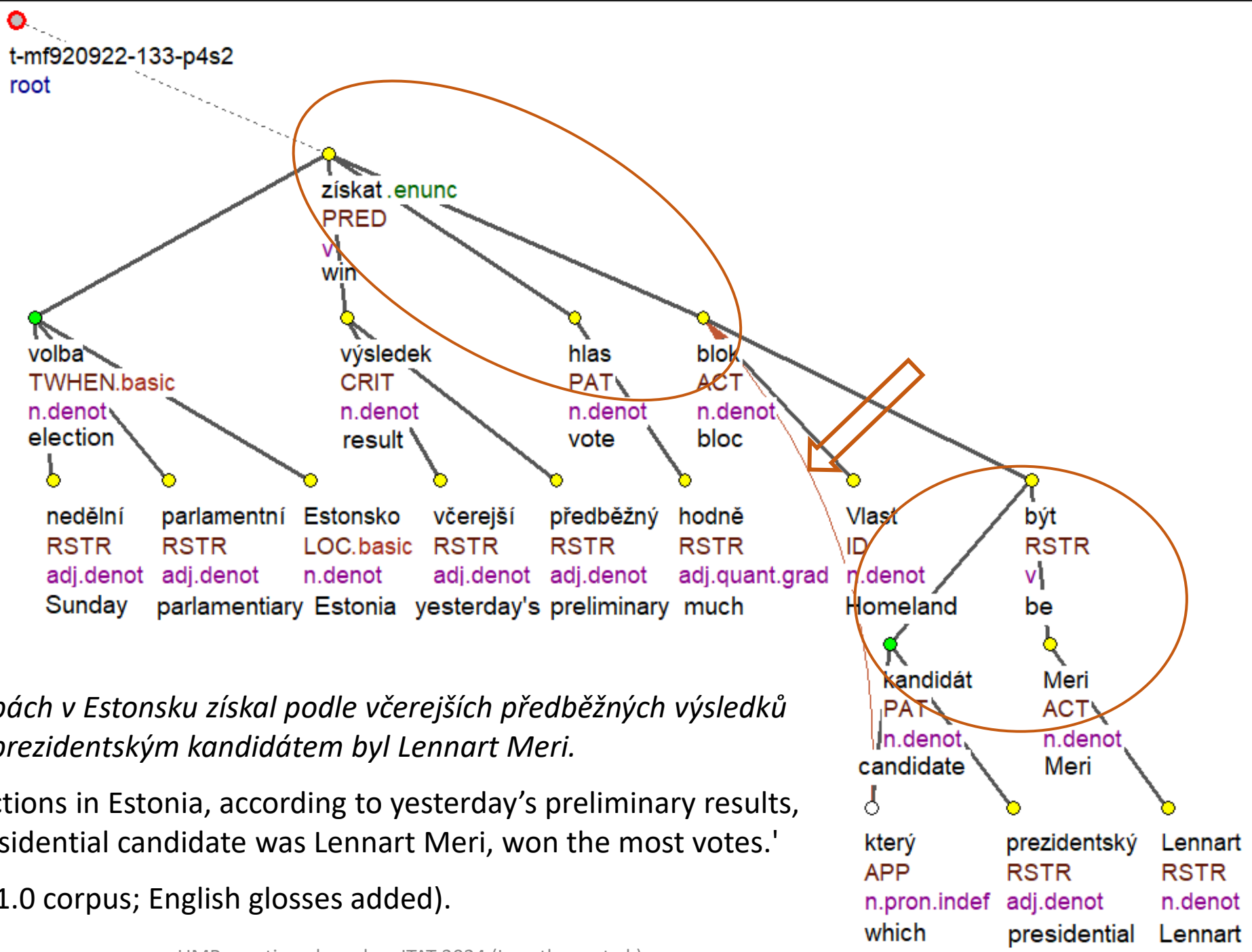
focus on **meaning as structured** by **the given language** more-or-less **directly reflects the text**

