LLMs and Symbolic Meaning Representations

Perspectives and Questions

Jan Hajič

hajic@ufal.mff.cuni.cz

June 26, 2025

Outline

- Generative Large Language Models
  - … And the OpenEuroLLM project
- Meaning Representations
  - Latest developments: PDT, UMR and beyond
- How are LLMs related to Meaning Representations
  - Basic Research Questions
- Conclusions

LLMs

What is a (Generative) Large Language Model

- Known to the public primarily as conversational LLM (e.g. ChatGPT, Llama-3.3-70B-Instruct)
- Technology
  - Deep Neural Networks
  - Trained from data (texts) – Machine Learning
  - Basic function: generate next word (segment, token) based on (long) sequence of previous words (tokens)
    - In interactive systems: start with a user „prompt“
    - Can be up to a million words (in some LLM systems)

What is a (base, foundation(al)) LLM?

- Model trained on running text only
  - i.e., not interactive
    (i.e., cannot answer questions [well])
  - Can be monolingual, multilingual, include (source) code
  - Can be multimodal (w/suitably encoded images, video, etc.)
- It is a basis for applications
  - Interactive (chatbot, conversational) LLM is created by
    - fine-tuning, continuous pre-training
    - human interaction – annotated data, relevance rating, etc.

How large is a Large Language Model?

- Model size is specified as
  - **Number of parameters** (weights), in millions (M) or [U.S.] billions (B)
    - Weight is a „real“ number in certain precision (from 32 down to 1.58 bit)
      - From that, **byte size** can be calculated: (1B weights à 8 bit = 1 GB)
- Known model sizes (open-weight models)
  - Llama 3.1: 405 B parameters (META)
  - Llama 3.3: 70 B parameters
    - **Quantized** (lower precision than original) e.g. to 6 bits: 53 GB size)
    - For inference („runtime“): 1 or more GPU cards
      - Context size matters: takes a large proportion of GPU card’s memory
- Model training:
  - Number of parameters fixed (in the standard setting)
  - Different **data (text) sizes** (in words/tokens: **tokenization** very important)
    - Llama 3.1: 15 T (trillion) tokens

OpenEuroLLM

Our goal:

Open Multilingual European Generative Foundational LLM

- Open Source (in full)
  including fully inspectable data
- 37+ languages
  EU + associated (+ business)
- High-quality
  standard and native benchmarks
- Compliant with EU regulations

OPEN EURO LLM

OpenEuroLLM PROJECT PARTNERS

Charles University

AMD SILO AI

alt-edic ELDA

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ellis INSTITUTE TÜBINGEN

Fraunhofer IAIS

JÜLICH Forschungszentrum

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF HELSINKI

UNIVERSITAS OSLOENSIS MDCCCXI

UNIVERSITY OF TURKU

ALEPH ALPHA

AI SWEDEN

ellamind

Lighton

prompsit

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

OpenEuroLLM PROJECT PARTNERS

Charles University

AMD SILO AI

alt-edic ELDA

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ellis INSTITUTE TÜBINGEN

Fraunhofer IAIS

JÜLICH Forschungszentrum

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF HELSINKI

UNIVERSITAS OSLOENSIS MDCCCXI

UNIVERSITY OF TURKU

ALEPH ALPHA

AI SWEDEN

ellamind

Lighton

prompsit

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

OpenEuroLLM PROJECT PARTNERS

Charles University

AMD SILO AI

alt-edic ELDA

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ellis INSTITUTE TÜBINGEN

Fraunhofer IAIS

JÜLICH Forschungszentrum

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF HELSINKI

UNIVERSITAS OSLOENSIS MDCCCXI

UNIVERSITY OF TURKU

ALEPH ALPHA

AI SWEDEN

ellamind

Lighton

prompsit

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

OpenEuroLLM PROJECT PARTNERS

Charles University

AMD SILO AI

alt-edic EL DA

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ellis INSTITUTE TÜBINGEN

Fraunhofer IAIS

JÜLICH Forschungszentrum

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF HELSINKI

UNIVERSITAS OSLOENSIS MDCCCXI

UNIVERSITY OF TURKU

ALEPH ALPHA

AI SWEDEN

ellamind

Lighton

prompsit

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

Wider context

- Programme: Digital Europe (25/50% co-funding), 3 years
- Set of AI-06 calls (projects started Jan-Mar 2025):
  - Two large projects: OpenEuroLLM and LLMs4EU
  - Coordination (ALT-EDIC4EU), total **~80 mil. EUR + HPC**
  - Strong cooperation (Deploy AI, TAILOR, TrustLLM, HPLT, …)
- Goals
  - Develop open LLMs, including conversational
  - Adapt them to applications in all areas, from commerce to e-government and education
  - Contribute to EU's digital sovereignty

OPEN EURO LLM

Collaborations, Compute, and Data

- Open source community members
  - Experts on LLMs (incl. from non-EU ones, informal)
    - LAION, open-sci, Common Crawl, …
  - Experts on legal issues
- Computing power: 5 EuroHPC centers on board (project partners)
  - Now: 3m GPU hours on Leonardo (CINECA), 1.5m GPU on LUMI-G (CSC)
    - For generating synthetic data, MoE experiments, multilinguality
- Data (w/CommonCrawl, Internet Archive, OpenWebSearch)
  - From previous projects (HPLT) and other sources – 37+ languages!
    - Cleaning, language ID, topic detection essential

OPEN EURO LLM

Evaluation and Benchmarking

- For initial experiments:
  - Standard benchmarks for base models
- Project longer-term goal
  - Benchmarks for all languages in native form
    - i.e., manually translated or inspected, incl. contents
- Continuous evaluation
- Tests for evaluation data purity
  - I.e., not used in training/SFT/…
- Models released based on evaluation results

OPEN EURO LLM

Meaning Representation(s)

Meaning Representation is…

- Symbolic system for describing “meaning”…
  - Usually in the form of a graph
    - Nodes: units of meaning
    - Edges: relation between such units
    - Attributes – additional information at nodes and/or edges
- … related to the (surface, running) text that expressed that
  - Aligned to text
    - At a sentence level only
    - With word/token granularity (“anchored”)

