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, Malostranské náměstí 25, Prague, Czechia

Abstract

For centuries, linguists have deliberated on how to represent meaning. Recently, this inquiry has been pursued not only as an intriguing theoretical problem but also due to its practical implications for applications. Here we provide a comparison of two meaning representations rooted in two different linguistic traditions and based on different theoretical assumptions: the meaning representation used in the family of Prague Dependency Treebanks and the Uniform Meaning Representation. We discuss the possibility and limitations of an automatic "translation" between these two formalisms, focusing esp. on selected deep syntactic phenomena affecting the shape of sentence graphs. Specifically, we concentrate on predicates and their argument structures, lexicons available for both approaches, levels of abstraction, and on coreference. We believe that the mutual inspiration of both approaches can lead to a substantially deeper understanding of language semantics.

Keywords

Prague Dependency Treebank, Uniform Meaning Representation, graph representation, events and entities, predicates and their argument structure, abstract concepts, reification, coreference

1. Motivation

For centuries, linguists have deliberated on how to represent meaning. Recently, this inquiry has been pursued not only as an intriguing theoretical problem but also due to its practical implications for applications, since meaning representation can serve, for example, as an interlingua for machine translation or as a basis for knowledge representation and knowledge systems. Today, large language models (LLMs) dominate the field; however, a significant issue persists with these systems, as they tend to fabricate information unscrupulously. Therefore, a sound and reliable representation of meaning, suitable as a basis for logical inference, remains a compelling task.

1.1. Related work

Naturally, dozens and dozens of formalisms for meaning representation have been proposed in recent decades (if we limit ourselves to the most influential ones). In this contribution, we focus only on *graph-based* and *dependency-oriented* representations, typically involving tree-like structures.[1] These frameworks offer a visual and structural approach to encoding of meaning, allowing for clear delineation of relationships and hierarchical organization of concepts.

Among many available overviews presenting most significant meaning representation formalisms, let us mention at least two prominent articles [4, 5]. Žabokrtský et al. [4] give an enlightening and informative overview of 11 influential frameworks as instantiated in available treebanks. They introduce, among others, the following approaches: the Paninian framework, the Meaning-Text Theory [6, 7], the meaning representation used in the family of Prague Dependency Treebanks (here abbreviated as PDT-MR, see Sect. 2.1 for references), the Abstract Meaning Representation (AMR, see Sect. 2.2) and Enhanced Universal Dependencies [8]. After describing basic features of the chosen approaches, the authors change their perspective and show how selected language phenomena are treated across these frameworks. They focus esp. on the following features of the compared approaches: on the formal structure (whether they use (rooted) trees or more general graphs, how they deal with coordination, etc.), on the characteristics of nodes and edges (esp. types of relations encoded), and on the "depth" of their position on the deep-syntax—semantics scale (this refers, e.g., to treatment of content vs. function words or to valency issues). Further, they discuss whether the frameworks cover also semantically relevant morphological

24th Conference ITAT (Information Technologies—Applications and Theory), 2024: Drienica, Čergovské vrchy, Slovakia

*Corresponding author.
lopatkova@ufal.mff.cuni.cz (M. Lopatková)
0000-0002-3833-9611 (M. Lopatková); 0000-0001-7895-8890 (E. Fučíková); 0000-0003-3632-0594 (F. Gamba); 0009-0000-3101-573X (J. Štěpánek); 0000-0002-5791-6568 (D. Zeman); 0000-0002-7805-9649 (Š. Zikánová)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] In particular, we do not deal here with primarily logical representations such as the Minimal Recursion Semantics (MRS) [1], Discourse Representation Theory (DRT) [2], or Groningen Meaning Bank [3] as they offer representations that are rather distant from sentence structure.

categories (surprisingly, even basic categories like tense or number are mostly ignored), coreference (covered by most frameworks), and discourse relations. As a conclusion on possible convergence of the introduced frameworks, the overview recommends establishing a baseline that should be common to all meaning representations.

While [4] approaches the question of meaning representation more-or-less from the theoretical perspective, the effort of Oepen et al. [5] is oriented more practically, towards semantic parsing. The authors report on the *Shared task on Cross-Framework Meaning Representation Parsing (MRP 2020)*[2] at the Conference on Computational Natural Language Learning (CoNLL). Still, they present five different frameworks for meaning representation that use directed graphs; the PDT meaning representation (converted into the so-called Prague Tectogrammatical Graphs) and Abstract Meaning Representation (AMR, see Sect. 2.2) being among them. Given the purpose, the overview focuses primarily on the formal structure (which concepts are represented as graph nodes, which relations correspond to edges, how coordination is represented), on their potential alignment to sub-strings of surface sentence structure, and on the level of lexicalization. Instead of converging the frameworks, the project attempts to explore systems with a shared implementation that can generate representations in any framework (at least to some extent) and utilize information across individual frameworks through "cross-fertilization".

1.2. Why PDT-MR and UMR?

The goal of this contribution is to provide a more detailed comparison of two meaning representations rooted in two different linguistic traditions and based on different theoretical assumptions: the meaning representation used in the family of Prague Dependency Treebanks (PDT-MR, references in Sect. 2.1 below) and the Uniform Meaning Representation (UMR, references in Sect. 2.2 below).

The choice of the first formalism is clear from our perspective: The three most prominent PDT treebanks, namely the original PDT, PDiT (Czech texts with discourse annotation) and PCEDT (Czech portion of parallel Czech and English texts) represent the most extensive and well-developed datasets available for Czech (see Sect. 2.1 for references). Furthermore, the PDT formalism has been applied not only to Czech and English: a PDT-like annotation is available, among others, for Latin texts as well.[3]

Moreover, we are familiar with this approach and possess well-functioning processing tools (like a graph viewer and editor serving also as a powerful annotation tool, a pipeline for tokenization, tagging, lemmatization and dependency parsing, a tool processing named entities, etc.).

Based on this perspective, we find UMR as an approach with a high potential to enrich our research of language semantics, for the following reasons: First, UMR offers a more abstract representation, which is less dependent on a specific language and its structure. Second, UMR anchors concepts within a knowledge base (utilizing the English Wikipedia). Third, UMR also aims to support logical inference, which falls outside the scope of PDT-MR. Last, but not least, UMR is being used for a variety of typologically diverse languages, including Arapaho, Navajo, Kukama, and Sanapaná. This approach and its rich data may facilitate understanding some features of the Czech and Latin languages from the typological point of view.

In this contribution, we present the basic characteristics of the two approaches (Sect. 2), and then we focus on selected features affecting the shape of sentence graphs (Sect. 3), namely on their formal representation (Sect. 3.1), on the way how predicate verbs and their argument structure are captured there, including so-called abstract predicates (Sect. 3.2), and on the treatment of the coreference chains (Sect. 3.3). We conclude with a summary and short outline of the future work (Sect. 4).

2. Basic Characteristics of the Two Approaches

2.1. PDT-MR: PDT meaning representation

The Prague Dependency Meaning Representation (PDT-MR) originates primarily in the tectogrammatical layer of language description [9, 10, 11], as designed within the theoretical approach of the Functional Generative Description (FGD) [12, 13] and instantiated in the family of Prague Dependency Treebanks, esp. the Prague Dependency Treebank (PDT) [14, 15] and Prague Discourse Treebank (PDiT) for Czech [16, 17], and the parallel Prague Czech-English Dependency Treebank (PCEDT) [18, 19].

PDT-MR is a dependency-oriented complex annotation scheme covering deep syntax, with predicate-argument structure forming a core of the dependency representation. It presents also meaning-relevant morphological information (like tense, number, gender, or (deontic) modality), information structure and discourse relations, including coreference annotation. Fig. 1 exemplifies slightly simplified meaning representation following the PDT-MR guidelines.

