Beyond the UMR Parsing Shared-Task Rankings: A Closer Look at Czech Results

Markéta Lopatková*, Hana Hledíková, Daniel Zeman and Jan Štěpánek

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské náměstí 25, Prague, Czechia

Abstract

Uniform Meaning Representation (UMR) is a cross-linguistic semantic formalism that captures both sentence-level meaning and document-level phenomena. The First Shared Task on UMR Parsing (2026) yielded two participating systems, ORANGE and SEMA. In this paper, we examine their performance on Czech data in greater detail.

Using the official shared-task results, we analyze the relationship between overall scores, node–token alignment quality, and the quality of generated UMR structures. Our analysis shows that relatively low official scores are strongly influenced by incomplete or inaccurate alignment, which limits the mapping between system-generated and gold-standard graphs during the evaluation. To measure this effect, we manually corrected the predicted alignments for a subset of Czech test set and recomputed the scores. The improvement indicates that, in the case of Czech, the underlying sentence-level graphs are substantially more accurate than the original rankings suggest.

We also provide a qualitative analysis of the strengths and weaknesses of both systems and compare their outputs with UMR structures automatically converted from the Prague Dependency Treebank. Our findings suggest that, for Czech, alignment prediction remains a major bottleneck for both participating systems and that they already produce semantic structures that constitute a promising basis for further development.

Keywords

UMR parsing, sentence-level graph, document-level graph, node–token alignment, quantitative evaluation, qualitative evaluation, alignment as evaluation bottleneck

1. Uniform Meaning Representation (UMR) and UMR parsing

UMR. UMR is a graph-based semantic formalism designed to represent the meaning of natural language texts, with a focus on cross-linguistic applicability [1, 2]. The UMR annotation consists of three layers. The first layer is the *sentence-level semantic graph*, where individual concepts (events, entities, properties) are represented as nodes. The semantic relationships between these concepts, particularly participant and non-participant roles, are depicted as graph edges. Additionally, nodes can be described by attributes and their values, such as, e.g., those related to aspect, quantification, or polarity. To facilitate automatic data processing, each graph node is linked to the surface tokens it represents, forming an *alignment block*. Finally, the third layer, known as *document-level annotation*, includes temporal relations between events, modal information (indicating whether an event is asserted, negated, hypothetical, believed, etc.), and coreference. An example of UMR annotation is illustrated in Example 1 and Figure 1.

Example 1.

context: *Albert Reynolds byl informován o příměří vyhlášeném IRA; následně o tomto příměří* (= variable s16p2) *zpravil* (= s16z2) *britského premiéra* (= s16p5)

'Albert Reynolds was informed on a truce declared by IRA; then he notified (= variable s16z2) the British prime minister (= s16p5) of the truce (= s16p2)'

Sentence: *John Major prohlásil, že to je nejlepší novinka, kterou slyšel za poslední tři roky.*

Sentence Gloss (en): John Major stated that it was the best news he had heard in the last three years.

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia

*Corresponding author.

lopatkova@ufal.mff.cuni.cz (M. Lopatková)

0000-0002-3833-9611 (M. Lopatková); 0000-0002-7970-1170 (H. Hledíková); 0000-0002-5791-6568 (D. Zeman); 0009-0000-3101-573X (J. Štěpánek)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4323/paper18.pdf