UMR

- semantics, abstracting away from syntax
  (esp. van Gysel et al, 2021; 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-TR

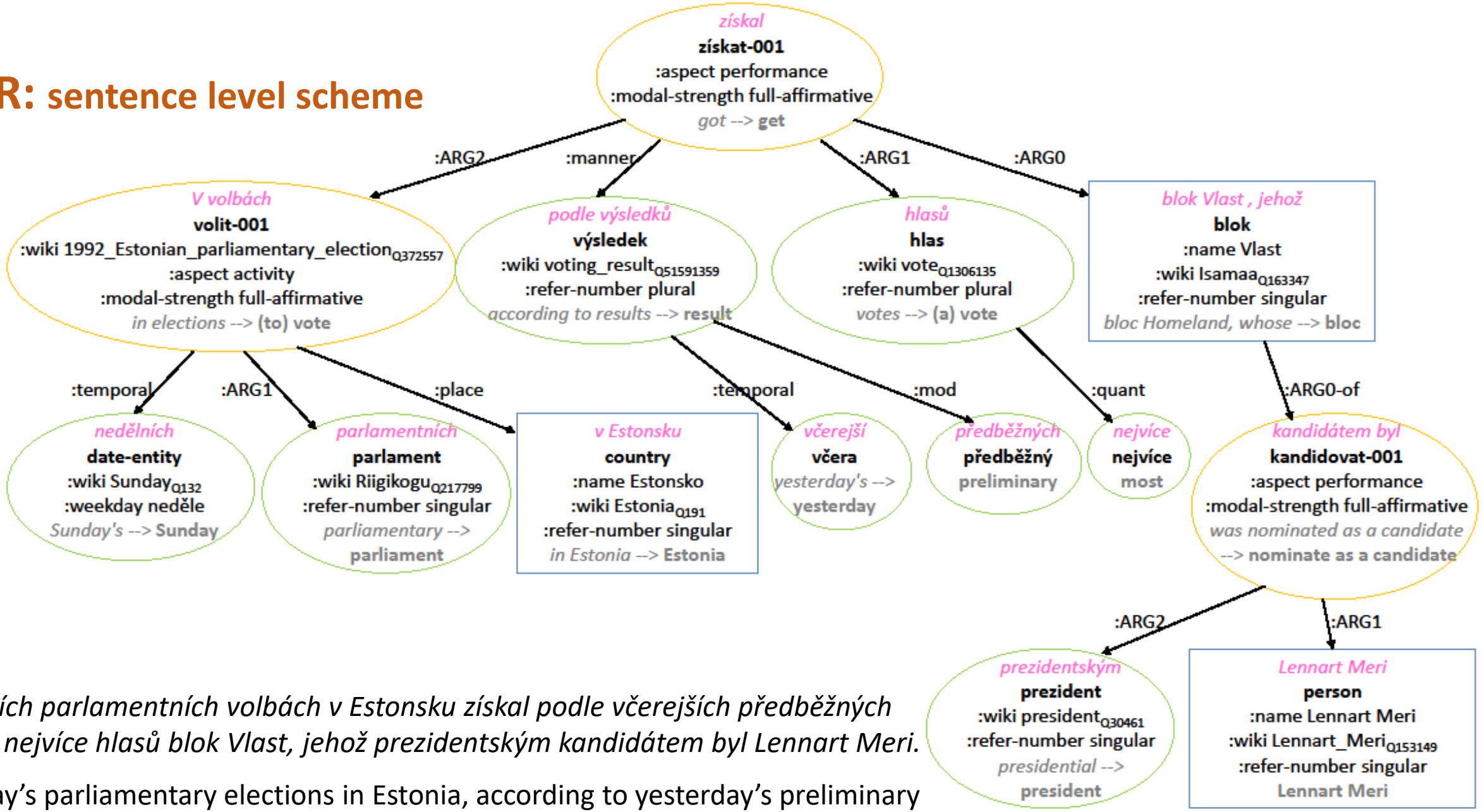


V nedělních parlamentních volbách v Estonsku získal podle včerejších předběžných výsledků nejvíce hlasů blok Vlast, jehož prezidentským kandidátem byl Lennart Meri.