[2] http://mrp.nlpl.eu/2020/index.php

[3] The texts annotated in the PDT style are the Index Thomisticus Treebank (ITTB) and a portion of the Latin Dependency Treebank (LDT); https://itreebank.marginalia.it/view/download.php

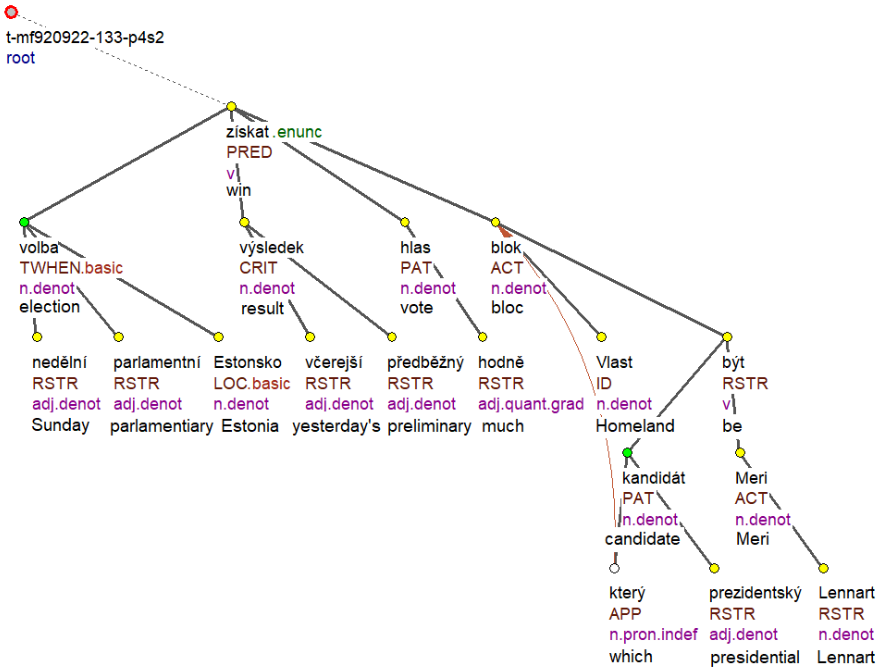


Figure 1: PDT-MR representation of the sentence *V nedělních parlamentních volbách v Estonsku získal podle včerejších předběžných výsledků nejvíce hlasů blok Vlast, jehož prezidentským kandidátem byl Lennart Meri.* 'In Sunday's parliamentary elections in Estonia, according to yesterday's preliminary 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).
Here we point out the most relevant PDT-MR principles:
(i) Only content (= autosemantic) words have their own nodes, function words (like prepositions) are represented as attributes of the relevant content words.
(ii) The valency characteristics of the predicate verb—the verb *získat* 'win; get' and its two actants ACT (`Actor/Bearer`), and PAT = (`Patient/Deep Object`)—create the core of the deep structure. The predicate verb is interlinked with the PDT-Vallex valency lexicon indicating the particular verb sense (`v#v-w9501f1`, not displayed here).
(iii) The relative clause is represented as a subtree headed by the copula verb *být* and its actants (ACT, PAT), the possessive relative pronoun *jehož* (normalized as *který* 'which') is interlinked with its antecedent *blok* 'bloc', forming a coreference link.
(iv) The information structure is encoded by the ordering of the nodes and by their colors. The more to right a node is placed, the higher communicative dynamism it has. White color is used for contextually bounded nodes (*který* 'which'), the green one for a contrastive topic (e.g., *volba* for *volby* 'election'), and the yellow one for contextually unbounded nodes (e.g., *výsledek*, *včerejší*, and *předběžný*). From these values, topic (= *what is being talked about*) and focus (= *what is being said about the topic*) of the sentence can be derived.
(v) For simplicity, grammatemes (= counterparts for morphological features that are relevant for the sentence meaning) are not displayed here. For example, the main predicate verb *získat* 'win; get' has the annotation informing about its aspect (`cpl` as it presents the event as completed/as a whole), tense (`ant` for preceding/anterior event), deontic modality (`decl` for basic, unmarked modality), diathesis (`act` for active diathesis) and epistemic modality (`asserted` as the event is presented as given, signaled by the indicative form of the verb); in addition, the sentence modality (here `enunc` for declarative modality, i.e., assertions) is encoded with the main predicate.

One of the main features of the PDT-MR approach is its *focus on linguistically structured meaning* (rather than on semantics or even pragmatics beyond the language structure). As such, the goal of the PDT-MR annotation of a sentence is to capture especially the lexical choice (content words), the deep syntactic relations among them, the meaning-relevant morphological categories, coreference relations, and information structure. Consequently, the PDT-MR annotation more-or-less directly refers to the annotated text.

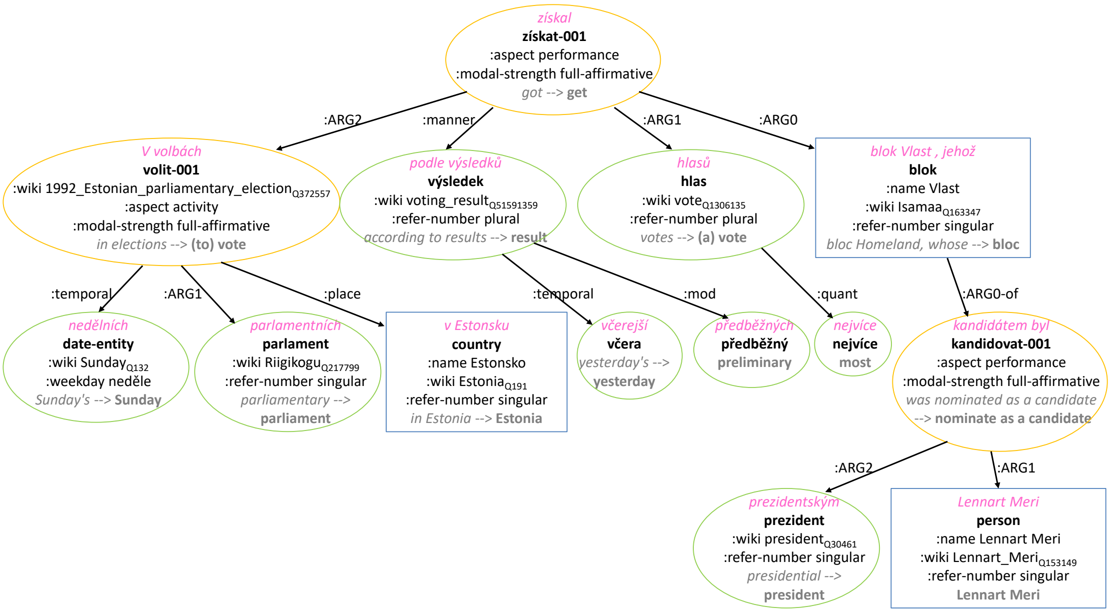


Figure 2: The ***sentence-level UMR graph*** of the same sentence as in Fig. 1, exemplifying the most relevant UMR principles:
(i) Similarly to PDT-MR, the core of the representation is formed by the predicate-argument structures: the predicate *získat-001* 'get' and its arguments ARG0 (for *blok Vlast* 'bloc Homeland'), ARG1 (for *hlas* 'vote'), and ARG2 (for *volby* 'elections', see below) form the upper part of the graph; further, there are two embedded predications here, one formed by the *volit* '(to) vote' predicate and its ARG1 (for *parlament* 'parliament'), the second formed by the predicate *kandidovat-001* 'to nominate as a candidate' and its ARG1 (for *Lennart Meri*) and ARG2 (for *prezident* 'president').
(ii) The concept of *volby* 'elections' is understood as an event (rather than an entity) and thus it is conceptualized as the predicate *volit-001* '(to) vote'; similarly the predicate *kandidovat-001* 'to nominate as a candidate' stands for the concept of *kandidát* 'candidate', as discussed in Sect. 3.2.1 and 3.3.1.
(iii) The named entity *Lennart Meri* is anchored in the Wikidata and it is classified as a *person* (which is an abstract concept); similarly, e.g., the event of Estonian elections (among others) is identified there as well.
The visualization is provided by [23]; to simplify the graph, UMR leaf-nodes for concepts and their respective variable nodes are merged (as suggested in Sect. 3.1, variable names are not displayed here). Predicates, corresponding to events, are marked by yellow ellipses, entities by green ellipses, and named entities are in blue boxes.

2.2. UMR: Uniform Meaning Representation

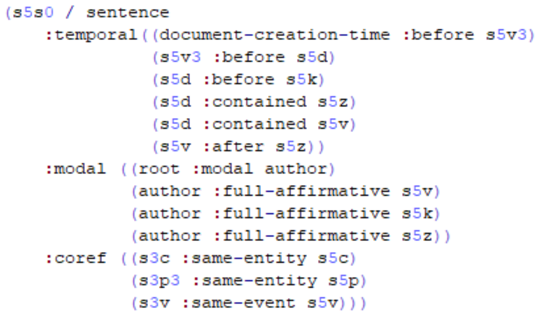
The Uniform Meaning Representation [20, 21, 22] is a semantic annotation schema that presents sentence meaning while abstracting away from syntax (and thus it is designed specifically for cross-lingual applications).

UMR elaborates the Abstract Meaning Representation (AMR) [24, 21] that focuses primarily on predicate-argument structures and was developed first for English but later applied also to other languages, incl. Czech [25]. This part of the UMR representation, referred to as the *sentence-level representation*, captures—in addition to predicate-argument structures—esp. representation of multi-word expressions and named entities (including their anchoring in the English Wikipedia), and aspect annotation for predicate verbs. Fig. 2 illustrates the UMR sentence level representation.