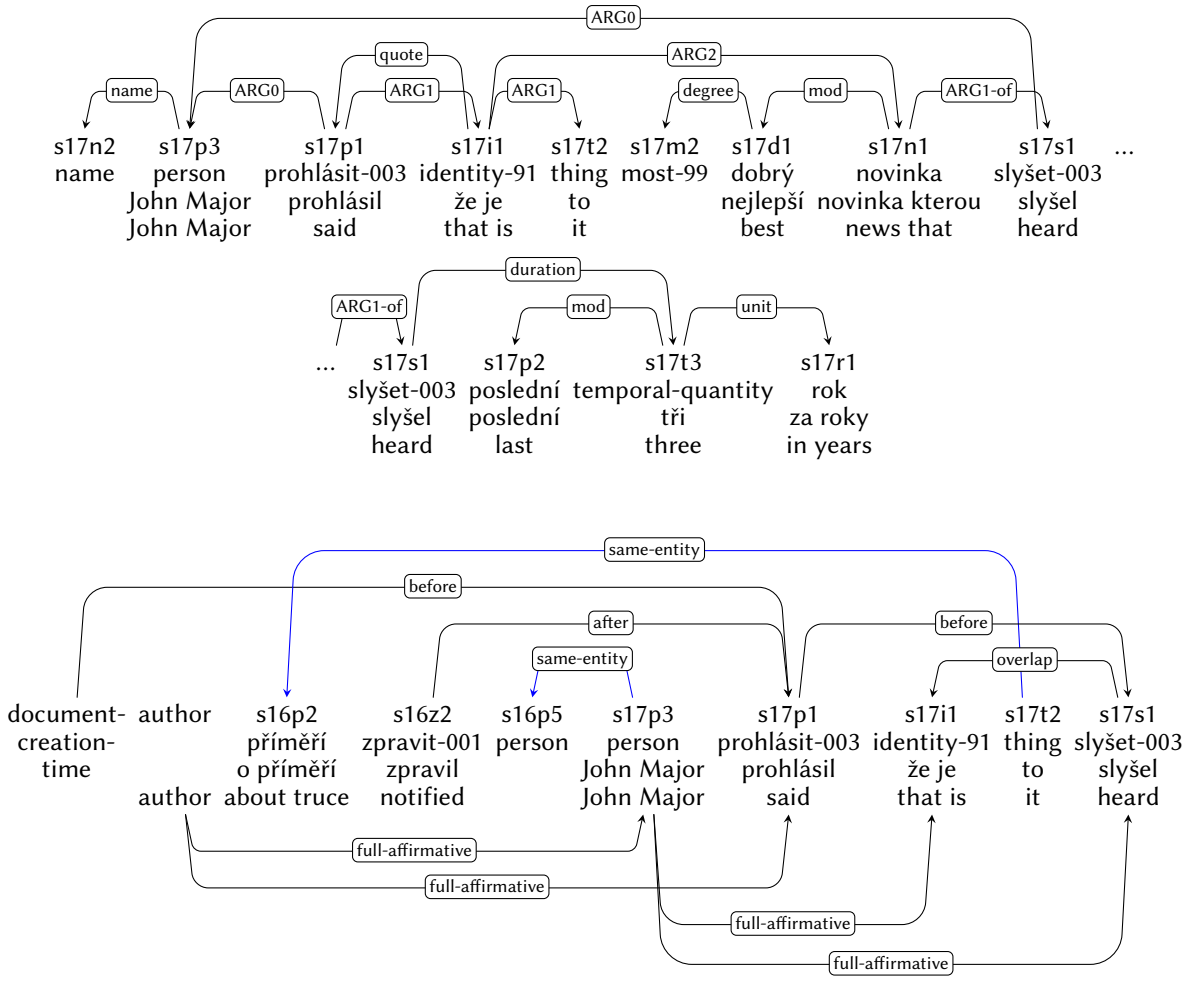


Figure 1: Graph representation of Example 1. Sentence-level graph (the upper part) is split to two parts due to space limit (marked with '...'). Document-level graph (the bottom part) includes relevant nodes from the previous sentence (e.g., variables 's16p5' and 's17p3' refer to the same entity, the British prime minister John Major, see one of the blue arrows). For each node, we give its variable, concept, aligned Czech words and their English glosses. Node attributes (e.g., `:aspect` of events) are not shown.

UMR Parsing. Research on UMR parsing is rather limited – until recently, only a few attempts have been published to develop a UMR parser, focusing on English [3, 4] and Chinese [5].

The situation changed with the *First Shared Task on UMR Parsing*[1] launched in January 2026 to attract researchers to focus on this field [6]. The task consisted in enriching plain texts in seven languages from four different language families with UMR annotation: sentence-level graphs, node–token alignment blocks, and document-level relationships. While the shared-task organizers provided (at least some) annotated training data for six languages (collecting all available UMR-related datasets for these languages), there was no training data for one language; thus, a zero-shot scenario was foreseen.

The multi-language task setup reflects the focus of UMR on cross-linguistic applicability, and has a practical implication: it can, at least partially, address the lack of large monolingual UMR training data.

Two parsers emerged from the Shared Task, ORANGE [7] and SEMA [8]. In this paper, we thoroughly analyze their outputs on Czech data and compare their performance with the PDT-to-UMR system, the automatic conversion from the richly annotated Prague Dependency Treebank [9, 10].

The paper is structured as follows: In Section 2 we present the data discussed in this paper (2.1) and briefly introduce the two parsers from the Shared Task, along with the PDT-to-UMR conversion (2.2).

[1] https://ufal.mff.cuni.cz/umr-parsing

Section 3 first introduces the *ju:mæʧ* metric (3.1) and addresses the performance of the systems on the original Czech shared-task test set (3.2). Then it presents a new experiment with corrected node–token alignment, providing both quantitative results (3.3) and qualitative analysis (3.4). Section 4 briefly summarizes our findings.

2. Data and systems overview

This paper makes use of the data and tools developed for the *First Shared Task on UMR Parsing*[2] [6]. The complete dataset prepared for the task, along with the outputs of the two competing systems, has been released through the Lindat repository[3] [11].

2.1. Data

Training data. The UMR data available for training includes six languages: Arapaho, Chinese, Czech, English, Latin, and Navajo. The datasets are derived from two UMR releases, UMR 2.0, which covers Arapaho, Chinese, English, and Navajo [12], and UMR 2.1, which covers Czech and Latin [13]. For the Shared Task, the data was slightly adjusted as needed to conform to the newly introduced UMR format specification (in addition, for Czech, so-called substitute lemmata from the source corpus were replaced with the corresponding UMR concepts). Table 1 displays the data statistics, with two types of training data available for use.

First, *dirty data* refers to larger datasets, particularly in the case of Czech and English, which are, however, imperfect or incomplete. These datasets are typically (semi-)automatically converted from their source corpora, as described in [6]. They vary in the relations and attributes annotated at the sentence-level. Moreover, node–token alignment and document-level annotation are typically incomplete or even missing.

Second, *clean data* consists of manually annotated datasets of reasonable quality that generally comply with the UMR specification[4] although certain phenomena may be left out of the annotation. These datasets are fairly similar to the gold-standard test data; however, they are typically considerably smaller than the *dirty data*.

Test data. For our analysis, we utilize two datasets as gold-standard test data. First, the *original shared-task test set* comprises manually created test sets for seven languages, which were prepared for the Shared Task. These include test data for the languages with training data, along with an Italian test set. Detailed statistics can be found in Table 1.

Second, the *Czech PDT test set* is a subset of the Czech data from the original shared-task test set, consisting of four files (120 sentences containing 2,194 tokens).[5] These files consist of data from the PDT-C corpus [14, 15] and cover the following genres:

- Two original Czech documents comprising general journalistic texts (71 sentences);
- One document professionally translated from English, comprising business and finance news (24 sentences);
- One file with spontaneous speech, a part of a testimony, originally recorded for the Shoah Memory Project[6] (25 sentences).