There are many Meaning Representations…

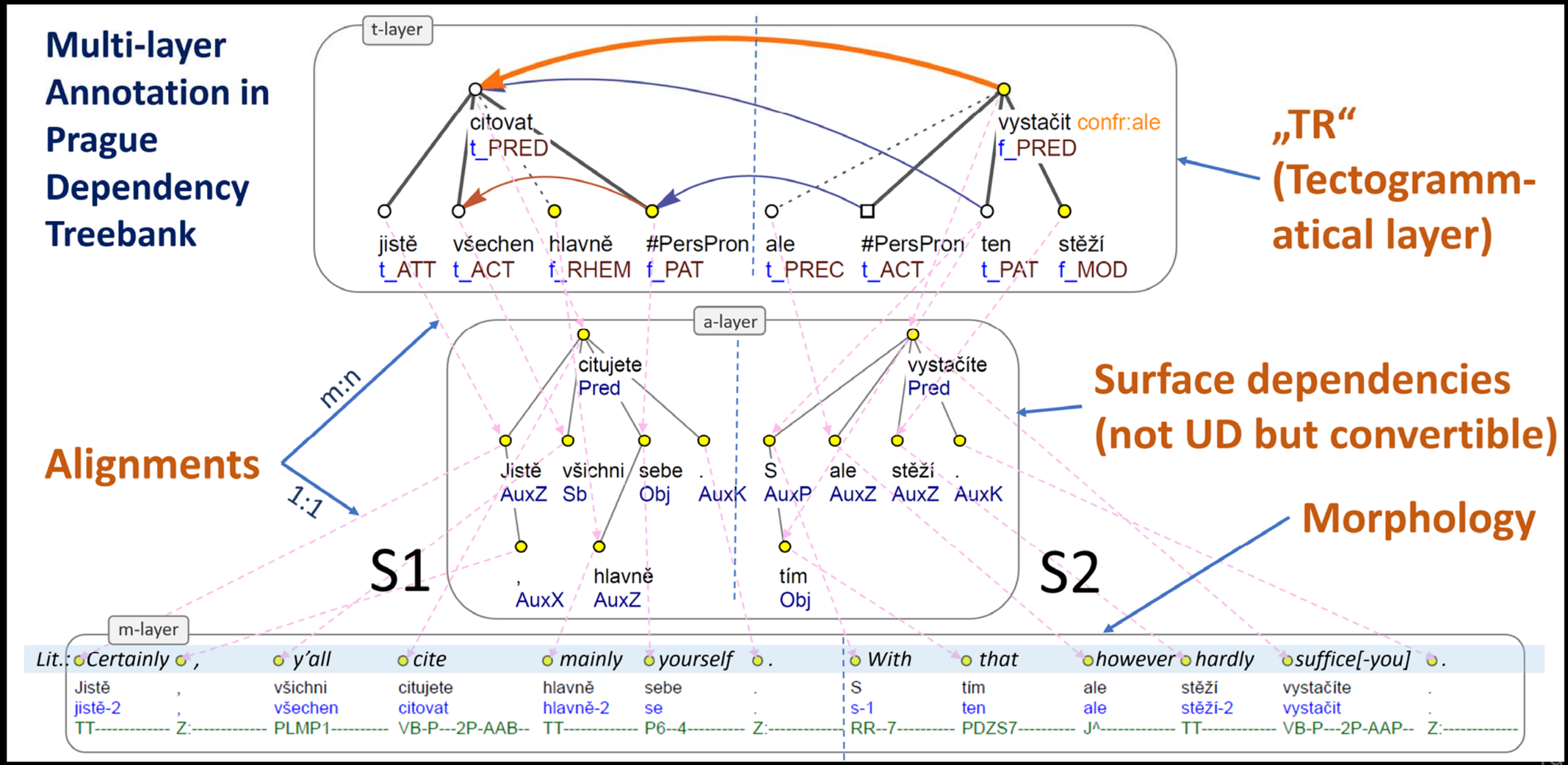
- Meaning representations vary along many dimensions
  - How meaning is connected to text
    - Anchoring, alignment, multi-layer vs. text-span only
  - Relationship to logical and/or executable form
  - Mapping to Lexicons/Ontologies
    - General, task-oriented
  - Relationship to discourse and discourse-like phenomena
    - Including co-reference, information structure, scoping, etc.
  - … and other inter-sentential relations