Besides that, UMR provides a more comprehensible annotation of epistemic modality, and marks temporal and coreference relations (both intra- and inter-sentential); this forms the *document-level representation*, as illustrated in Fig. 3.

UMR also aims to capture quantification and scope for the benefit of logical inference [20]; however, this kind of annotation is not available in the released dataset [22].

One of the main UMR features is a looser relation to syntax and the primary focus on semantics. Consequently, it provides the same representation for all possible (syntactic) variants of a statement, including its restructuring or splitting into more sentences. In fact, this approach allows for much broader interpretation of sentences to be represented, compared to the PDT-MR approach.



```
(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)))
```

Figure 3: The ***document-level UMR annotation*** of the same example sentence indicates temporal relations, modal dependencies and coreference chains identified in this sentence:
(i) The temporal annotation determines mutual temporal relations for all temporal expressions and all events identified in the sentence and relates them to the date of the document creation (e.g., the variable s5v3, standing for the relative temporal expression *včera* 'yesterday', refers to the particular time period before the document were created; further, the event conceptualized by the predicate *získat-001* 'get', variable s5z, happened after the event identified by the *volit-001* 'vote', variable s5v, etc.) .
(ii) The modal annotation indicates that the author of the text is sure that all three events identified in the sentence have happened (encoded as the ':full-affirmative' relation).
(iii) As this sentence is a part of a longer document, the annotation identifies which events and which entities has been already mentioned in the document (e.g., the *volit-001* 'vote' event, variable s5v, is the same event as the one with the variable s3v mentioned in one of the previous sentences).

3. Selected Features in More Detail

3.1. Graph structure

Both UMR and PDT-MR employ directed graphs for meaning representations. However, they differ in the way how graph nodes and edges are used to represent sentence meaning.

PDT-MR. In PDT-MR, the graph reflects deep syntactic structure of a sentence. Its nodes represent content words (or, better to say, their deep syntactic counterparts, see Fig. 1). Edges stand for deep syntactic relations between content words. The only exemptions are (i) the technical root node (containing metadata such as the ID of the sentence) serving as the mother node of the main predicate in a sentence and (ii) special nodes and edges used for the representation of paratactic structures (coordination, apposition). In fact, the PDT-MR structures are trees (when ignoring coreference links); i.e., any lexical content that should be repeated in the sentence structure (calls for "re-entrancy") is represented as two (or more) nodes linked by coreference arrow(s).

UMR. Compared to PDT-MR, UMR aims at a more abstract depiction of sentence meaning. Following its AMR predecessor, it seeks rather for a logical representation than for a syntactic one. According to the AMR 1.2.6 Specification (dated May 1, 2019),[4] it adopts a simplified, standard neo-Davidsonian semantics [26, 27, 24].

In UMR, two types of nodes are distinguish, as illustrated in Fig. 4. Non-leaves (inner nodes) stand for *variables*. The variables are instances of *concepts*, which are represented as leaves in the UMR graph. They represent primarily entities (as *man*, *parlament* 'parliament', or *blok* 'bloc') and events (as *získat-001* 'get', *volit-001* '(to) vote', or *taste-01*) but there are also special keywords for entity types (as date-entity in Fig. 2),[5] quantity types (e.g., temporal-quantity) and for discourse relations (as, e.g., conjunction and) and other operators (e.g., more-than).

Having two types of nodes, there are also two types of edges in UMR: First, the edges connecting leaf nodes with corresponding variables, representing thus the instance relation.

The second type of edges, those connecting non-leaf nodes represent semantic relations between concepts (instantiated as respective variables).

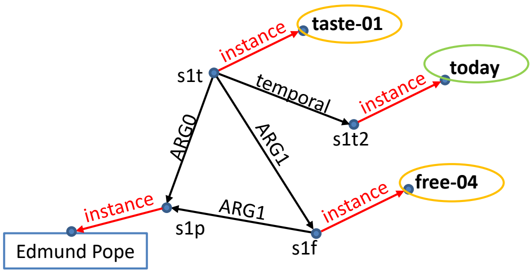


Figure 4: The AMR graph representation of the sentence *Edmund Pope tasted freedom today* (the sentence is taken from the released UMR data, simplified; it is also used as an example sentence in the UMR 0.9 Specification). Variables are non-leaves, concepts leaves; red arrows indicate the instance relation, black arrows stand for semantic relations between concepts. Eventive concepts are in yellow ellipses, the entity concept in a green one, and the named entity marked is by a blue box (for simplicity, its inner structure is hidden here). According to the UMR principles, the core structure is formed by the predicate *taste-01*, its arguments *Edmund Pope* (ARG0) and *freedom*, which is conceptualized as the predicate *free-04* (ARG1), and by ARG1 of the latter predicate. Thus, the single node for *Edmund Pope* serves here as an argument of two predicates (*taste-01* and *free-04*).

[4] https://github.com/amrisi/amr-guidelines/blob/master/amr.md
[5] These include also abstract entities (as person for *Lennart Meri*) identifying types of named entities, i.e., expressions referring to concrete persons, institutions, places, etc., see also footnote 15.

There is one important feature of the AMR (and UMR) formal representation: it allows for "re-entrancies", i.e., a variable can appear as a child of more than one semantic relation; thus the structure is a directed, rooted graph, which is typically acyclic, as illustrated in Fig. 4 and in a text-friendly way below (the slashes "/" represent the instance relation).

```
(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))
```

Conversion. As for the graph structure, transformation from PDT-MR to UMR is relatively easy: Each non-root PDT-MR node retains its id (= variable) and its lexical content is moved to a newly created leaf node, connected with the original one by an edge representing the instance relation. In this way, the requirements of the formal structure of the sentence-level annotation are secured. In addition, nodes within a single sentence that are marked as coreferential should be merged, as is discussed in Sect. 3.3 (under type A).

As for the opposite transformation (from UMR to PDT-MR), the situation is more complicated: First, each UMR leaf-node and the respective variable node must be merged to a single PDT-MR node, with the variable serving as the unique identifier of the node and with the concept having the role of the content word. Further, nodes with re-entrancies must be split (each of the original relations going to a different clone of the child node, and the children will be linked as coreferential). In this way, the adjusted sentence-level graph meets the PDT-MR formal requirements. However, there is no straightforward way how to transfer the information stored in the UMR document-level representation.

3.2. Events: Predicate and its argument structure

3.2.1. Concepts vs. content words as nodes

UMR. One of the crucial distinctions UMR works with is the conceptual distinction between entities (objects), states (properties), and events (processes). The aim is to abstract from morphological characteristics, i.e., whether the given concept appears in the surface text as a morphological noun, adjective, or verb. For example, think of a *driving* event as a concept represented by the *drive* predicate (*drive-01* in the PropBank lexicon [28]); then, *drive*, *driving* as well as *driver* are represented by this predicate and its argument structure (the last one as somebody that acts as ARG0 of the *driving* event).

PDT-MR. PDT-MR implements only the very first steps of such an abstraction, as represented by the concept of the so-called *t-lemma* (understood as a meaning counterpart of a lexical unit present in a surface sentence) and the *sempos attribute* (semantic part of speech). For example, the morphological possessive adjective *matčin* 'belonging to mother' is represented by the same t-lemma as the morphological and semantic noun *matka* 'mother' and it is characterized as a possessive form; similarly, the relative possessive pronoun *jehož* is captured as *který* in Fig. 1. However, different t-lemmas are supposed for *bojovat* 'to fight', *bojování* 'fighting', *boj* '(the) fight', and *bojovník* 'fighter'.[6]

Conversion. As for the event-entity distinction, transforming the PDT-MR data to UMR presents a challenge due to two main reasons: **A.** a lack of such information in the PDT-MR data and **B.** an insufficient definition of these concepts in UMR, as we discuss below.

A. Lack of information in PDT-MR: First, the relevant information—*which lexemes (words) are related to event concepts (verbal predicates)* (and how they are related) and *which rather represent entities*—is only very partially available in the PDT-MR data (in the form of a note in the related lexicons, see below). Thus, we also experiment with external lexical resources (in combination with some heuristics or manual annotation).

Selected types of derivational information are stored in the MorfFlex dictionary [29, 30], which covers general Czech morphology. Even more, this information is formalized so it can be used to identify base lexemes automatically. However, the lexicon is focused mainly on (inflectional) morphology; thus, the derivational information might not be entirely complete in some cases.