The UMR manual annotation includes both sentence-level and document-level annotation and provides node–token alignment. However, the sentence-level graphs omit certain UMR attributes (esp.

[2] https://ufal.mff.cuni.cz/umr-parsing

[3] http://hdl.handle.net/11234/1-6132

[4] https://umr4nlp.github.io/web/UMRSchemaPages/index.html

[5] For this analysis, we exclude the shared-task test file containing 100 sentences from the Parallel Universal Dependencies corpus (PUD), as its UMR annotation slightly differs from that in the other test files (esp. in the relations for predicate-argument structure). More importantly, our aim is to compare the output of the two available parsers with the automatic PDT-to-UMR conversion, which is not available for this file.

[6] https://www.ushmm.org/online/world-memory-project/, https://vha.usc.edu/home

Table 1
Statistics of the training and testing data used in the Shared Task; the table is taken from [6].

	Training data						Testing data		
	Clean			Dirty					
	Files	Sents.	Tokens	Files	Sents.	Tokens	Files	Sents.	Tokens
Arapaho	1	53	274	2	292	1,627	2	55	274
Chinese	20	557	15,375	1	1,435	18,656	15	236	6,467
Czech	6	103	1,587	6,516	159,906	2,537,317	5	220	4,048
English	4	180	2,023	570	29,872	292,023	5	195	4,092
Italian	–	–	–	–	–	–	1	100	2,212
Latin	–	–	–	3	1,049	19,139	1	50	889
Navajo	1	5	60	3	337	2,663	1	163	1,194

the `:wiki` attribute that links named entities to entries in the WikiData knowledge base). Further, the document-level graphs cover only coreferential relations, while modal relations are (partially) covered at the sentence-level (attribute `:modal-strength`); temporal annotation is completely missing.

There is no overlap between the training data (whether dirty or clean) and the test data.

2.2. Systems

This section briefly introduces the two parsers from the Shared Task, along with the PDT-to-UMR conversion.

ORANGE. The ORANGE team tackles UMR parsing using a two-stage multilingual pipeline designed to cope with very limited annotated data [7]. First, they fine-tuned two monolingual and one multilingual T5-family models to generate an enriched *UMR graph* from each sentence. Then the best model per language was selected – for Czech, the monolingual mT5 model yielded the best results (in fact, the authors report the best results for Czech across all 7 languages on their development set). For *document-level parsing*, a separate model was trained (Qwen3 8B, with 3 setups). Unfortunately, no inter-sentence annotation was generated for the original Czech shared-task test set (probably due to the scarcity of the training data). Lastly, a heuristic lemma-to-concept matching strategy based on UDParse is employed to identify *node–token alignment*, with special handling for named entities.

SEMA. The SEMA team [8] characterizes their approach as "multistage LoRA fine-tuning of Qwen3-4B, using capped dirty-data training, continuation on the clean split, a dedicated alignment generation stage, and rule-based post-processing."[7] At inference time, *sentence-* and *document-level* annotation and alignment generation are performed by separate model checkpoints, followed by rule-based post-processing that should repair malformed outputs, convert alignments back to the official format, and ensure evaluator compatibility.

PDT-to-UMR automatic conversion. The PDT-to-UMR conversion is a recursive, rule-based transformation that uses the Prague Dependency Treebank (PDT) annotation to incrementally construct Uniform Meaning Representation (UMR) graphs, see esp. [9, 16, 17]. Thus, instead of building UMR structures from (segmented and tokenized) texts, as required in the Shared Task, the procedure makes use of the rich manual annotation available in PDT (namely PDT-C [14, 15]).

During the conversion, the PDT representation of individual syntactic and semantic constructions (such as coordination, coreference, relative clauses, raising and control verbs) is adjusted to comply

[7]To optimize performance across languages, the dirty training data was limited to 2,000 examples per language. For Czech, this means that only a fraction of the available dirty data was used to train the system.

Table 2
Results computed for the *original Czech shared-task test set*; figures for ORANGE and SEMA are taken from [6] (Tables 2, 7, 8, and 9).

	Shared Task			
	Czech		Average (7 languages)	
	ORANGE	SEMA	ORANGE	SEMA
Overall *ju:mætʃ*	15.87	26.52	19.35	19.43
F_1 of tokens aligned to a node	43.08	49.57	49.16	50.78
Proportion of mapped nodes*	39.33	54.88	45.00	48.06
F_1 on mapped nodes**	52.10	48.72	48.82	39.81

* Computed as $F_1 = 2PR/(P+R)$, where R is the number of gold nodes mapped to system nodes, divided by the total number of gold nodes, and P is the number of system nodes mapped to gold nodes, divided by the total number of system nodes.
** Triples governed by unmapped nodes are ignored.

with the UMR Specification. Furthermore, it relabels deep-syntactic nodes as UMR concepts, maps PDT functors to UMR semantic relations (using verb-specific or default mappings) and enriches the resulting graph with attributes such as `:aspect`, `:polarity`, `:refer-person`, and `:refer-number`. A key principle is that multiple linguistic layers – deep and surface syntax, lexical information, morphology, and coreference – must be considered jointly, since phenomena such as ellipsis, pronouns, and secondary dependencies often interact and require coordinated graph transformations rather than isolated rule application.

Importantly, the output of the PDT-to-UMR automatic conversion served as *Czech dirty training data* during the Shared Task (see Sect. 2.1), which combines the converted training and development data in the official PDT-C split. Note that the clean training data (with manual UMR annotation) is a subset of the PDT-C development data; hence, these files are available both with clean and dirty annotation.