'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.'

(borrowed from the PDiT-EDA 1.0 corpus; English glosses added).

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-TR 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. Coreference

coreference ≈ 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** → text coherence

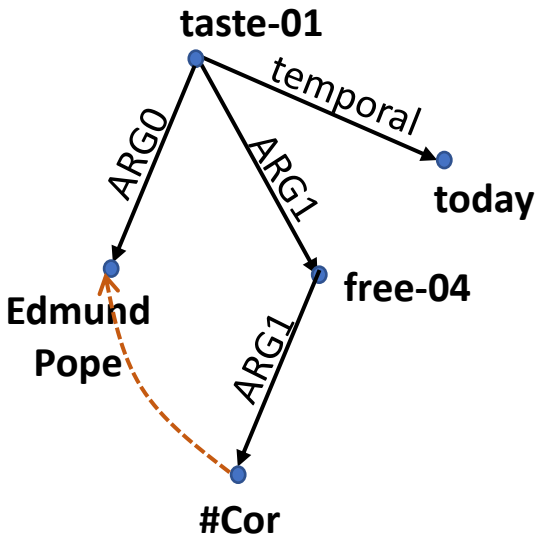
Mary lives in Prague. She likes ice-cream. The girl decided Ø to go for a trip.

antecedent **anaphor**

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

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

Coreference of 2 nodes in PDT-TR



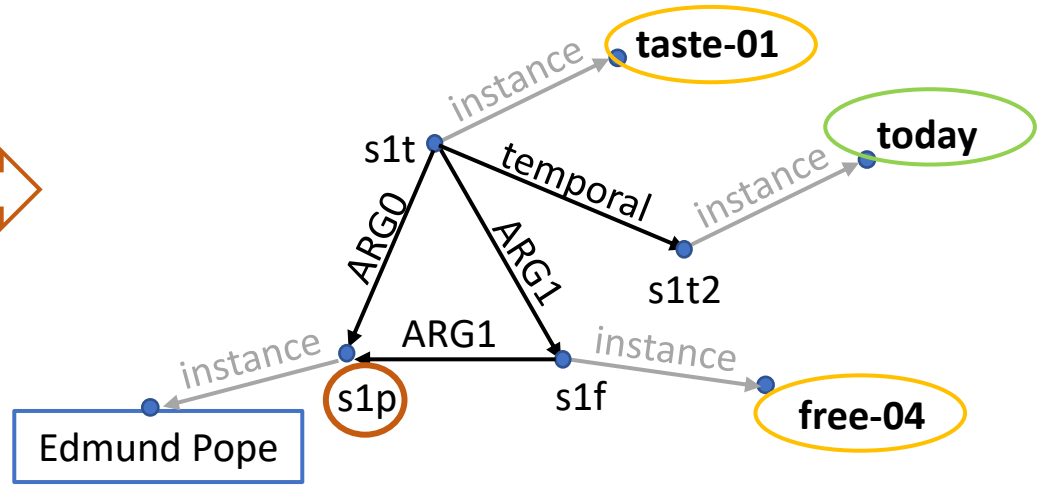
Edmund Pope tasted freedom today.

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

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

```
(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))
```