There is one more language resource very relevant for this task, namely DeriNet, the Lexical Network of Word-Formation Relations in Czech [31, 32]. It has a form of a network where nodes represent Czech lexemes and edges correspond to derivational links, i.e., relations between derivatives and their base lexemes. Thus DeriNet can reveal a relationship among, e.g., *bojovat* '(to) fight', *bojování* 'fighting', *boj* '(the) fight', *bojovník* 'fighter', *bojující* 'fighting (Adj)', *bojiště* 'battlefield', *bojovný* 'fighting (Adj)', *bojůvka* '(storm) troop', but also *zabojovat* 'to fight shortly', *odboj* 'resistance movement', *odbojář* 'resistance fighter', *zbojník* 'brigant', *souboj* 'combat', and many others; all of them share some aspects of the base lexeme *bojovat* '(to) fight'.

[6] The underlying Functional Generative Description [12, 13] stipulates that derived surface forms that preserve their semantic part of speech are represented by their base words, e.g., *bojovat* 'to fight', *bojování* 'fighting' and *boj* '(the) fight' (being semantic verbs) share the same lexical concept and thus should have the same representation, contrary to *bojovník* 'fighter', which is classified as semantic noun (agent noun). However, in the PDT data, only limited number of types are covered in this way.

However, neither MorfFlex nor DeriNet provide an information on the type of derivation—whether a particular derivative is an event nominal, agent noun, property, place, tool, etc. Thus, some heuristics must be applied to identify the particular type of derivation, as it has an impact on the argument structure of the derivative.

To start with the simplest and most systematic class of deverbal nouns, we focused on nouns ending with *-ní/-tí* first. In total, 1,690 such nouns are identified in the PDT-MR data (source: the PDT-Vallex lexicon, see Sect. 3.2.2). We combined this information with the derivational information stored in DeriNet and in MorfFlex.

We learned that even for this type of nouns, the derivation information is not complete in these resources in all cases. For example, ambiguous *dojetí* 'arrival' or 'emotion' is only shown as derived from *dojet* 'arrive' in DeriNet; in addition, MorfFlex identifies also *dojmout* 'touch; affect' as the base verb. With the help of DeriNet and MorfFlex, we were able to process 1,668 *-ní/-tí* nouns and identify the base verb lexemes. Seven of the remaining 22 nouns have an explanatory note in PDT-MR identifying the base verb, reducing the number of unidentified nouns to 15.

The second necessary step in the noun-to-verb conversion process is to identify the relevant sense among all senses of the identified base verb (as stored in the PDT-Vallex lexicon), which is a necessary step allowing us to capture the event argument structure of the derivative. This task still needs to be completed.

B. Unclear boundary between entities and events in UMR: Furthermore (and even more importantly), UMR does *not sufficiently define* the crucial *boundary between entities and events.* The UMR 0.9 Specification (dated August 8, 2022) simply states: [7]

> "[E]vent identification is based on a combination of semantic type and information packaging [33]. Semantic type refers to the difference between entities (or, objects), states (or, properties), and processes; this can be thought of as a categorization of things in the real world. Information packaging (also called discourse function or information structure), on the other hand, characterizes how a particular linguistic expression "packages" the semantic content."

However, the Specification suggests that it is (at least to some extent) the English grammar and English word-formation processes that are used as criteria to set the boundary (which, of course, contradicts the basic UMR principles).

For example, in the UMR approach, so-called stative verbs (as, e.g., *love*) are treated differently than verbs denoting processes (as, e.g., *run* or *damage*). According to the Specification, these verbs indicate events only if packaged as predication (and non-events in modification or reference packaging). Consequently, their annotations in a main clause and in a relative clause differ: For example, in *My cat loves wet food*, the verb *love* denotes an event of *loving*; in *My cat, that loves wet food, is beautiful*, the verb *love* is packaged as a modification, thus it is not considered as an event (with all the consequences for annotation). This distinction, however, is questionable for Czech and Latin, where statives represent a blurred category; thus, operative criteria for their identification cannot be applied (in contrast to English, where stative verbs exhibit specific syntactic behavior).

In any case, the criteria proposed in the Specification are not directly applicable to Czech and Latin data in many cases. This leads to the need to specify the boundary differently for those two languages. Tentatively, all concepts represented as predicate verbs and concepts used in predication are considered events in Czech and Latin UMR data.[8] As such, they are characterized by valency frames (∼ rolesets) stored in the PDT-Vallex lexicon (see Sect. 3.2.2). Further, abstract predicates, selected implicit rolesets (Sect. 3.2.3), and reified relations (Sect. 3.2.4) are treated as events in Czech and Latin UMR data.

3.2.2. PropBank lexicon vs. PDT valency lexicon

Core argument structure

UMR. UMR adopts the same principles for the representation of predicate-argument structure as used in AMR, relying on the PropBank lexicon (called "PropBank Frame Files") [34]. This lexicon stores predicates (mainly verbs, but also nouns and adjectives) subdivided into individual senses, assigned with a set of arguments and their coarse-grained semantic arguments[9] [28]. Originally designed for English,[10] it has been later used for a number of other treebanks of different languages (e.g. for Hindi, Chinese, Arabic and others); see also "IBM Universal Proposition Banks" project[11] [35].

The PropBank lexicon uses ARG0 to ARG5 labels to identify semantic roles of arguments, with ARG0 reserved for the `Prototypical_Agent` and ARG1 for the `Prototypical_Patient` or `Theme`[12] [36]. In addition, the ARGM label for adjunct-like arguments is being used (with several subtypes, as e.g., location). For example, the

[7] https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md#part-3-1-umr-concepts

[8] Thus, stative verbs are treated in the same way as verbs denoting processes, as we cannot set reliable criteria for their identification.

[9] http://propbank.github.io/

[10] http://propbank.github.io/v3.4.0/frames/index.html

[11] https://universalpropositions.github.io/

[12] According to [28], "No consistent generalizations can be made across verbs for the higher-numbered arguments, . . . "

PropBank lexicon contains two rolesets for the English verb *base*,[13] thus identifying its two senses, 1. 'be located in' and 2. 'justified by, made up of'; the first roleset, marked as *base-01*, describes three arguments:

ARG0-PAG: agent basing something somewhere,
ARG1-PPT: institution,
ARGM-LOC: for location, where ARG1 is based.

The argument labels are preserved across the rolesets of a given predicate (wherever relevant), disregarding the sentence structure (e.g., compare *John*.ARG0 *broke the window*.ARG1 to *The window*.ARG1 *broke*).

PDT-MR. PDT-MR adheres to the original valency theory [37], as instantiated in several valency lexicons, VALLEX [38, 39], PDT-Vallex [40, 41] (both for Czech), EngVallex [42, 43] (for English), and Latin VALLEX v1[14] [44] (for Latin). The PDT-Vallex and EngVallex serve for annotation of the PDT and PCEDT corpora, while Latin VALLEX v1 was built upon the tectogrammatical layer of Latin texts annotated in the PDT style.

Similarly as the PropBank lexicon, the lexicons of the Vallex family provide valency frames ($\sim$ rolesets) for individual predicate (primarily verb) senses. Instead of numbered arguments, they use five labels for the so-called actants (ACT for `Actor/Bearer`, PAT for `Patient`, ADDR for `Addressee`, ORIG for `Origin`, and EFF for `Effect`). However, their specification differs from the PropBank approach: the first two arguments (ACT and PAT) have no specific semantics, ACT being assigned typically to the argument in the subject position (in an active sentence), PAT to the argument in the object position (for verbs with the only object position). Only with verbs with three and more arguments, semantics of individual arguments plays role. As a consequence, the labels are not preserved in lexical alternations with different syntactic structure (e.g., compare *Jan*.ACT *rozbil okno*.PAT, 'John.ACT broke the window.PAT' and *Okno*.ACT *se rozbilo*, 'The window.ACT broke').

Conversion. As illustrated above, the PDT-MR and UMR approaches differ in the argument labeling style and in the specification of individual argument roles (manifested esp. in the treatment of the first two arguments). Fortunately, Hajič et al. [45] provide a partial mapping of PDT-MR rolesets to PropBank-based UMR rolestes. Based on existing resources, they have been able to convert automatically and with high certainty about 43% of PDT-Vallex argument labels, so-called functors (out of 42,116 PDT-Vallex functors) to PropBank argument labels. In this way, the core of the PropBank-like lexicon for Czech has been established. The lexicon has a form of a table with *verb specific argument mapping* (when available) that can be used for the automatic conversion of the PDT-MR to UMR representation. Additionally, the table stores information on candidate mappings and includes supplementary valency information that can be used for future manual extensions of the mappings.