Further, the same PDT-to-UMR procedure is used to process the *Czech PDT test set* in our experiment described in Sect. 3.3.

3. Analysis of the system outputs

3.1. Metric

To evaluate the outputs of the systems, we use the *ju:mætʃ* metric [18, 6], which operates in two steps. First, the corresponding nodes in the system output (system nodes) and the test data (gold nodes) must be identified. The *ju:mætʃ* mapping algorithm relies primarily on the node–token alignment predicted by the systems; for nodes without alignment, those with the same concepts are mapped, too. The *ju:mætʃ* algorithm outputs a symmetric one-to-one mapping of nodes whenever possible.

Second, the F_1 score is calculated on triples of the following types:

- parent node – relation – child node;
- node – attribute – value (concept is also treated as an attribute).

It is important to stress the role of the node–token alignment predicted by the systems for a thorough and fair evaluation of the competing systems. Missing or incorrect alignment (the first step) will inevitably result in lower *ju:mætʃ* scores computed in the second step.[8]

3.2. Overall results on the original Czech shared-task test set

Table 2 presents the scores as published in [6]. The first row provides the overall *ju:mætʃ* score – we can observe the following:

[8]Participants of the Shared Task were explicitly warned that "token-node alignment should not be omitted, as it affects mapping of system and gold nodes, and, consequently, evaluation of all relations and attributes". For more details, refer to Sect. Evaluation, https://ufal.mff.cuni.cz/umr-parsing.

- On average, the systems gain almost the same macro-average scores across 7 languages, with ORANGE reaching 19.35 % and SEMA 19.43 %. These scores indicate rather low performance of both systems.
- SEMA performs much better on Czech, with overall *ju:mætʃ* reaching 26.52 %; in fact, Czech scored the best, as reported in [6] (followed by Chinese with 25.85 %).
- ORANGE provides considerably lower results for Czech (15.87 %); [6] reports the best ORANGE results for Chinese (36.51 %) and the second best for English (22.19 %).
- Unfortunately, we cannot compare the overall performance of the systems on Czech data with the results produced by the PDT-to-UMR automatic conversion, as there is no PDT source annotation for one of the testing files (see footnote 5 in Sect. 2.1).

However, the figures in the second row suggest that the main reason for the unsatisfactory performance might lie only in the suboptimal node–token alignment predicted by the systems:

- The alignment for Czech covers only 43.08 % (ORANGE) and 49.57 % (SEMA) of the tokens.
- The alignment for all 7 languages covers roughly half of the tokens on average; nevertheless, results across languages are uneven (reaching over 70 % for Chinese and English (ORANGE) and Arapaho (SEMA)).

As a result, the proportion of corresponding nodes that have been successfully identified by the *ju:mætʃ* metric is relatively low:

- For Czech, the proportion of mapped SEMA and gold nodes reaches 54.88 %.
- In comparison, the same proportion for ORANGE is only 39.33 % for Czech.
- We observe that the node mapping for SEMA is better in Czech than in all languages on average, while ORANGE for Czech is below the average.

The last row provides information about the performance of the two systems when computed only on system nodes successfully mapped to gold nodes (i.e., when triples governed by unmapped nodes are ignored) – here, the picture changes:

- In this setting, ORANGE scores better both across all 7 languages (+9.01 % over SEMA) and in Czech (here the difference between ORANGE and SEMA is lower, just +3.38 %).
- Interestingly, this setting is the only one in which the performance of both systems for Czech is better than the average performance across all languages (+3.28 % for ORANGE and +8.91 % for SEMA).

3.3. Results with corrected alignment on the Czech PDT test set

New experiment. The analysis of the results discussed in the previous section, in particular the unsatisfactory prediction of the node–token alignment, inspired us to carry out an additional experiment. We know from the system descriptions [7, 8] that both teams approach the node–token alignment subtask in the later phases of the parsing process. Importantly, this information is not used internally in the parsers proper.[9] Therefore, our additional experiment consists in manual correction of the node–token alignment predicted by the two systems and re-calculation of the scores with the corrected alignment.

For this experiment, we used data from the *Czech PDT test set* (a subset of the original Czech test set used in the Shared Task, 4 files with 120 sentences, 2,194 tokens in total, see Sect. 2.1). This choice has several reasons:

- The annotation of the Czech PDT test set perfectly matches the clean training data provided for the Shared Task (while the Czech PUD test set deviates in certain aspects).
- For this test set, the PDT-to-UMR automatic conversion is available. It is thus possible to compare the performance of the shared-task systems (which only need tokenized and sentence-segmented

[9] In ORANGE, only a basic token–node alignment guesser has been developed, which is based on a lemmatizer and subsequent mapping of lemmas to concepts. Within SEMA, node–token alignment is established in the last-but-one stage, followed only by final rule-based post-processing.

Table 3
Results for the systems with the corrected alignment, computed for the *Czech PDT test set*.
While ORANGE and SEMA processed tokenized texts, PDT-to-UMR processed their PDT structures.

	ORANGE		SEMA		PDT-to-UMR
ju:mætʃ	original	corrected	original	corrected	
Overall *ju:mætʃ*	16.85	48.59	25.73	35.57	51.38
F_1 of tokens aligned to a node	43.84	86.23	46.71	76.21	84.31
Proportion of mapped nodes*	40.62	81.53	51.28	65.44	80.70
F_1 on mapped nodes**	56.01	60.52	50.38	53.71	62.16

*, ** Computed as in Table 2.

text on input) with the conversion that makes use of rich manual annotation in the original PDT-C corpus.
- Finally, the size of this test set is smaller, and therefore manageable, given the necessity of time-consuming manual processing (which must be done for each system separately).