There are many Meaning Representations…

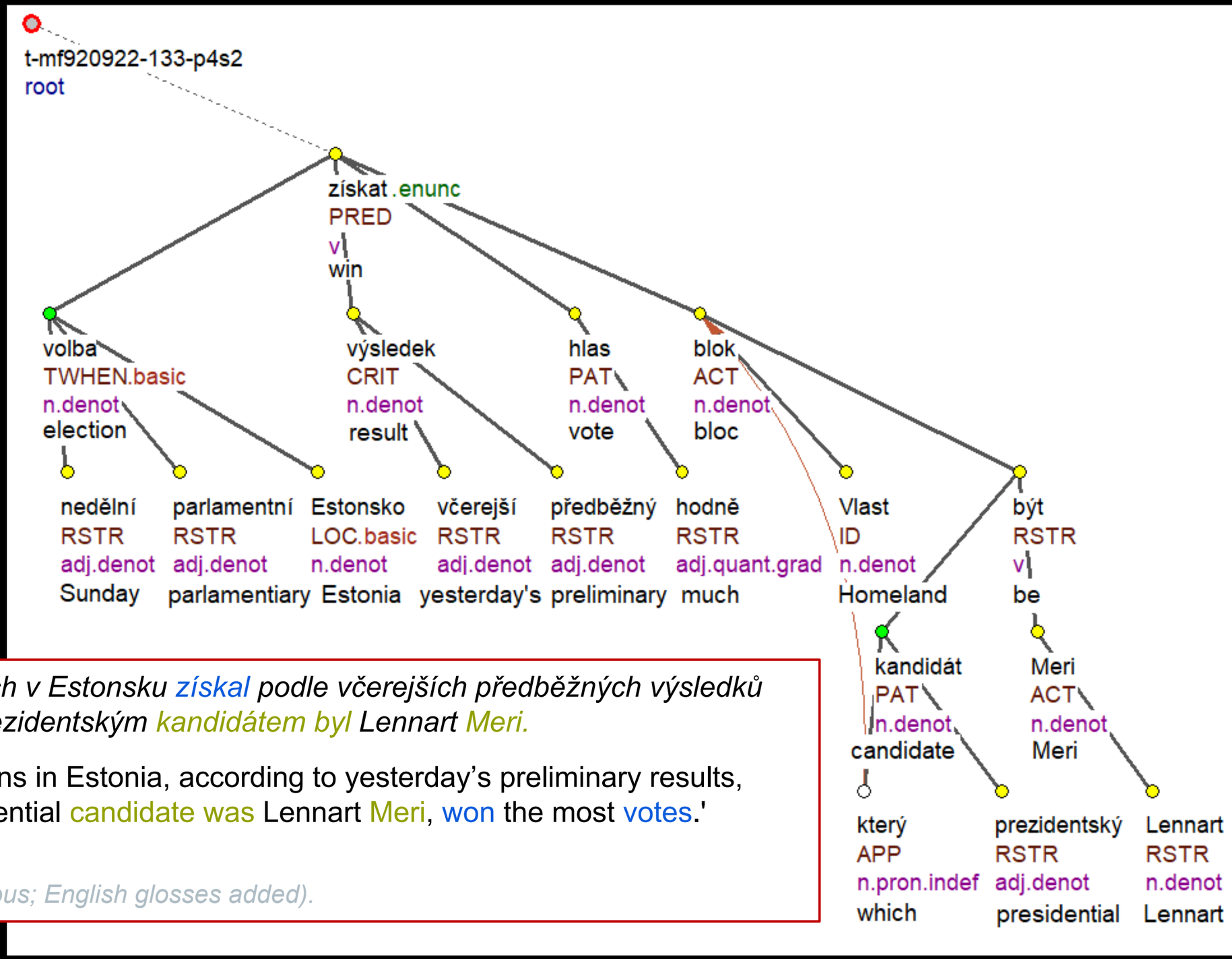
- Me

	Alignment	Logical Scoping & Interpretation	Ontologies and Task-Specific	Discourse-Level
DRS (Groningen / Parallel)	Compositional /Anchored	Scoped representation (boxes)	Rich predicates (WordNet), general roles	Can handle referents, connectives
MRS	Compositional /Anchored	Underspecified scoped representation	Simple predicates, general roles	n/a
UCCA	Anchored	Not really scoped	Simple predicates, general roles	Some implicit roles
Prague (PDT) Tectogrammatical Representation Layer	Anchored	Not really scoped with exceptions (negation)	Rich predicates, semi-lexicalized roles	Rich multi-sentence conference, discourse
AMR	Unanchored except ZH	Not really scoped yet	Rich predicates, lexicalized roles	Rich multi-sentence coreference
UMR	Anchored	Scoped (quantifiers, negation)	Rich predicates, lexicalized roles	Rich multi-sentence conference, discourse

Prague Dependency Treebank (PDT) in a nutshell



PDT-TR example

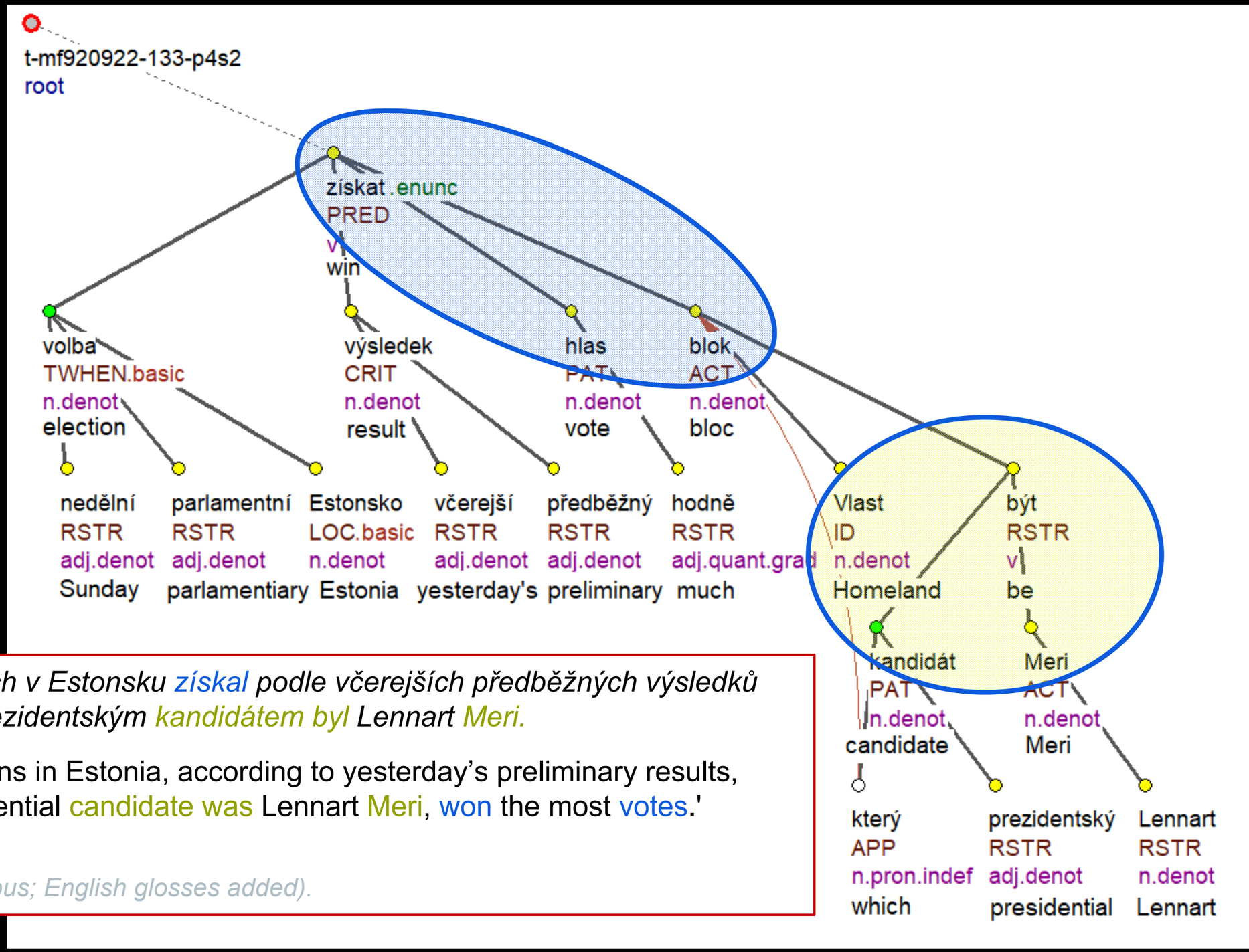


V nedělních parlamentních volbách v Estonsku získal podle včerejších předběžných výsledků nejvíce hlasů blok Vlast, jehož prezidentským kandidátem byl Lennart Meri.