As a fallback solution for predicate verbs without proposed mappings of PDT-MR functors to PropBan arguments, *a default mapping* can be used, as suggested by [46, 47]. Based on introspection,they hypothesized that ACT typically corresponds to the ARG0 argument, PAT is most often ARG1, ADDR is typically ARG2, and so on. For example, the verb *živit* 'nourish' has two actants, which by default get the following roles:

ACT (`Agent/Causal agent`) → ARG0,
PAT (`Entity fed or maintained`) → ARG1.

In this case, the argument labels agree with the argument specification and argument labels of the English verb *nourish* (as provided in the PropBank lexicon), thus the default mapping is correct. However, for verbs with more than two arguments, the default mapping is not satisfactory enough. For example, the verb *nachystat* 'prepare' has three actants, which by default get the following roles:

ACT (`Creator`) → ARG0,
PAT (`Thing made ready`) → ARG1,
ORIG (`Created_from`) → ARG3.

The mapping of the first two argument labels seems correct (as they agree with the arguments of the verb *prepare*); however, the semantics of the last argument (`Created_from`) correspond rather to ARG2 of *prepare* (thus, the default ARG3 label is inconsistent with PropBank).

To evaluate the proposed approach, we compared the default mapping and the verb-specific mapping presented by [45] on the available 10,426 functor-argument pairs; the results are in Table 1. The table reveals that the default mapping represents a relatively good approximation for the first three actants (reaching accuracy of 86.8%). However, it is a futile attempt to use the proposed mapping for the last two actants. Instead, we suggest to convert them to the general (verb non-specific) UMR roles `Source` (for ORIG) and `Goal` (for EFF).

To summarize, the combination of the mapping procedure proposed in [45] (for predicate verbs with the verb-specific mapping available) and the default arguments mapping for ACT, PAT and ADDR actants (for other verbs) can serve as a good starting point for future manual refinement of UMR argument labeling for Czech predicate verbs.

Non-core arguments and adjuncts

Both UMR and PDT-MR distinguish a wide range of labels for non-core arguments, adjuncts, and other relations beyond the scope of the (core) argument structure.

[13] http://propbank.github.io/v3.4.0/frames/base.html
[14] https://github.com/CIRCSE/Latin-VALLEX

Table 1
Accuracy of the default actant to argument mapping.

PDT-MR → UMR mapping	correct (%)	incorrect
ACT → ARG0	4,355 (82.6%)	918
PAT → ARG1	3,829 (92.5%)	310
ADDR → ARG2	464 (84.4%)	86
ORIG → ARG3	51 (20.6%)	197
EFF → ARG4	0 (0.0%)	216
total	8,699 (83.4%)	1,727

UMR adopts rather coarse-grained labels for adverbial modifications; compare, e.g., two general labels for temporal relations (`temporal`, `duration`) and nine more fine-grained temporal labels used in PDT-MR (distinguishing, e.g., relations like `When?`, `From_when?`, `To_when?`, or `Till_when?`). Given this, PDT-MR labels for adverbial modifications can be (at least tentatively) mapped to UMR relations based on a simple translation table.

However, UMR introduces also some relations that are more specific than those used in PDT-MR, as, e.g., `quantity`, `age`, `topic`, or `medium`. Thus, in the data converted from PDT-MR, these labels are not identified correctly; instead, more general UMR labels are used, as, e.g. mod (when used as a nominal modifier).

3.2.3. Abstract concepts vs. strong lexicalization

One of the main goals of the UMR approach is to provide a meaning representation usable for various languages allowing for cross-linguistic comparability of annotations. This is supported by the introduction of the concepts of *abstract predicates* and *implicit rolesets*.[15] PDT-MR, on the other hand, can be characterized as strongly lexicalized approach, relying on the predicate structure of individual (mostly verbal) predicates.

Abstract predicates

UMR. UMR introduces nine abstract (also referred to as "non-verbal") predicates. These predicates are used for representing the predication of property, possession, and location. They are identified by special labels (serving as artificial lemmas) equipped with their own rolesets.

For example, the abstract predicate *have-place-91* is used for a (predicative) location; it has two argument roles, ARG1 for `Theme` and ARG2 for `Location`. This predicate applies, e.g., in sentences like *Brambory*.ARG1 *jsou ve sklepě*.ARG2, 'The potatoes.ARG1 are in the cellar.ARG2'. Similarly, the *exist-91* predicate represents a thetic location, characterized by a reverse role semantics (applicable, e.g., for *Na obzoru*.ARG1 *je Sněžka*.ARG2, 'There is the Sněžka mountain on the horizon').

Another example of constructions that should be treated using abstract predicates are constructions with the copula verb, corresponding to *být* 'be' in Czech and *sum* 'be' in Latin. In those cases, the following abstract predicates should be used:

- *have-mod-91*
  (as in cs. *Podle čeho soudíte, že v tom jste nejlepší?*, 'Why do you think you are the best at that?'; lat. *Vita ipsa brevis est*, 'Life itself is short'),
- *have-role-91*
  (as in cs. *Vinken je prezidentem společnosti Elsevier N. V.*, 'Mr. Vinken is a chairman of Elsevier N.V.'; lat. *Cato quaestor fuit*, 'Cato was quaestor'),
- *identity-91*
  (as in cs. *USA jsou jedinou zemí, kde . . .*, 'The US is the only country where . . . ').

Conversion. Identifying candidate constructions in PDT-MR data that should be represented by abstract predicates is a challenging task. We can indicate a tentative list of Czech predicates (e.g., the respective senses of the verbs *mít* 'have', *patřit* 'belong', *vlastnit* 'own', etc.) as well as other relations (as, e.g., constructions with possessive forms, like *Mariina/její taška*, 'Maria's/her bag') that express possession with reasonable certainty. However, the cases of location predication and property predication are more complex as all their subtypes are typically expressed by the verb *být* 'be' in Czech, which is categorized either (i) as the copula or (ii) as the existential or so-called substitute verb (subsumed under the single PDT-Vallex entry in the PDT-MR data). Thus, it is not possible to automatically distinguish more subtle senses as required in UMR.

The same holds true for Latin, where the verb *sum* 'be' can correspond to several UMR abstract predicates.

Implicit rolesets

UMR works with a list of other implicit rolesets that conceptualize various linguistic constructions. We can distinguish two main types here, rolesets for special linguistic constructions and those used for the analysis of structured texts.

[15] In addition, UMR employs a set of abstract entities identifying entity types. They serve several purposes:
(i) they stand for arguments in case of not overtly present arguments (or arguments present just as pronouns),
(ii) they are used for classification of named entities (e.g., *Lennart Meri* is classified as a `person` in Fig. 2), and
(iii) they provide an identification of structured data as special "entities" (as, e.g., `date-entity`, further structured with attributes like day, month, year, century, etc.) or "quantities" (as, e.g., `monetary-quantity`, `temporal-quantity-quantity`, both with the attributes quant and unit).
Focusing on the predicate-argument structure in UMR and PDT-MR, we leave abstract entities aside here.

A. First, to exemplify the **rolesets for special linguistic constructions**, we can list the following:

- *have-degree-91* as in comparison constructions (e.g., *Dívka je vyšší než chlapec*, 'The girl is taller than the boy'),
- *include-91*, as in the part-whole relation (e.g., *Pro blok Vlast hlasovalo asi 20.5 procenta z celkového počtu 457 319 voličů*, 'About 20.5 percent of the total number of 457,319 voters voted for the Homeland bloc.')
- *resemble-91* is used for analogies (e.g., *It was like mud running down the mountain…*)

Identification of similar constructions in the PDT-MR data requires future examination; we postpone this task to the (near) future.

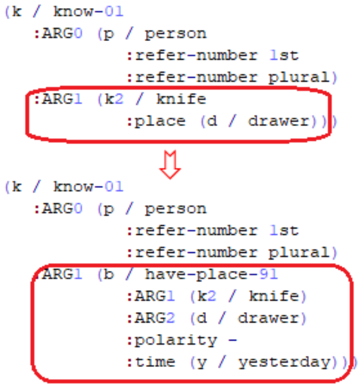
B. Second, the **structured texts** can be exemplified with, e.g., *cite-91* for citations, *hyperlink-91*, or *street-address-91.* In PDT-MR, there are special rules for structured text. However, the representation adhere to the language (deep syntactic) structure: it is governed by the general rules for distinguishing between verbal clauses and non-verbal clauses. Thus, we assume that there is only a very limited possibility to automatically convert the PDT-MR data to the UMR-compliant form.

3.2.4. Reification

UMR. UMR fully adopts the AMR approach, within which reification is understood as a technique to *convert a relation into an (abstract) predicate.* This technique makes it possible to put focus on the (original) relation itself, to modify or to negate it. Then, the concepts interlinked by the original relation are converted to the arguments of the resulting abstract predicate.