Results. Indeed, the manual alignment correction has resulted in substantially better results, as presented in Table 3:[10]
- For ORANGE, the overall *ju:mætʃ* score nearly tripled, reaching 48.59 %, while SEMA improved by approximately by two-thirds, reaching 35.57 %.
- These results indicate that the unsatisfactory alignment had a much stronger impact on the former system, which, after manual alignment correction, emerges as the more successful of the two.

In particular, the changes in the node–token alignment are quantified in the second row in Table 3, while the third row indicates their impact on the proportion of system nodes mapped to gold nodes:
- ORANGE exhibits an almost doubled F_1 of tokens aligned to a node, and the same increase holds for the proportion of mapped nodes.
- For SEMA, the alignment improves by approximately two-thirds, while the proportion of mapped nodes by almost one-third.
- ORANGE achieves both a higher proportion of tokens aligned to a node (by approximately 10 %) and a higher proportion of system nodes mapped to gold nodes (by more than 16 %).

Finally, both systems show a modest increase in F_1 on mapped nodes, with improvements of 4.51 % for ORANGE and 3.33 % for SEMA.

Comparison with PDT-to-UMR scores. The final column of Table 3 reports, for comparison, the performance of the PDT-to-UMR conversion (which reuses manually annotated PDT data). In particular, ORANGE with the manually corrected node–token alignment nearly matches the performance of the rule-based conversion, trailing by less than three percentage points.[11]

3.4. Qualitative analysis of the outputs

We conducted an exploratory manual inspection of one file (a journalistic text, 28 sentences, 550 tokens covered by 439 gold nodes) to identify the strong features of the competing systems on the one hand and the primary failure modes on the other. Table 4 summarizes the scores for the analyzed phenomena.

[10]However, it is important to note that evaluating the outputs using data from a single language (Czech) is somewhat unfair, as the systems are not optimized specifically for that language. Instead, they are designed to achieve the best overall performance across six languages, which necessarily involves trade-offs.

[11]This difference between the ORANGE and PDT-to-UMR scores comes close to the level of scores computed for inter-annotator agreement (IAA); e.g., [9] reports 93 % IAA on mapped nodes and 90 % IAA on all nodes on similar data.

3.4.1. ORANGE

Alignment. Although the original node–token alignment is only partial, the output successfully covers entire sentences with only negligible omissions. To overcome the limitation of the alignment, the scores provided below are counted on the ORANGE output for the given file using the manually corrected alignment.

Sentence-level graph. We can observe that structure of the sentence-level graphs produced by the system nicely mirrors the dirty training data (i.e., the data produced by the PDT-to-UMR automatic conversion). Scores between the ORANGE and PDT-to-UMR outputs confirm this introspective insight (with the overall *ju:mæʧ* being 81.91 % and F_1 on mapped nodes even 88.08 %). The unexpectedly strong structural similarity between the ORANGE and PDT-to-UMRs outputs – which also holds for syntactically complex structures such as coordination, relative clauses, or foreign phrases – deserves further investigation.

In particular, the system was able to learn lemmatization (with occasional errors such as *vyhlašovat* 'declare' –> *vyhazovat* 'throw'). The lemmata serve as (lexicalized) `concepts`, similarly as in the Czech dirty training data. It typically sticks to the surface (as applied in the Czech dirty training data, unlike in the Czech clean training and testing data); we do not spot any strong effort to use, e.g., verbal concepts to denote events in cases of deverbal nouns. For abstract concepts (e.g., person, event, etc.), agreement is lower. This obviously negatively affects the *concept* scores (given in Table 4). The system is also able to determine mutual *relations* between concepts with sufficient quality; again, the outputs reflect rather (deep) syntax than semantics.

Regarding UMR attributes, the system attempts to identify `:aspect` values for appropriate concepts (this attribute belongs only to events); it reaches relatively high precision (68.75 %) but lower recall (48.89 %). However, it completely fails to predict another event-specific attribute, namely `:modal-strength`. The prediction of `:polarity` attribute seems to be easy; ORANGE correctly identifies all 9 of its occurrences (plus one extra for double negation construction, as in *nedovolí si žádnou akci* '(they) do not dare (to do) any action', with both the negated verb form *nedovolí* and the negative form of the pronoun *žádnou*). Scores for the attribute `:refer-number`, capturing the number of entities, are lower than expected (esp. precision reaching 68.70 % on mapped nodes deserves closer examination). The same applies to `:refer-person`, limited to pronouns (with precision just 37.50 %, but here a low number of instances – just 5 occurrences in the gold file – might influence the scores).

Other UMR attributes are either rare in the file under inspection (`:quant`, `:quote`, `:degree`) or even absent (`:wiki`, `:mode`, `:polite`).

Document-level graph. The ORANGE system does not produce any document-level annotation. The *coreference* annotation focuses solely on capturing intra-sentence relations within sentence-level graphs, which appears to be of reasonable quality at first glance. *Modal annotation* is not included, which also affects the partial modal annotation (i.e., annotation within the `:modal-strength` attribute), as noted above. Similarly, there is no *temporal* annotation at the document-level.

3.4.2. SEMA

Alignment. Although the original proportion of aligned nodes was slightly higher than that of the previous system, SEMA outputs often do not cover the entire processed sentence (an extreme case is a sentence with 16 tokens, of which only 2 are covered). Further, the system tends to hallucinate (introducing nodes for tokens not contained in original sentences) or duplicate sentence fragments. For some sentences, it was difficult (or impossible) to reliably correct the alignment since the provided graph structure is not compatible with the meaning of the sentence.