'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).

PDT-TR example



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

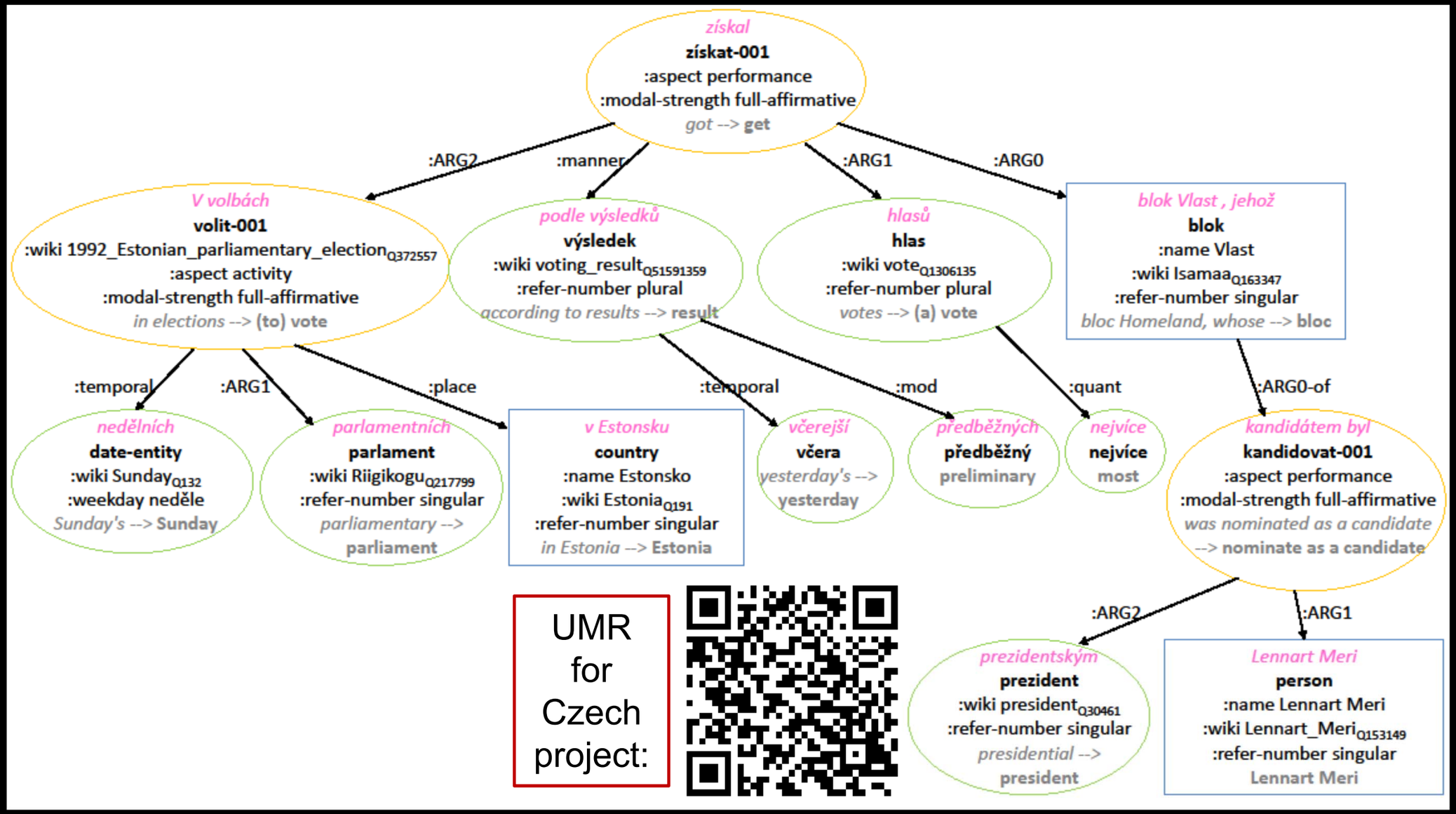
'In Sunday's parliamentary elections in Estonia, according to yesterday's preliminary results, the Homeland bloc, whose presidential candidate was Lennart Meri, won the most votes.'

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

Uniform Meaning Representation (UMR)