To illustrate the technique, compare the annotation of the sentence *We know the knife that is in the drawer* (focusing on the *knife*, Fig. 5, upper part) and its modification *We know the knife was not in the drawer yesterday*; as the `place` relation is modified by the temporal concept and negated, it asks for reification, i.e., it is converted to the *have-place-91* predicate, as shown in Fig. 5, lower part.[16]

According to the AMR 1.2.6 Specification (dated May 1, 2019), "AMR with reification" is considered "real AMR" (with "non-reified relations as semantic sugar"). To put it differently, following the "reify all the time" principle, it would eliminate almost all relations[17] and replace them with abstract predicates. However, as this would be an inconvenient technicality, the AMR Specification prefers non-reified relations when annotating data, aiming to support the corpus consistency—unless reification is needed (i.e., unless focusing on the relation, modifying or negating it). Further, it suggests that the AMR representation "will be normalized into reified form behind the scenes."



```
(k / know-01
   :ARG0 (p / person
            :refer-number 1st
            :refer-number plural)
   :ARG1 (k2 / knife
            :place (d / drawer)))
```

```
(k / know-01
   :ARG0 (p / person
            :refer-number 1st
            :refer-number plural)
   :ARG1 (b / have-place-91
            :ARG1 (k2 / knife)
            :ARG2 (d / drawer)
            :polarity -
            :time (y / yesterday)))
```

Figure 5: The UMR representation of the `place` relation (above, the sentence *We know the knife that is in the drawer*) and its reification using the *have-place-91* abstract predicate (below, the sentence *We know the knife was not in the drawer yesterday*).

PDT-MR. PDT-MR does not allow for negating or modifying a relation itself – the underlying principles suppose that a speaker will overtly mark such communication needs, thus they will express the focus on the relation itself by choosing different syntactic structure and/or lexicalization, accompanied it with the relevant information structure.

Conversion. Given the fact that AMR (and thus also UMR) relies on the data post-processing within which the AMR/UMR representations are converted into the reified forms, we give up attempts to identify constructions in the PDT data that call for reification and leave them to be handled in the subsequent phases of the project.

3.3. Coreference

Coreference is generally understood as a relation between two or more expressions in a text that refer to the same concept, seen as a mental concept of a real-world entity or event. Such expressions usually form coreferential chains, which make the text(s) coherent. Members of the coreferential chain are connected by an *anaphoric relation*, i.e., the intra-textual relation that is bilateral and asymmetric, having an *anaphor* (a pronoun in most typical case) and its *antecedent/postcedent* (usually a content word). Primarily expressions referring to entities,

[16] The example is borrowed from the AMR Specification, https://github.com/amrisi/amr-guidelines/blob/master/amr.md#reification, and converted to follow the UMR principles.

[17] With the exception of argument relations, relations used in conjunctions, and relations or attributes related to abstract entities, see also footnote 15.

but also those referring to events can be interlinked with the anaphoric relation.

PDT-MR. In the PDT-MR approach, all types of coreference are treated in the same way. Each expression in the anaphoric reation is typically represented as a single node in the graph.[18] The (node for the) anaphor bears a set of coreference attributes identifying esp. the ID(s) of the antecedent/postcedent node(s), the type of the coreference[19] and the type of reference (e.g., specific or generic).

For example, in Fig. 1, the relative pronoun *jehož* (normalized as *který* 'which') coreferes with *blok* 'bloc' (i.e., it identifies its ID in the respective attribute), which is visualized as the brown arrow (brown color stands for grammatical coreference).

From the technical point of view, there is no difference in treating coreference (and bridging relations) within a sentence and these relations crossing sentence boundaries.

UMR. UMR offers three ways how to capture coreference relations (two applicable within a single sentence and one for inter-sentential relations); we will briefly sketch them now and comment on the possibilities of the PDT-MR to UMR conversion.

3.3.1. Coreference within a single sentence

Re-entrancy of a variable

As already discussed in Sect. 3.1, there is a possibility of a re-entrancy of a variable within a sentence in UMR, exemplified by Fig. 4 representing the sentence *Edmund Pope tasted freedom today*. Here the *Edmund Pope* entity serves as ARG0 of the *taste-01* predicate and at the same time as ARG1 of the *free-04* predicate. This type is strictly limited to intra-sentential relations and applies to the sentence-level representation.

In PDT-MR, there are two nodes representing this entity, one in the argument structure of the predicate *taste-01* (with the t-lemma *Edmund Pope*) and one in the argument structure of *free-04* (with the special coreferential t-lemma substitute `#Cor` and the coreference attribute storing the ID of the antecedent node); the relation is visualized as an arrow interconnecting these two nodes, see Fig. 6.

The information provided in PDT-MR is sufficient for the sound identification of nodes that should be merged when converting the PDT-MR representation to the UMR one.

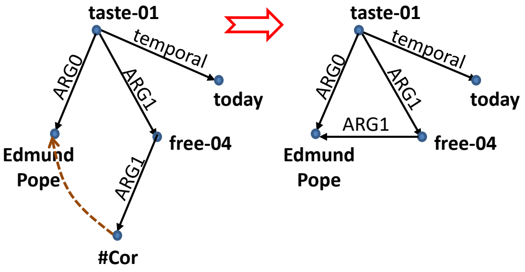


Figure 6: Coreference in PT-MR (left) transformed to re-entrancy in UMR (rifgt), illustrated with the sentence *Edmund Pope tasted freedom today*.

Inverse roles

Further, UMR employs the idea of *inverse roles* (already introduced in AMR), which are used primarily for the annotation of relative clauses (type A below), nominalizations (type B below), and embedded interrogatives (which we leave aside for the time being). The inverse role makes it possible to represent an embedded event as an event modifying one of the arguments; thus it is relevant for the sentence-level representation.

A. In PDT-MR, a **relative clause** is represented as a subtree rooted at its verbal predicate, which itself is a child of (the node for) the modified concept; further, the relative expression (a relative pronoun or pronominal adverb) typically serves as an argument or adjunct in the valency structure of the clause predicate.[20] At the same time, there is a coreferential link between the relative expression and the modified concept. For example, consider the following sentence: *Lidé, kteří bydlí v blízkosti závodu, si stěžují na zdravotní potíže*, 'People who live near the factory have been complaining of health problems'. Its (simplified) PDT-MR representation is presented in Fig. 7 (left). The tree for the relative clause is rooted at the node for the verb *bydlet* 'live', which is treated as the head of the attribute clause (the RSTR functor) modifying the expression *člověk* 'person' (standing for *lidé* 'people'); the relative pronoun *který* 'who' is ACT of the *bydlet* 'live' predicate; this pronoun is interconnected with its antecedent *člověk* 'person' with a coreferential link.

When converting to UMR, the node for the relative expression and the one for the modified expression are merged; further, the relation between (the node for) the predicate verb of the relative clause and (the one for)

[18] The antecedent/postcedent node(s) may stand for the whole subtree(s) it/they govern(s). Further, the PDT-MR annotation schema also allows for exophora, i.e., a type of coreference with a pronoun referring to a situation or reality external to the text; we will leave such cases aside here.

[19] Apart from grammatical and textual coreference, relations of bridging anaphora are also distinguished, incl. relations such as set-subset, part-whole or function-object.

[20] The relative expression can be embedded more deeply in the sentence (e.g., in the case of the possessive personal pronoun); in such cases, the conversion follows the same principles.

the modified expression gets the relation inverse to the original relation between the predicate and the relative expression (i.e., typically the inverse role to the original argument or adjunct relation). Thus, in the example sentence above, the predicate *bydlet* 'live' remains a node modifying *člověk* 'person'; the node for the relative pronoun *který* 'who' disappears (= is merged with the node for *člověk* 'person') and the respective relation it represented (ACT of *bydlet* 'live') is preserved as the ACT-of relation between *člověk* 'person' (mother node) and *bydlet* 'live' (daughter node) (indicating that *člověk* 'person' is an ACT-of *bydlet* 'live'),[21] see Fig. 7 (right).

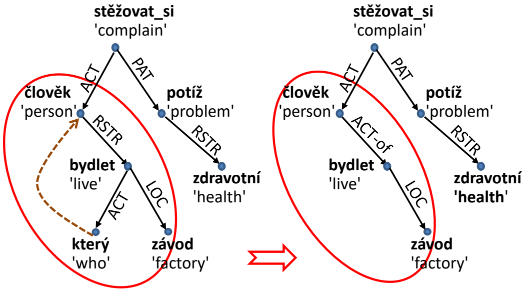


Figure 7: Example relative clause in PDT-MR (left) and in UMR (right), illustrated with the sentence *Lidé, kteří bydlí v blízkosti závodu, si stěžují na zdravotní potíže* 'People who live near the factory have been complaining of health problems'.