Concept of re-entrancy in UMR

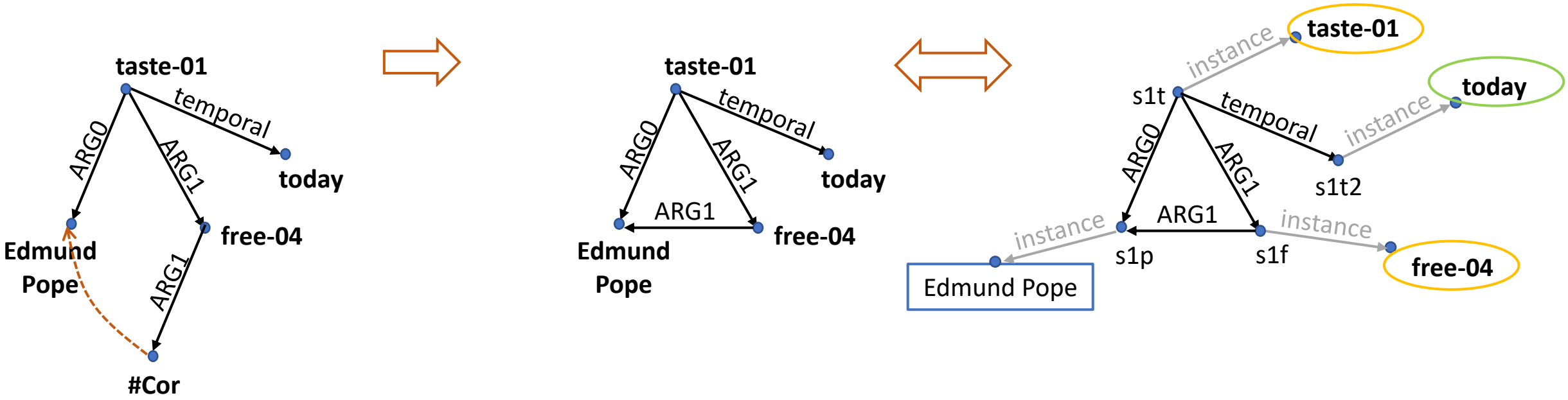


Edmund Pope tasted freedom today.

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

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

Conversion: Merging 2 nodes in PDT

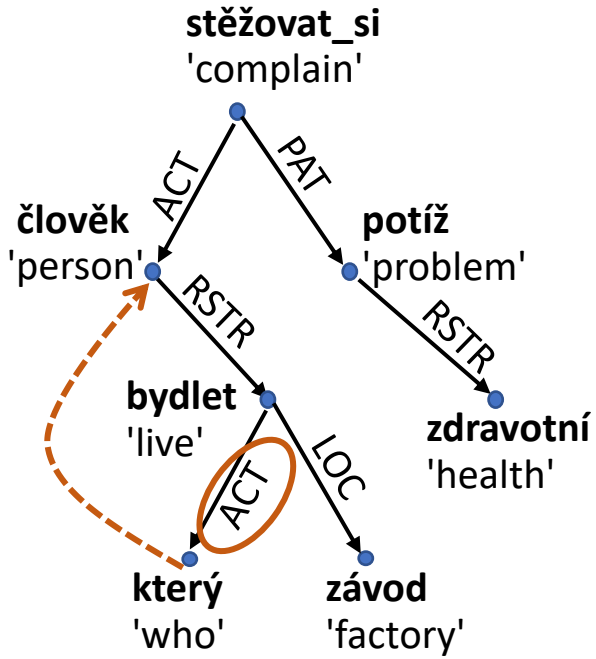


Edmund Pope tasted freedom today.

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

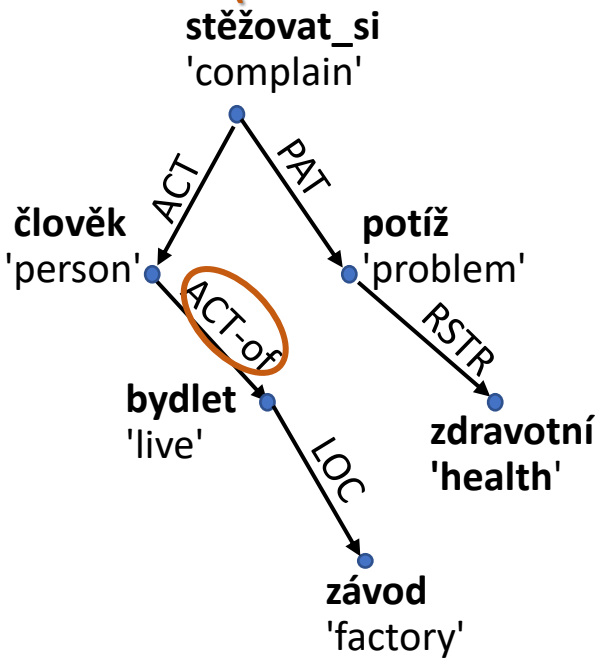
Ib. PDT-TR Coreference → UMR Inverse Role