UMR project

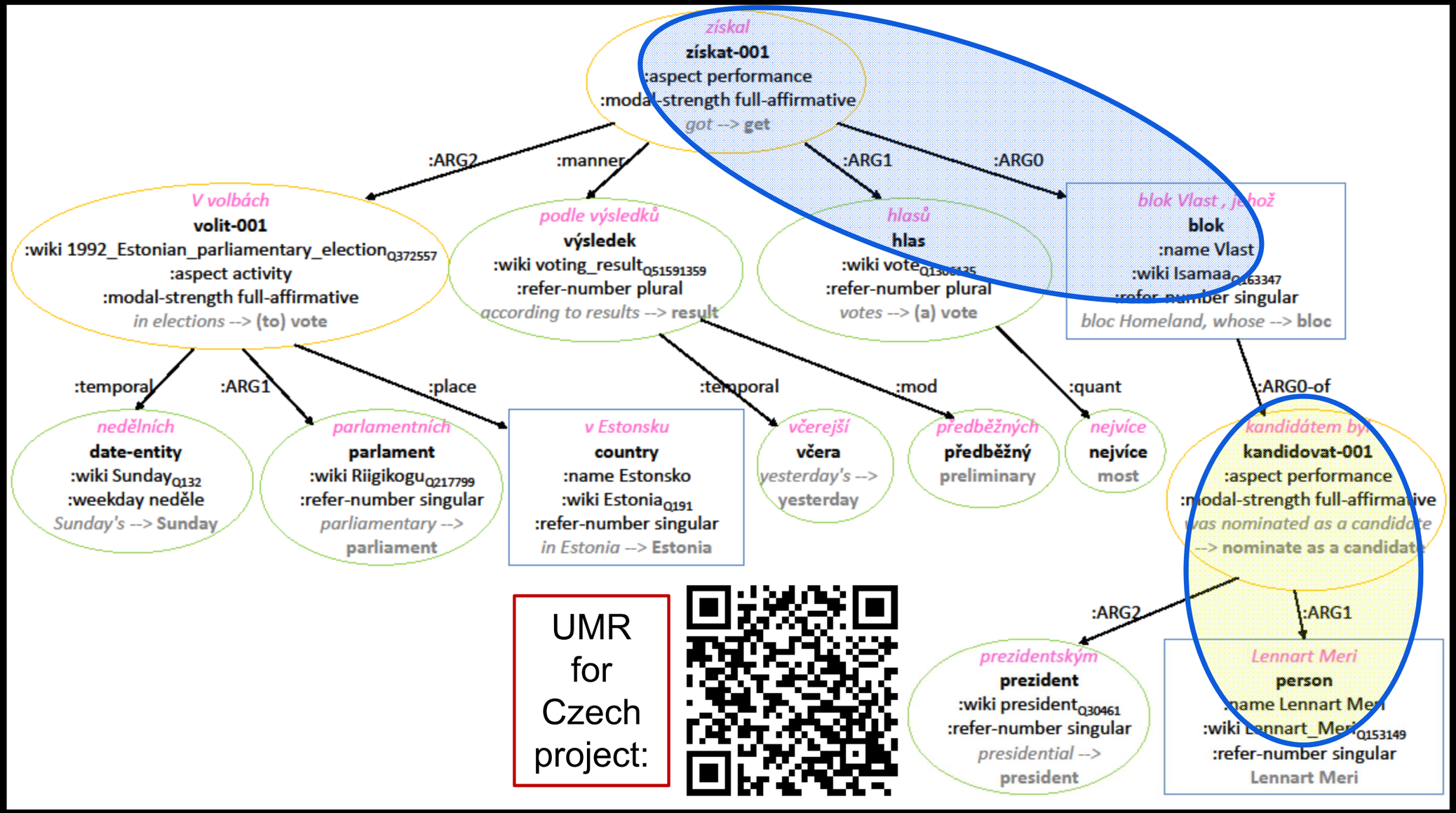
UMR 2.0 data



Uniform Meaning Representation (UMR)

UMR project

UMR 2.0 data

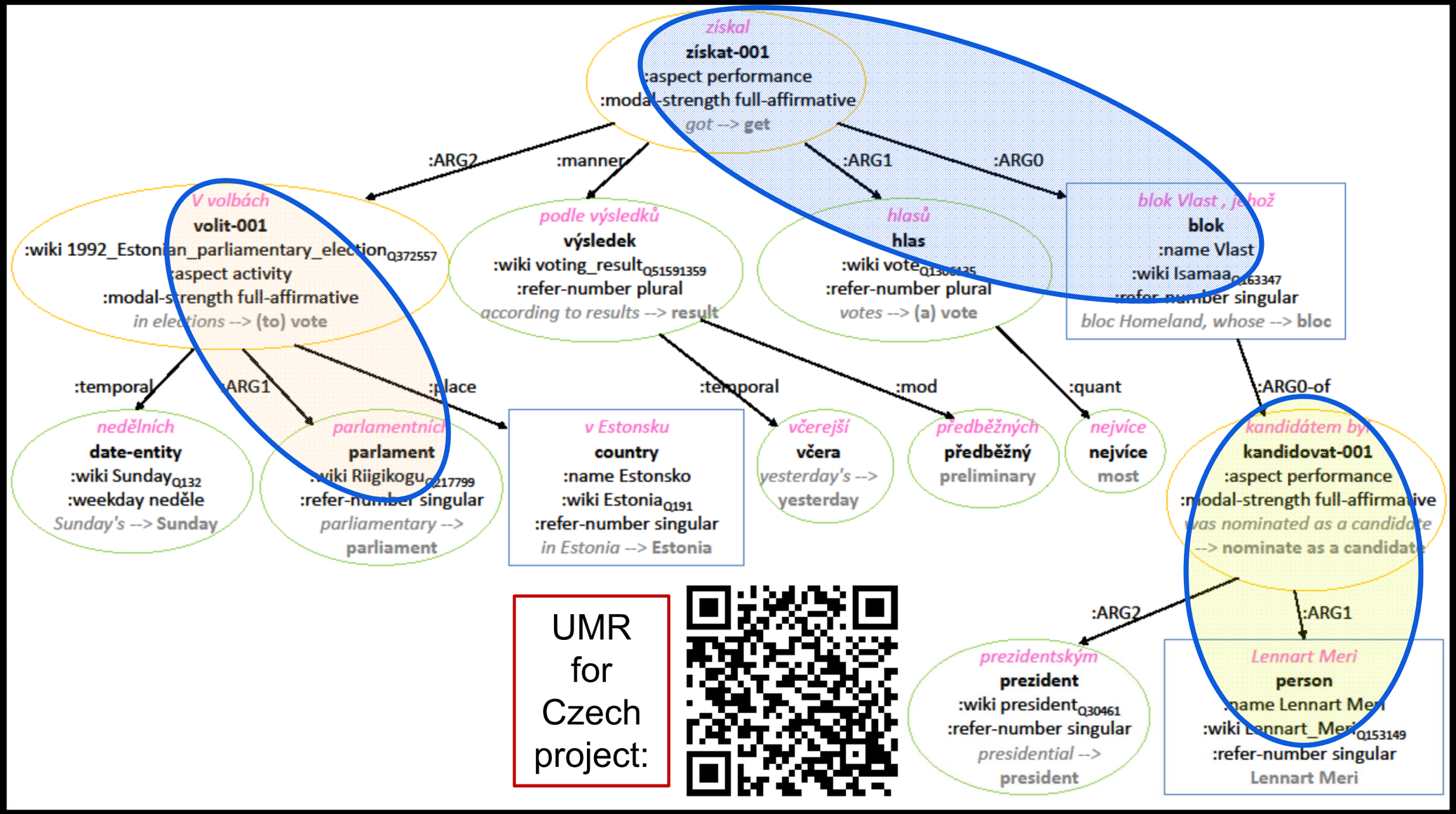


UMR for Czech project:

Uniform Meaning Representation (UMR)

UMR project

UMR 2.0 data



UMR for Czech project:

Beyond Predicate-Argument Structure

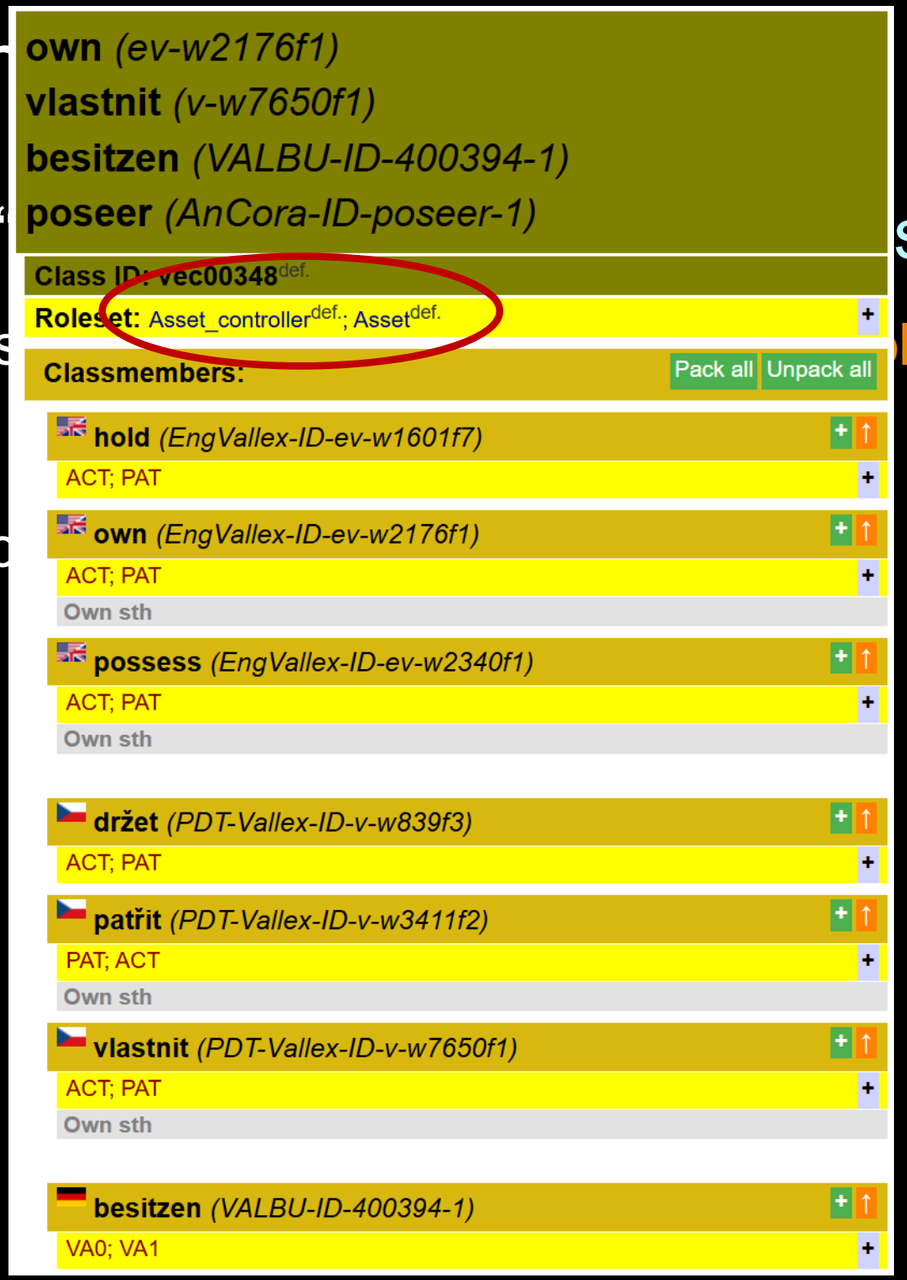
- Current Predicate-Argument lexicons
  - PropBank (AMR/UMR), (PDT-)Vallex (PDT and similar)
- Known issues
  - Argument labels not always “semantic”
    - ?semantics of PAT in PDT, Arg2 in PropBank
  - Synonymy?
    - Is there a difference, in a particular context, among “inform”, “announce” and “tell”? If yes, in what exactly?
    - Not having it makes inference more difficult
  - Hierarchy (IS-A, or general/specific relation)?
    - More general terms often used but actual meaning is more specific

SynSemClass ontology of eventive types

- Class ~ eventive concept (“to have something in possession”)
  - Class members: words (senses) with argument structure and roles
    - A.k.a. synonyms
    - In multiple languages (concept is “language independent”)

SynSemClass ontology of ever

- Class ~ eventive concept (“ssion”)
  - Class members: words (sens les
    - A.k.a. synonyms
    - In multiple languages (c