B. As for **nominalizations**, the situation is more tricky. For example, let us have the nominal group *představitel republiky* 'the representative of the republic': in UMR, the agentive noun *představitel* '(the) representative' should be seen as ARG0 of the predicate *představovat-003* 'represent' (i.e., it is 'the person who represents the republic'), while *republika* 'republic' serves as ARG1 of the predicate, as the following annotation shows:

```
(p/ person
   :ARG0-of (p2 / představovat-003 'represent'
                  :ARG1 republika 'republic')
```

Unfortunately, neither the PDT-MR data nor available external resources allow for a sound identification of nominalizations and their source predicates (as already discussed in Sect. 3.2.1). Thus, only for those entities that can be identified as related to events—i.e., we can identify the source predicate, the type of derivation, and the respective valency frame (∼ roleset) in the PDT-Vallex lexicon identifying the argument structure, as was discussed in Sect. 3.2.1 and 3.2.2—we can modify the source PDT-MR representation to comply with the target UMR principles.

3.3.2. Coreference crossing a sentence boundary

Finally, the *inter-sentential coreference relations* are treated within the document-level representation in the UMR. For each sentence, all nodes with a coreferential link indicating a node (or nodes) outside the sentence must be collected and the respective pairs of coreferring nodes added to the document-level part of the sentence annotation. Further, the proper relation between the pair members must be identified, reflecting (1) whether they refer to the same entity or to the same event and (2) whether their mutual relation is a relation of the identity (both nodes represent the same referent) or it is a relation between a set and its (proper) subset / event and its subevent.

The event vs. entity distinction is crucial for the UMR approach and as such, it is reflected in the PDT-MR conversion, see Sect. 3.2.1. Thus, coreferential nodes identified during the conversion as the identical events are linked with the same-event relation, the ones identified as identical entities get the same-entity relation.

As for the identity vs. subset distinction, the coreferential links in the PDT-MR data capture primarily the relation of identity; therefore, these links are treated as the same-entity (or same-event) relation. In the PDT-MR data, the subset relations can be extracted from the annotation of bridging relations. We postpone the work on this type of relations to the future.

It is important to note that coreferential relations between events are only sporadically captured in the PDT-MR data. As a result, these relations cannot be identified automatically—additional analysis and manual annotation are needed to comply with the UMR principles.

4. Conclusion and Plans for the Future

In this paper, we have outlined the basic characteristics of the two approaches to meaning representation, namely the Uniform Meaning Representation (UMR) and meaning representation used in the Prague Dependency Treebank (PDT-MR).

We have concentrated esp. on those features that influence the structure of sentence graphs. We have started with the formal properties of the graph and then continued with the main linguistic phenomena relevant to the graph structure: the event/entity distinction, the two approaches to capturing predicate verbs and their argument structures, lexical as well as abstract ones (including lexical resources storing this type of information), and the handling of coreference chains. We have discussed the possibilities of the automatic conversion of these phenomena from the PDT-MR approach to UMR. We have also presented the first results of this conversion for indi-

[21] For simplicity, the PDT-MR-like labels are kept in both graphs in Fig. 7 (supposing that they will be converted to the UMR labels in the following steps).

vidual subtasks, which can serve as a baseline for future improvements.

It is clear that *the possibility of a direct automatic "translation" between these two representations* is significantly limited, despite the availability of rich linguistic resources capturing the semantics of Czech and Latin. Even in this case, it is necessary to search for appropriate heuristics for the linguistic phenomena under study and to supplement the automatic procedure with a substantial amount of manual annotation.

As for the *future plans*, we want to focus primarily on the complex phenomena identified above. First, we will examine the possibility of an automatic identification of other nominal and adjectival derivatives that should be treated as events in UMR. Abstract predicates and implicit rolesets represent the second area that calls for detailed investigation—our goal for the near future is to prepare (at least a tentative) list of Czech predicate verbs that should be converted to UMR abstract predicates as well as to detect candidate linguistic constructions for conversion using UMR implicit rolesets.

Further, there are many other linguistic phenomena covered in the PDT data that we left aside in the current stage of the PDT-MR to UMR conversion. Among the most relevant ones, let us mention at least three areas: (i) the elaborated structure of grammatemes (= counterparts for morphological features that are relevant for the sentence meaning, as tense, aspect, or deontic modality) that should inform the related UMR attributes, and (ii) the detailed discourse annotation that should be converted to the UMR discourse rolesets. Moreover, (iii) the identification of named entities and their inner structure following the UMR principles, as well as their proper anchoring in Wikipedia remain a challenging task to be addressed.

Given the complexity of the PDT-MR and UMR representations, the paper primarily compares selected language phenomena and their treatment in both representations. It describes the first experiments aiming at the PDT-MR to UMR conversion. For the time being, we cannot provide *an evaluation of the conversion results*, not even for the selected phenomena, due to the fact that there are no UMR golden data available for Czech or Latin (or for other languages with available PDT-MR representation). Thus, the only way we can carry on the evaluation is to compare the generated structures with ad-hoc manually annotated data for Czech/Latin and measure, e.g., how different the generated data are from the manually created ones. Hence, the evaluation represents the most urgent task for the near future.

Acknowledgments

The work described herein has been supported by the grants *Language Understanding: from Syntax to Discourse* of the Czech Science Foundation (Project No. 20-16819X) and *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic. It has also been partially supported by the Charles University, GAUK project No. 104924, and SVV project No. 260 698.

The project has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

[1] A. Copestake, D. Flickinger, C. Pollard, I. A. Sag, Minimal Recursion Semantics: An Introduction, Research on Language and Computation 3 (2005) 281–332. doi:10.1007/s11168-006-6327-9.

[2] H. Kamp, U. Reyle, From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory, Studies in Linguistics and Philosophy, Springer, Dordrecht, 1993. doi:10.1007/978-94-017-1616-1.

[3] J. Bos, V. Basile, K. Evang, N. Venhuizen, J. Bjerva, The Groningen Meaning Bank, in: Handbook of Linguistic Annotation, Springer, 2017, pp. 463–496. doi:10.1007/978-94-024-0881-2.

[4] Z. Žabokrtský, D. Zeman, M. Ševčíková, Sentence meaning representations across languages: What can we learn from existing frameworks?, Computational Linguistics 46 (2020) 605–665. doi:10.1162/coli_a_00385.

[5] S. Oepen, O. Abend, L. Abzianidze, J. Bos, J. Hajic, D. Hershcovich, B. Li, T. O'Gorman, N. Xue, D. Zeman, MRP 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing, in: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, Association for Computational Linguistics, 2020, pp. 1–22. doi:10.18653/v1/2020.conll-shared.1.

[6] A. K. Žolkovskij, I. A. Mel'čuk, O vozmožnom metode i instrumentax semantičeskogo sinteza (On a possible method and instruments for semantic synthesis), Naučno-texničeskaja Informacija (1965) 23–28.

[7] I. A. Mel'čuk, Dependency syntax: Theory and practice, SUNY Press, Albany, NY, 1988.

[8] S. Schuster, C. D. Manning, Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks, in: Proceedings of LREC 2016, ELRA, Portorož, Slovenia,

2016, pp. 2371–2378. URL: https://aclanthology.org/L16-1376.

[9] E. Hajičová, Dependency-Based Underlying-Structure Tagging of a Very Large Czech Corpus, Special Issue of TAL Journal, Grammaires De Dépendence / Dependency Grammars (2020) 57–78.

[10] J. Hajič, E. Hajičová, J. Mírovský, J. Panevová, Linguistically Annotated Corpus as an Invaluable Resource for Advancements in Linguistic Research: A Case Study, The Prague Bulletin of Mathematical Linguistics (2016) 69–124. doi:10.1515/pralin-2016-0012.

[11] D. Zeman, J. Hajič, FGD at MRP 2020: Prague Tectogrammatical Graphs, in: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, Association for Computational Linguistics, 2020, pp. 33–39. URL: https://aclanthology.org/2020.conll-shared.3. doi:10.18653/v1/2020.conll-shared.3.

[12] P. Sgall, Generativní popis jazyka a česká deklinace (Generative Description of a Language and the Czech Declension), Academia, Praha, 1967.

[13] P. Sgall, E. Hajičová, J. Panevová, The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Reidel, Dordrecht, 1986.