Sentence-level graph. At first glance, SEMA outputs reveal that the system struggles to generate adequate sentence-level semantic structures. In addition, input sentences are often processed only

Table 4
Detailed results for the systems with the corrected alignment (computed for the analyzed Czech file).

	ORANGE		SEMA		PDT-to-UMR	
	F_1**	*ju:mætʃ*	F_1**	*ju:mætʃ*	F_1**	*ju:mætʃ*
overall score	59.94	50.10	55.02	31.11	61.55	53.02
concept	69.63	57.58	62.06	33.69	69.63	57.04
aspect	57.14	51.46	69.39	44.93	57.52	51.16
modal-strength	–	–	62.50	40.18	–	–
polarity	94.74	94.74	57.14	47.06	94.12	94.12
refer-number	72.58	68.18	66.67	37.50	76.98	73.12
refer-person	54.55	46.15	42.11	21.05	53.33	47.06

** Triples governed by unmapped nodes are ignored.

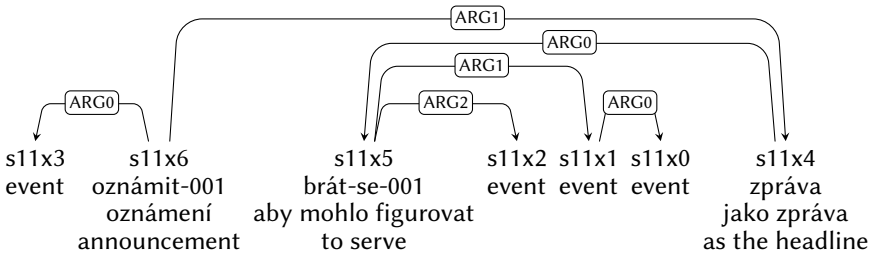


Figure 2: SEMA sentence-level graph of Example 2; to simplify the graph, the attributes `:aspect`, `:modal-strength`, `:refer-number` are not shown. The graph covers only part of the sentence; despite this, the original SEMA generated alignment for the two lexicalized nodes (those with concepts *oznámit-001* and *zpráva*) covers all the sentence tokens. The meaning of the provided UMR structure is hard to interpret.

partially (as noted earlier). These issues are illustrated in Example 2 and Figure 2. As a result of these deficiencies, the overall score is lower, as indicated in the first row of Table 4 (first row).[12] To cope with the effect of structural problems, we report both the overall *ju:mætʃ* scores and the F_1 scores on mapped nodes in this table.

SEMA inclines to represent deverbal nouns using their verbal concepts, e.g., *oznámení* 'announcement' –> *oznámit* 'announce' (for Czech, the same principle is applied in the clean training data, but not in the dirty training data). Although this feature must be evaluated positively, this design is often a source of structural issues in sentence representation, as illustrated in Example 2 and Figure 2.

Example 2.
Sentence: *Oznámení o bezpodmínečném klidu zbraní bylo načasováno tak, aby mohlo figurovat jako hlavní zpráva ...*
Sentence Gloss (en): The announcement of an unconditional ceasefire was timed so as to serve as the main headline ...

We appreciate the SEMA effort to construct structures involving abstract concepts; this can be exemplified by introducing a node with the *person* abstract concept, with the name of the person encoded as sub-roles (compare Example 1). However, not all such occurrences are adequate. In addition, structures with abstract concepts are sometimes recursively embedded; in the extreme case, a node representing the concept *person* with the role *have-rel-role-91* is recursively embedded 57 times (!) under the main predicate in one of the examined sentences.

SEMA also faces minor problems with lemmatization, such as introducing English lemmata (e.g., *author* and *veteran* instead of the Czech *autor* and *veterán*), wrong lemmatization (e.g., *(dát) pokyn*

[12]They also explain a lower impact of the alignment correction for SEMA for the full Czech PDT test set, compared to ORANGE, as displayed in Table 3.

‘(give an) instruction’ –> *pokývat* ‘beckon’) or more obscure cases (e.g., *svatozář* ‘aureole’ –> *svatý* ‘holy’ + *zář* ‘shine’; *pumový* ‘bomb (adj)’ –> *pum* ??). However, when limited to mapped nodes, the identification of concepts reaches reasonable quality (F_1 = 62.06 %).

As for the UMR attributes, SEMA outperforms ORANGE in the `:aspect` prediction, with substantially larger recall than ORANGE (on mapped nodes, SEMA reaches 71.83 % and ORANGE 48.89 %, with almost the same precision 67.11 % and 68.75 %, respectively). SEMA also achieves relatively good quality in `:modal-strength` prediction (reaching 62.50 %), where ORANGE completely fails (no prediction).

For the rest of assessed attributes (`:polarity`, `:refer-person`, `:refer-number`), SEMA falls behind ORANGE even on mapped nodes. Attributes `:quant`, `:quote`, `:degree`, `:wiki`, `:mode`, and `:polite` are either too rare or absent in the processed file.

Document-level graph. SEMA has not produced any inter-sentence *temporal* and *coreference* annotations for the file under inspection (the same is true for ORANGE). Intra-sentence coreference relations are represented within sentence-level graphs; due to the problems with the graphs’ structure (as mentioned above), it is not easy to assess the overall quality of coreference representation.

Regarding *modal* annotation, SEMA has generated – in addition to the `:modal-strength` attribute in the sentence-level graphs (see above) – this type of annotation for just 2 of 28 sentences of the inspected file. This produces duplicate annotation of epistemic modality for 20 events in these sentences (out of 24 events included in the sentence-level graphs for these 2 sentences) and one inappropriate modal annotation for a non-event; when duplicated, modal annotations of individual events match. However, for the majority (26 out of 28 sentences), modal annotation is covered only in sentence-level graphs.