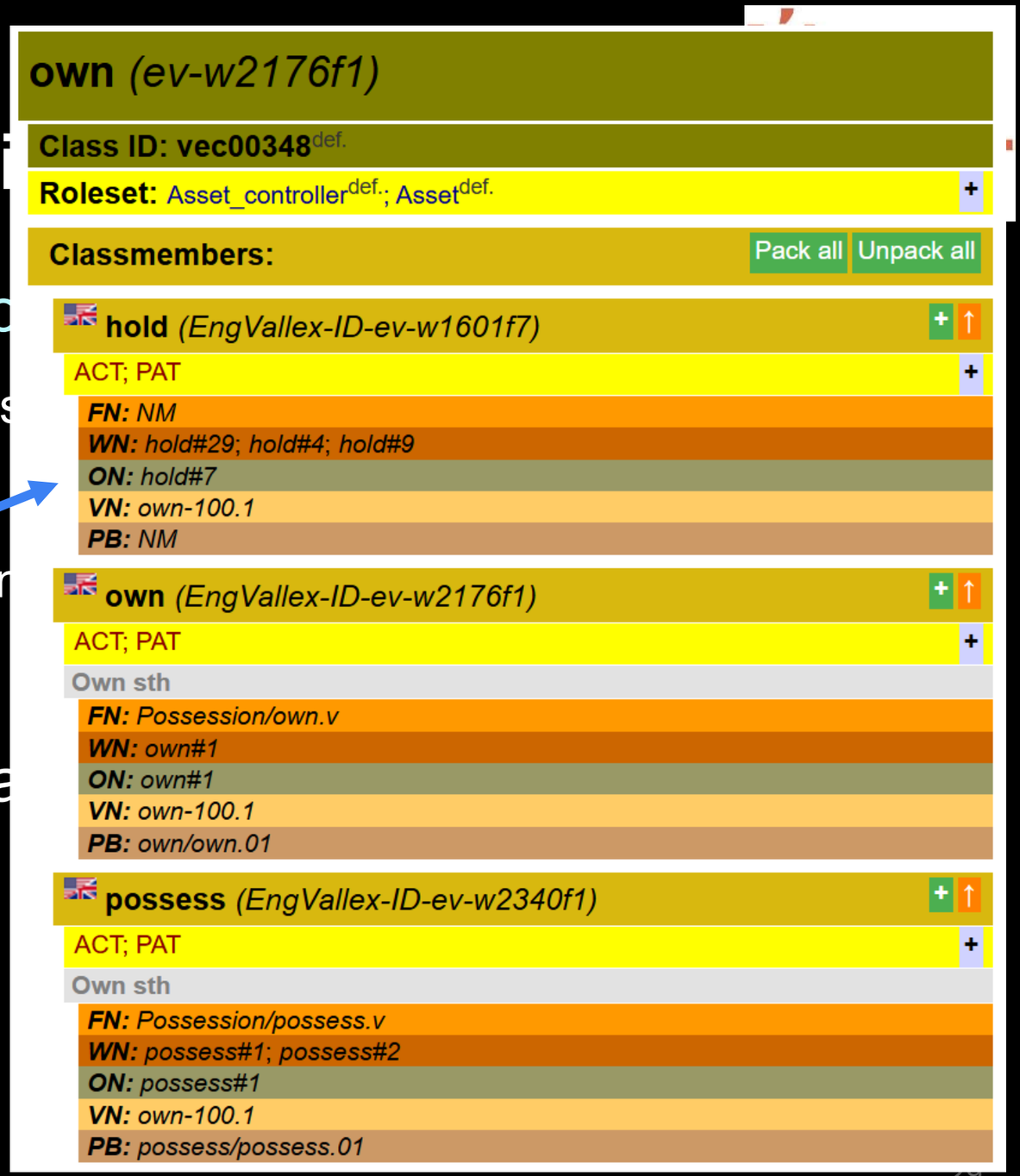


SynSemClass ontology of eventive types

- Class ~ eventive concept (“to have something in possession”)
  - Class members: words (senses) with argument structure and roles
    - A.k.a. synonyms
    - In multiple languages (concept is “language independent”)
- Definitions, examples
- Links to existing lexical-semantic resources
  - FrameNet, WordNet, VerbNet, OntoNotes and similar in other languages

SynSemClass ontology of eventi

- Class ~ eventive concept (“to
  - Class members: words (senses
    - A.k.a. synonyms
    - In multiple languages (cor
- Definitions, examples
- Links to existing lexical-sema
  - FrameNet, WordNet, VerbNet,