[14] J. Hajič, E. Bejček, J. Hlaváčová, M. Mikulová, M. Straka, J. Štěpánek, B. Štěpánková, Prague Dependency Treebank - Consolidated 1.0, in: Proceedings of LREC 2020, ELRA, Marseille, France, 2020, pp. 5208–5218. URL: https://aclanthology.org/2020.lrec-1.641/.

[15] J. Hajič, E. Bejček, A. Bémová, E. Buráňová, E. Fučíková, E. Hajičová, J. Havelka, J. Hlaváčová, P. Homola, P. Ircing, J. Kárník, V. Kettnerová, N. Klyueva, V. Kolářová, L. Kučová, M. Lopatková, D. Mareček, M. Mikulová, J. Mírovský, A. Nedoluzhko, M. Novák, P. Pajas, J. Panevová, N. Peterek, L. Poláková, M. Popel, J. Popelka, J. Romportl, M. Rysová, J. Semecký, P. Sgall, J. Spoustová, M. Straka, P. Straňák, P. Synková, M. Ševčíková, J. Šindlerová, J. Štěpánek, B. Štěpánková, J. Toman, Z. Urešová, B. V. Hladká, D. Zeman, Š. Zikánová, Z. Žabokrtský, Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0), 2020. URL: http://hdl.handle.net/11234/1-3185, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[16] P. Synková, J. Mírovský, L. Poláková, M. Rysová, Announcing the Prague Discourse Treebank 3.0, in: Proceedings of LREC-Coling 2024, ELRA, Torino, Italy, 2024, pp. 1270–1279. URL: https://aclanthology.org/2024.lrec-main.114.

[17] P. Synková, M. Rysová, J. Mírovský, L. Poláková, V. Sheller, J. Zdeňková, Š. Zikánová, E. Hajičová, Prague Discourse Treebank 3.0, 2022. URL: http://hdl.handle.net/11234/1-4875, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[18] J. Hajič, E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, Z. Žabokrtský, Announcing Prague Czech-English Dependency Treebank 2.0, in: Proceedings of LREC 2012, ELRA, Istanbul, Turkey, 2012, pp. 3153–3160. URL: https://aclanthology.org/L12-1280/.

[19] J. Hajič, E. Hajičová, J. Panevová, P. Sgall, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, Z. Žabokrtský, Prague Czech-English Dependency Treebank 2.0, 2012. URL: http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[20] J. van Gysel, M. Vigus, J. Chun, K. Lai, S. Moeller, J. Yao, T. O'Gorman, J. Cowell, W. Croft, C.-R. Huang, J. Hajič, J. Martin, S. Oepen, M. Palmer, J. Pustejovsky, R. Vallejos, Designing a uniform meaning representation for natural language processing, KI - Künstliche Intelligenz 35 (2021) 343–360. doi:10.1007/s13218-021-00722-w.

[21] S. Wein, J. Bonn, Comparing UMR and cross-lingual adaptations of AMR, in: Proceedings of the Fourth International Workshop on Designing Meaning Representations (DMR 2023), Association for Computational Linguistics, Nancy, France, 2023, pp. 23–33. URL: https://aclanthology.org/2023.dmr-1.3.

[22] J. Bonn, C. Ching-wen, J. A. Cowell, W. Croft, L. Denk, J. Hajič, K. Lai, M. Palmer, A. Palmer, J. Pustejovsky, H. Sun, R. Vallejos Yopán, J. Van Gysel, M. Vigus, N. Xue, J. Zhao, Uniform meaning representation, 2023. URL: http://hdl.handle.net/11234/1-5198, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[23] M. Novák, UMR Visualization, unpublished, 2023.

[24] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract Meaning Representation for Sembanking, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 178–186. URL: https://aclanthology.org/W13-2322.

[25] N. Xue, O. Bojar, J. Hajič, M. Palmer, Z. Urešová, X. Zhang, Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech, in: Proceedings of LREC 2014, ELRA, Reykjavik, Iceland, 2014, pp. 1765–1772. URL: https://aclanthology.org/L14-1332/.

[26] D. Davidson, The logical form of action sentences, in: N. Rescher (Ed.), The Logic of Decision and Action, University of Pittsburgh Press, 1967, pp. 81–95.

[27] J. Higginbotham, On semantics, Linguistic inquiry 16 (1985) 547–593.

[28] M. Palmer, D. Gildea, P. Kingsbury, The Proposition Bank: An Annotated Corpus of Semantic Roles, Computational Linguistics 31 (2005) 71–106. doi:10.1162/0891201053630264.

[29] J. Hlaváčová, M. Mikulová, B. Štěpánková, Konzistence morfologického slovníku MorfFlex, Jazykovedný časopis / Journal of Linguistics 72 (2021) 855–861.

[30] J. Hajič, J. Hlaváčová, M. Mikulová, M. Straka, B. Štěpánková, Morfflex CZ 2.0, 2020. URL: http://hdl.handle.net/11234/1-3186, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[31] L. Kyjánek, Z. Žabokrtský, M. Ševčíková, J. Vidra, Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources, The Prague Bulletin of Mathematical Linguistics 115 (2020) 5–30.

[32] J. Vidra, Z. Žabokrtský, L. Kyjánek, M. Ševčíková, Šárka Dohnalová, E. Svoboda, J. Bodnár, DeriNet 2.1, 2021. URL: http://hdl.handle.net/11234/1-3765, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[33] W. Croft, Radical Construction Grammar: Syntactic Theory in Typological Perspective, Oxford University Press, 2001. doi:10.1093/acprof:oso/9780198299554.001.0001.

[34] S. Pradhan, J. Bonn, S. Myers, K. Conger, T. O'Gorman, J. Gung, K. Wright-Bettner, M. Palmer, PropBank comes of age—larger, smarter, and more diverse, in: Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, ACL, Seattle, Washington, 2022, pp. 278–288. doi:10.18653/v1/2022.starsem-1.24.

[35] I. Jindal, A. Rademaker, M. Ulewicz, H. Linh, H. Nguyen, K.-N. Tran, H. Zhu, Y. Li, Universal proposition bank 2.0, in: Proceedings of LREC 2022, ELRA, Marseille, France, 2022, pp. 1700–1711. URL: https://aclanthology.org/2022.lrec-1.181.

[36] D. R. Dowty, Thematic proto-roles and argument selection, Language 67 (1991) 547–619. doi:10.2307/415037.

[37] J. Panevová, Valency Frames and the Meaning of the Sentence, The Prague School of Structural and Functional Linguistics 41 (1994) 223–243.

[38] M. Lopatková, V. Kettnerová, E. Bejček, A. Vernerová, Z. Žabokrtský, Valenční slovník českých sloves VALLEX (The Valency Dictionary of Czech Verbs VALLEX), Karolinum, Praha, 2016.

[39] M. Lopatková, V. Kettnerová, J. Mírovský, A. Vernerová, E. Bejček, Z. Žabokrtský, VALLEX 4.5, 2022. URL: http://hdl.handle.net/11234/1-4756, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[40] J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolářová, P. Pajas, 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, 2003, pp. 57–68.

[41] Z. Urešová, A. Bémová, E. Fučíková, J. Hajič, V. Kolářová, M. Mikulová, P. Pajas, J. Panevová, J. Štěpánek, PDT-Vallex: Czech Valency lexicon linked to treebanks 4.0 (PDT-Vallex 4.0), 2021. URL: http://hdl.handle.net/11234/1-3499, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[42] S. Cinková, From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description, in: Proceedings of LREC 2006, ELRA, ELRA, Genova, Italy, 2006, pp. 2170–2175.

[43] S. Cinková, E. Fučíková, J. Šindlerová, J. Hajič, EngVallex - English Valency Lexicon 2.0, 2021. URL: http://hdl.handle.net/11234/1-3526, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[44] M. Passarotti, B. G. Saavedra, C. Onambele, Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin, in: Proceedings LREC 2016, ELRA, Portorož, Slovenia, 2016, pp. 2599–2606. URL: https://aclanthology.org/L16-1414.

[45] J. Hajič, E. Fučíková, M. Lopatková, Z. Urešová, 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, 2024, pp. 88–100. URL: https://aclanthology.org/2024.dmr-1.10.

[46] J. Hajič, O. Bojar, S. Cinková, R. Sudarikov, Z. Urešová, M. Popel, O. Dušek, Tectogrammatical to AMR conversion: current status, unpublished, 2014. URL: https://www.clsp.jhu.edu/workshops/14-workshop/, JHU Summer Workshop 2014 project.

[47] O. Bojar, S. Cinková, O. Dušek, T. O'Gorman, M. Popel, R. Sudarikov, Z. Urešová, Tecto to AMR and translation, unpublished, 2014. URL: https://www.clsp.jhu.edu/workshops/14-workshop/, JHU Summer Workshop 2014 project.