4. Final remarks

This paper presented a detailed analysis of the outputs of the two systems participating in the First Shared Task on UMR Parsing, with a focus on Czech data. Although the official shared-task evaluation suggests that current UMR parsers achieve only modest performance, our results reveal a more nuanced picture.

Our analysis indicates the unsatisfactory node–token alignment as the main bottleneck in the shared-task rankings. Because alignment errors reduce the number of nodes that can be mapped between system and gold graphs, they negatively affect the evaluation of the assessed triplets (parent node – relation – child node or node – attribute – value). When alignments were manually corrected on the Czech PDT test set, the evaluation scores improved substantially, demonstrating that the semantic structures produced by the parsers are often of higher quality than the original scores imply.

The two systems demonstrate complementary strengths. SEMA achieved the highest official score in Czech on the original shared-task test set, while ORANGE consistently produced more coherent sentence-level semantic structures. Furthermore, the comparison with the PDT-to-UMR conversion indicates that the benefits of approaches relying on rich manual linguistic annotation may be smaller than expected. Although their outputs provide necessary training data (as shown by the ORANGE system), our analysis suggests that data-driven parsers have already reached a level at which their outputs can effectively serve as a useful starting point for downstream processing and annotation support. This finding, at least for Czech, contrasts with the conclusion of the ORANGE team [7] that UMR parsing is currently not exploitable due to its poor performance. Our introspective analysis suggests that ORANGE, by producing well-structured and easily interpretable sentence-level graphs, is a promising preprocessing tool for manual UMR annotation for Czech, with the potential to substantially accelerate and streamline the annotation process.

At the same time, our findings identify alignment prediction as one of the key challenges for future UMR parsing research. Improving this component, integrating alignment information more effectively into the parsing pipeline, and extending the analysis to more languages may yield substantial gains without requiring fundamentally different parsing architectures. More broadly, the results highlight the

importance of examining evaluation procedures alongside system rankings when assessing progress in multilingual semantic parsing.

Acknowledgments

The work described herein has been supported by the grants *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic, the Charles University Research Centre program No. 24/SSH/009, and SVV project No. 260 821.

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).

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI's ChatGPT (GPT-5, free access), Grammarly (Edu account provided by the university library), and Overleaf's Writefull model. These tools were used to paraphrase and reword the text, improve writing style, and check grammar and spelling; further, abstract and conclusion were drafted using these tools.

After using these services, the authors reviewed and edited the content as needed and take full responsibility for the content of the article.

References

[1] 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.

[2] J. Bonn, M. J. Buchholz, J. Chun, A. Cowell, W. Croft, L. Denk, S. Ge, J. Hajič, K. Lai, J. H. Martin, S. Myers, A. Palmer, M. Palmer, C. B. Post, J. Pustejovsky, K. Stenzel, H. Sun, Z. Urešová, R. Vallejos, J. E. L. Van Gysel, M. Vigus, N. Xue, J. Zhao, Building a Broad Infrastructure for Uniform Meaning Representations, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 2537–2547. URL: https://aclanthology.org/2024.lrec-main.229/.

[3] J. Chun, N. Xue, Uniform Meaning Representation Parsing as a Pipelined Approach, in: D. Ustalov, Y. Gao, A. Panchenko, E. Tutubalina, I. Nikishina, A. Ramesh, A. Sakhovskiy, R. Usbeck, G. Penn, M. Valentino (Eds.), Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 40–52. URL: https://aclanthology.org/2024.textgraphs-1.3/.

[4] E. Markle, J. G. Bach, S. Wein, SETUP: Sentence-level English-To-Uniform Meaning Representation Parser, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), ELRA, Palma, Mallorca (Spain), 2026, pp. 9710–9721. URL: https://aclanthology.org/2026.lrec-1.762/.

[5] H. Sun, N. Xue, J. Zhao, L. Yue, Y. Sun, K. Xu, J. Wu, Chinese UMR annotation: Can LLMs help?, in: C. Bonial, J. Bonn, J. D. Hwang (Eds.), Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 131–139. URL: https://aclanthology.org/2024.dmr-1.14/.

[6] J. Štěpánek, D. Zeman, M. Lopatková, F. Gamba, H. Hledíková, N. Xue, First Shared Task on UMR Parsing, in: J. Zhao, C. B. Post, E. Hoefer (Eds.), Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, ELRA, Palma, Mallorca (Spain), 2026, pp. 136–147. URL: https://aclanthology.org/2026.dmr-1.12/.

[7] J. Heinecke, M. Asadullah, Orange @ UMR Parsing Shared Task, in: J. Zhao, C. B. Post, E. Hoefer (Eds.), Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, ELRA, Palma, Mallorca (Spain), 2026, pp. 148–154. URL: https://aclanthology.org/2026.dmr-1.13/.

[8] R. de Vergnette, M. Amblard, Sema System for the DMR 2026 Shared Task: Multistage UMR Parsing with Qwen3-4B, in: J. Zhao, C. B. Post, E. Hoefer (Eds.), Proceedings of the Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, ELRA, Palma, Mallorca (Spain), 2026, pp. 155–159. URL: https://aclanthology.org/2026.dmr-1.14/.

[9] J. Štěpánek, D. Zeman, M. Lopatková, F. Gamba, H. Hledíková, Comparing Manual and Automatic UMRs for Czech and Latin, in: Proceedings of the Sixth International Workshop on Designing Meaning Representations (DMR 2025), Association for Computational Linguistics, Stroudsburg, PA, USA, 2025, pp. 1–12. URL: https://aclanthology.org/2025.dmr-1.1/.