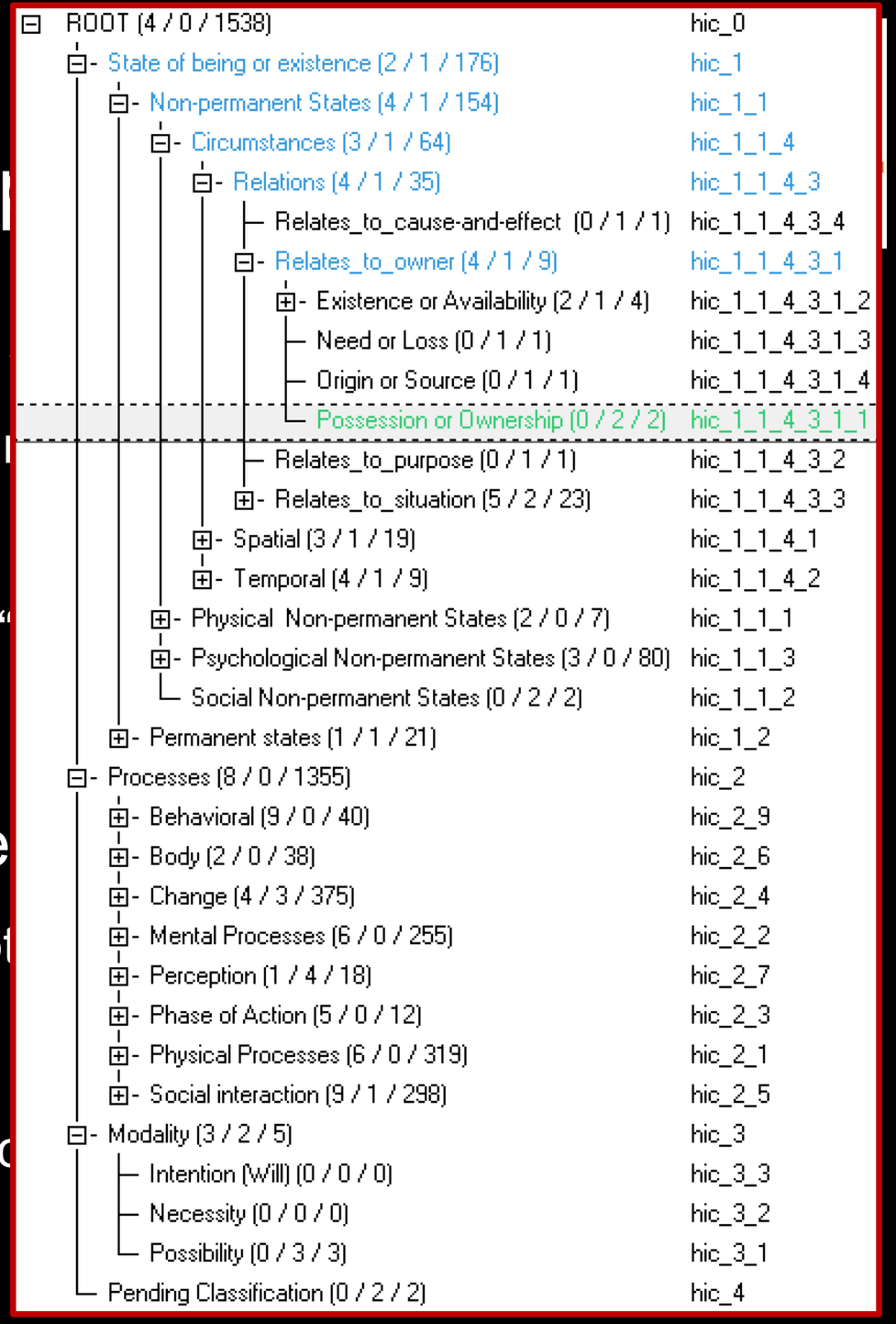


SynSemClass ontology of eventive types

- Class ~ eventive concept (“to have something in possession”)
  - Class members: words (senses) with argument structure and roles
    - A.k.a. synonyms
    - In multiple languages (concept is “language independent”)
- Definitions, examples
- Links to existing lexical-semantic resources
  - FrameNet, WordNet, VerbNet, OntoNotes and similar in other languages
- Hierarchy in classes
  - More general class / more specialized classes

SynSemClass ontology of eventive typ

- Class ~ eventive concept (“to have
  - Class members: words (senses) with ar
    - A.k.a. synonyms
    - In multiple languages (concept is “
- Definitions, examples
- Links to existing lexical-semantic re
  - FrameNet, WordNet, VerbNet, OntoNot
- Hierarchy in classes
  - More general class / more specialized c



LLMs
vs. (and?)
Meaning Representations

(Computational) Linguistics, NLP and LLMs

- Is NLP “solved” (by LLMs)?
  - No, not really, not yet – but there is great progress
  - Still hard problems remain
    - reasoning, explainability, interpretability; low-resourced languages
  - Can Semantic Representations be used with(in) LLMs to improve NLP?
- Is (Computational) Linguistics ”solved” (by LLMs)?
  - Not at all
  - We can ask LLMs questions about language
    - Answers come from the texts they were trained on …
      - … not on their “introspection”
  - Hard questions unresolved

Open questions

- Linguistics
  - Language structure (is there any…?)
  - What are symbolic representations actually telling us?
    - About morphology, syntax, meaning, …
- (Language) Learning
  - How do we learn language (mother tongue, 2nd)? Anything LLMs can teach us?
- Relation between language and the world around us
  - How is our knowledge (memory) structured? Any parallels with LLMs?
  - Relation between perception of language, vision and other senses
  - Why are we describing language as graphs or formulas?
  - What exactly is “grounding” (in perception, communication, society)?

LLMs vs. computational models

- How do they represent the language they were trained on?
- How do they generalize (in the sense humans do)? Do they?
- How do they learn about “concepts”, represent and reason over them?
- Can we learn anything from comparing humans and LLMs?
  - How can we make such comparisons?
    - Brain level, symbolic level, using logic, philosophy, …?
  - Is it fair comparison? (different learning mechanisms, language x other modalities, …)
- Should we collaborate with psycholinguists, neurolinguistics, cognitive scientists, logicians, philosophers? [yes, of course; but how exactly?]
- Where LLM biases come from and how they differ from ours?

Thank you!

Questions?

https://ufal.mff.cuni.cz/node/2540
UFAL MFF UK – UMR project

https://openeurollm.eu/

https://ufal.mff.cuni.cz/grants/lusyd
UFAL MFF UK – LUSyD project

Supported by the project OpenEuroLLM, GA No. 101195233, ALT-EDIC4EU, GA No. 101195344, Digital Europe Programme by European Commission and co-funded by the JU subprogramme of the MEYS CR and other MEYS CR and CSF programs.

Co-funded by the European Union

Open Source and Community

- Open Strategic Partnership Board (Strategic advisory role)
  - Open source community members
  - Experts on LLMs (incl. from non-EU ones)
    - Former commercial and/or open source model developers
- Experts on legal issues
- Informal cooperations
  - Data side: CommonCrawl, Internet Archive EU, OpenWebSearch
  - Open source models community
    - EuroLLM (Univ. of Edinburgh - UK, UnBabel - Portugal)
    - LAION, open-sci, …

OPEN EURO LLM

Computing facilities

- 5 EuroHPC centers on board (project partners)
  - Technical expertise - jumps start using the respective facilities
- Some compute available from previous projects
- Participation in EuroHPC calls in 2025
  - In line with project plan for the rest of 2025
- „Strategic“ allocations in the future („STEP“)
  - Using current facilities & new in AI Factories (2026/2027)
    - Just received 3m GPU hours for May-Nov. 2025 on Leonardo (CINECA)
      - For generating synthetic data
    - … and 1.5m GPU hours at CSC, on LUMI-G
      - Testing data staging, multilingual training, MoE (~scaling laws)

OPEN EURO LLM

Data for 37+ languages

- Using available Open Source data
  - HPLT 2.0 (HPLT 3.0, July-Aug 25), Fineweb2, Cultura-X, …
  - Mixtures to be experimentally determined
    - Ultimate (re)sources: CommonCrawl, Internet Archive, IA Europe
    - OpenWebSearch – negotiations ongoing
- Focus on low-resource languages for additional data
  - Incl. specific cases for very similar languages
- Additional data for
  - Fine-tuning, instruction-tuning, reasoning
    - … if necessary for benchmarking

OPEN EURO LLM