Coreference of 2 nodes in PDT-TR



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-TR Coreference → UMR Pairing

Inter-sentence coreference relation

PDT-TR

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

UMR

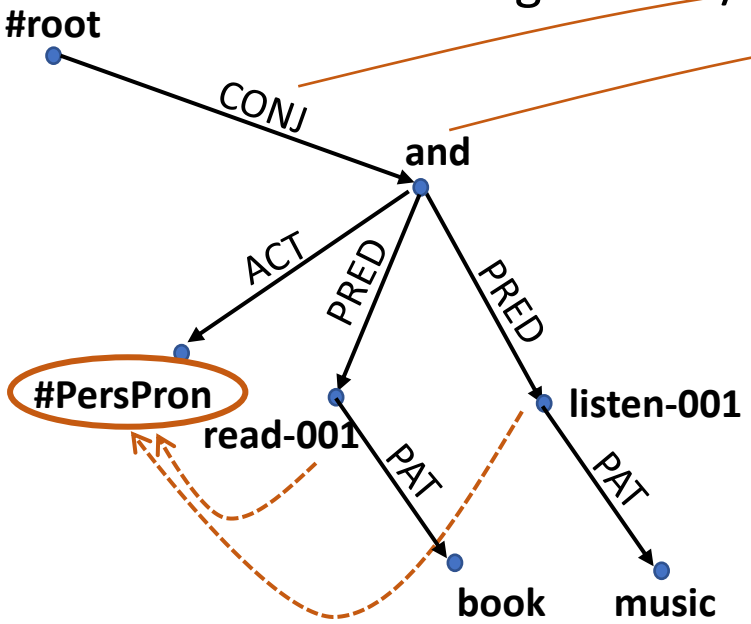
- lists pairs of coreferring concepts
  - ID of both concepts
    - event or entity … entities
    - identity or subset … identity

```
(s5s0 / sentence
 :coref ((s3c :same-entity s5c)
         (s3p3 :same-entity s5p)
         (s3v :same-event s5v)))
```

Id. Coordination

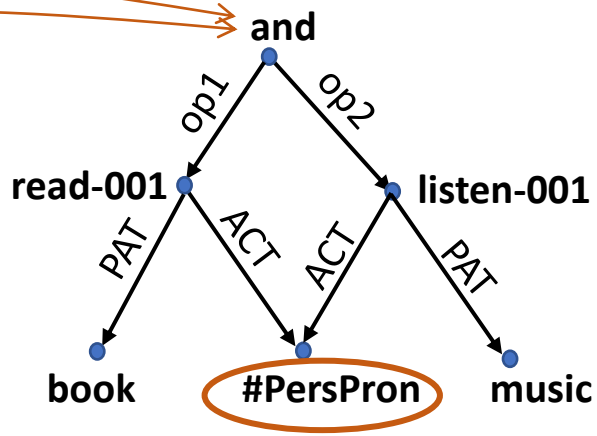
PDT-TR

- 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

PDT-TR

- verbs ≈ events and states → event annotation
- other nodes → entities or keywords

  with some degree of abstraction

  e.g., *matčin* 'mother's' → *matka* 'mother' + possesive

  "normalization", e.g., *jehož* → *který* 'who'
- refinement: lack of information

  even for most systematic changes

  e.g., *bojování* 'fighting' → *bojovat* '(to) fight'

  (*příjezd*) *přijíždění* 'coming' → *přijíždět* '(to) come'

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, no testable criteria
- skewed towards English (e.g., statives)
- big impact on annotation
  - modal, temporal, aspectual for events