[10] M. Lopatková, E. Fučíková, F. Gamba, J. Štěpánek, D. Zeman, Š. Zikánová, Towards a Conversion of the Prague Dependency Treebank Data to the Uniform Meaning Representation, in: L. Ciencialová, M. Holeňa, R. Jajcay, T. Jajcayová, F. Mráz, D. Pardubská, M. Plátek (Eds.), Proceedings of the 24th Conference Information Technologies – Applications and Theory (ITAT 2024), Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia, CEUR-WS.org, Košice, Slovakia, 2024, pp. 62–76. URL: https://ceur-ws.org/Vol-3792/paper7.pdf.

[11] J. Bonn, C. Bonial, M. Buchholz, H.-J. Cheng, A. Chen, C. wen Chen, A. Cowell, W. Croft, L. Denk, A. Elsayed, E. Fučíková, F. Gamba, C. Gomez, J. Hajič, E. Hajičová, J. Havelka, L. Havenmeier, H. Hledíková, A. Kilgore, V. Kolářová, L. Kučová, K. Lai, B. Li, J. Li, M. Lopatková, M. MacGregor, M. Mikulová, J. Mírovský, A. Nedoluzhko, S. Myers, M. Novák, T. O'Gorman, P. Pajas, A. Palmer, M. Palmer, J. Panevová, B. Post, J. Pustejovsky, P. Sgall, J. Song, L. Song, M. Ševčíková, J. Štěpánek, Z. Urešová, H. Sun, Y. Sun, R. Vallejos Yopán, J. Van Gysel, M. Vigus, K. Wright-Bettner, J. Wu, N. Xue, D. Xing, K. Xu, Z. Xu, L. Yue, D. Zeman, J. Zhao, Šárka Zikánová, Z. Žabokrtský, Uniform Meaning Representation 2.2, 2026. URL: http://hdl.handle.net/11234/1-6132.

[12] J. Bonn, C. Bonial, M. Buchholz, H.-J. Cheng, A. Chen, C. Chen, A. Cowell, W. Croft, L. Denk, A. Elsayed, E. Fučíková, F. Gamba, C. Gomez, J. Hajič, E. Hajičová, J. Havelka, L. Havenmeier, A. Kilgore, V. Kolářová, L. Kučová, K. Lai, B. Li, J. Li, M. Lopatková, M. MacGregor, M. Mikulová, J. Mírovský, A. Nedoluzhko, S. Myers, M. Novák, T. O'Gorman, P. Pajas, A. Palmer, M. Palmer, J. Panevová, B. Post, J. Pustejovsky, P. Sgall, J. Song, L. Song, M. Ševčíková, J. Štěpánek, Z. Urešová, H. Sun, Y. Sun, R. Vallejos Yopán, J. Van Gysel, M. Vigus, K. Wright-Bettner, J. Wu, N. Xue, D. Xing, K. Xu, Z. Xu, L. Yue, D. Zeman, J. Zhao, Š. Zikánová, Z. Žabokrtský, Uniform Meaning Representation 2.0, 2025. URL: http://hdl.handle.net/11234/1-5902, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[13] J. Štěpánek, M. Lopatková, D. Zeman, F. Gamba, H. Hledíková, E. Fučíková, M. Novák, Š. Zikánová, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Mikulová, J. Mírovský, A. Nedoluzhko, P. Pajas, J. Panevová, P. Sgall, M. Ševčíková, Z. Urešová, Z. Žabokrtský, J. Hajič, Uniform Meaning Representation 2.1 (Czech and Latin), 2025. URL: http://hdl.handle.net/11234/1-5951.

[14] 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 2.0 (PDT-C 2.0), 2024. URL: http://hdl.handle.net/11234/1-5813, LINDAT/CLARIAH-CZ Digital Library, ÚFAL, MFF UK, Prague, Czechia.

[15] J. Hajič, E. Bejček, J. Hlaváčová, M. Mikulová, M. Straka, J. Štěpánek, B. Štěpánková, Prague Dependency Treebank - Consolidated 1.0, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference

(LREC 2020), European Language Resources Association, Marseille, France, 2020, pp. 5208–5218. URL: https://aclanthology.org/2020.lrec-1.641/.

[16] M. Lopatková, H. Hledíková, J. Štěpánek, D. Zeman, From the Prague Dependency Treebank to the Uniform Meaning Representation: Gold-Standard Czech UMR Data and Partial Automatic Conversion, in: L. Ciencialová, M. Čechová, L. Denti, M. Holeňa, F. Mráz, R. Ostertág, D. Pardubská, M. Plátek, M. Stanek (Eds.), Proceedings of the 25th Conference Information Technologies – Applications and Theory (ITAT 2025), Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia, CEUR-WS.org, Košice, Slovakia, 2025, pp. 179–190. URL: https://ceur-ws.org/Vol-4092/paper14.pdf.

[17] M. Lopatková, E. Fučíková, F. Gamba, J. Hajič, H. Hledíková, M. Mikulová, M. Novák, J. Štěpánek, D. Zeman, Š. Zikánová, UMR 2.0 - Czech: Release Notes, Technical Report TR-2025-74, ÚFAL MFF UK, Prague, Czechia, 2025. URL: https://github.com/ufal/UMR/blob/main/papers/2025-data-release-notes-2-0.md.

[18] D. Zeman, F. Gamba, Word Alignment-Based Evaluation of Uniform Meaning Representations, 2026. URL: https://arxiv.org/abs/2603.26401.