- fuzzy boundary btw. entities and events
- big space for different interpretations
- intuitive decisions

III. Graph labeling

PDT-TR

arguments:

- PDT-Vallex valency lexicon (Hajič et al., 2003, Urešová et al., 2021))
  - verbs, nouns (adjectives)
  - elaborated valency theory
  - 5 "arguments": ACT, PAT, ADDR, ORIG, EFF

UMR

arguments:

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

partial verb-specific mapping
~ 43% of PDT-Vallex labels (out of 42,116) (Hajič et al, 2024)
default mapping for the rest of verb senses
most frequent argument mappings

adjuncts:

default mapping based on their semantics
further refined where necessary

What We Have Learned

PDT-TR

- **theory**:
  meaning **as structured** by **the particular language**
  THUS: too close to the text?
  → How different for various languages?

- **data annotation:**
  refined criteria how to annotate
  many "running text" examples
  stress on consistency of annotation
  (→ consequences for ML)
- **"LR technology":**
  massive consistency checking
  well-defined data format
  formal validation
  many tools (editing, visualization)

UMR

- **theory:**
  meaning representation **as language independent**
  THUS: broad interpretation
  → should serve **as a basis for logical inference**
  BUT not much investigated so far
- **data annotation:**
  vague description
  small number of examples (to illustrate the theory)
  interest in the annotator's understanding
  (→ consequences for logical inference ?)
- **"LR technology":**
  Weaker consistency checking, data validation

Future Work

- Refining the conversion of illustrated phenomena
  - focus on abstract predicates and rolesets (language-independent predicates)
  - nouns/adjectives to predicative verbs
- PDT-TR grammatemes 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 !!!

References

- Bonn, J. et al (2023): Uniform meaning representation, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.
- Uniform Meaning Representation (UMR) 0.9 Specification (dated August 8, 2022)
- van Gysel, J. et al. (2021): Designing a uniform meaning representation for natural language processing, *KI - Künstliche Intelligenz*, 35, p. 343-360.
- Hajič J. et al. (2003): PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation, in: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of Mathematical Modeling in Physics, Engineering and Cognitive Sciences, Vaxjo University Press, Vaxjo, Sweden, p. 57-68.
- Hajič, J. et al. (2020): *Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)*, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia (project URL)
- Hajič, J. et al. (2024): Mapping Czech Verbal Valency to PropBank Argument Labels, in: *Proceedings of the Fifth International Workshop on Designing Meaning Representations (DMR 2024), ELRA and ICCL*, Torino, Italia, p. 88-100.
- Hajičová, E. (2020): Dependency-Based Underlying-Structure Tagging of a Very Large Czech Corpus, *Special Issue of TAL Journal, Grammaires De Dépendence / Dependency Grammars*, p. 57-78.
- Lopatková, M. et al. (2024): Towards a Conversion of the Prague Dependency Treebank Data to the Uniform Meaning Representation. in: Proceedings of the 24th Conference Information Technologies – Applications and Theory (ITAT 2024), CEUR-WS.org, Košice, Slovakia, p. 62-76.
- Palmer, M. and Gildea, D. and Kingsbury, P. (2005): The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 31, p. 71-106.
- Pradhan, S. et al. (2022): PropBank comes of age—larger, smarter, and more diverse, in: *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, ACL, Seattle, Washington, p. 278-288.
- Sgall, P. (1967): *Generativní popis jazyka a česká deklinace* (Generative Description of a Language and the Czech Declension), Academia, Praha.
- Sgall, P. and Hajičová, E. and Panevová, J. (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Reidel, Dordrecht.
- Synková, P. et al. (2022): *Prague Discourse Treebank 3.0 (PDiT 3.0)*, LINDAT/CLARIAHCZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.
- Urešová, Z. et al. (2021): *PDT-Vallex: Czech Valency lexicon linked to treebanks 4.0 (PDT-Vallex 4.0)*, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

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).