First Shared Task on UMR Parsing

Jan Štěpánek*, Daniel Zeman*, Markéta Lopatková*
Federica Gamba*, Hana Hledíková*, Nianwen Xue†
*Charles University, Faculty of Mathematics and Physics, ÚFAL
Prague, Czechia
{stepanek, zeman, lopatkova, gamba, hledikova}@ufal.mff.cuni.cz

†Brandeis University
Waltham, MA, USA
xuen@brandeis.edu

Abstract

The paper presents the first shared task on parsing Uniform Meaning Representation (UMR), a graph-based framework for cross-linguistic semantic annotation of typologically diverse languages. The task requires systems to enrich plain text with sentence-level structure, node–token alignment, and document-level relations. It involves processing data for seven languages from four language families (Indo-European, Sino-Tibetan, Na-Dene, and Algic). Six languages have at least some training data; for one language, data is not available, leading to a zero-shot scenario. The training dataset as well as the gold-standard test set for all seven languages is released and made available for follow-up research. We present the task setup and evaluation methodology, using two graph matching approaches – a traditional, and an alignment-sensitive one, tailored specifically for UMR. Two participating systems are compared, each representing different modeling approaches. Results highlight the challenges of UMR parsing, particularly for alignment prediction and document-level semantics, and reveal substantial variation across languages and annotation conditions.

Keywords: uniform meaning representation, parsing, evaluation, shared task

1. Uniform Meaning Representation

UMR is a graph-based framework that is semantically grounded and designed for cross-linguistic applicability (Van Gysel et al., 2021; Bonn et al., 2024). Based on Abstract Meaning Representation (AMR, Banarescu et al., 2013), it abstracts from surface syntax to represent concepts—such as entities and events—as graph nodes. The relationships between these concepts are captured as graph edges, and their attributes are included in a normalized, language-independent format. Notably, all syntactic variations of a statement are represented uniformly within this framework. These graphs constitute the *sentence-level annotation*.

In addition, UMR provides a comprehensive annotation of epistemic modality, and marks temporal and coreference relations (both intra- and inter-sentence); this forms the *document-level representation*.

Furthermore, to support automatic processing of the data, the graph nodes are aligned with surface tokens; this *node-to-token alignment* forms an additional annotation block for each sentence. Figure 1 shows an example of UMR annotation with alignment and document-level relations.

The goal of the shared task on UMR parsing is to attract researchers to focus on automatic prediction of UMR annotation across typologically diverse languages, which so far remains largely unexplored.

This paper presents the shared task data (Sect. 2) and briefly describes the task settings (Sect. 3). Section 4 first discusses the metrics used to evaluate the competing systems (Sect. 4.1) and then provides a summary of these systems (Sect. 4.2). Finally, the systems' performance is compared (Sect. 4.3). Section 5 summarizes the task findings, highlighting key challenges and outlining directions for future research.

1.1. Previous Work

While recently developed UMR datasets have become available for several languages, providing an essential foundation for cross-linguistic semantic analysis and modeling, the task of automatically parsing raw text into UMR structures remains relatively underexplored, with only limited efforts dedicated to building robust, generalizable UMR parsers.

The first published UMR parsing model for English was introduced by Chun and Xue (2024). This approach uses a pipeline that leverages existing AMR parsers to generate AMR structures, which are then converted into UMR sentence-level graphs using linguistically motivated heuristics. Document-level annotation is learned inde-

Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, pages 136–147

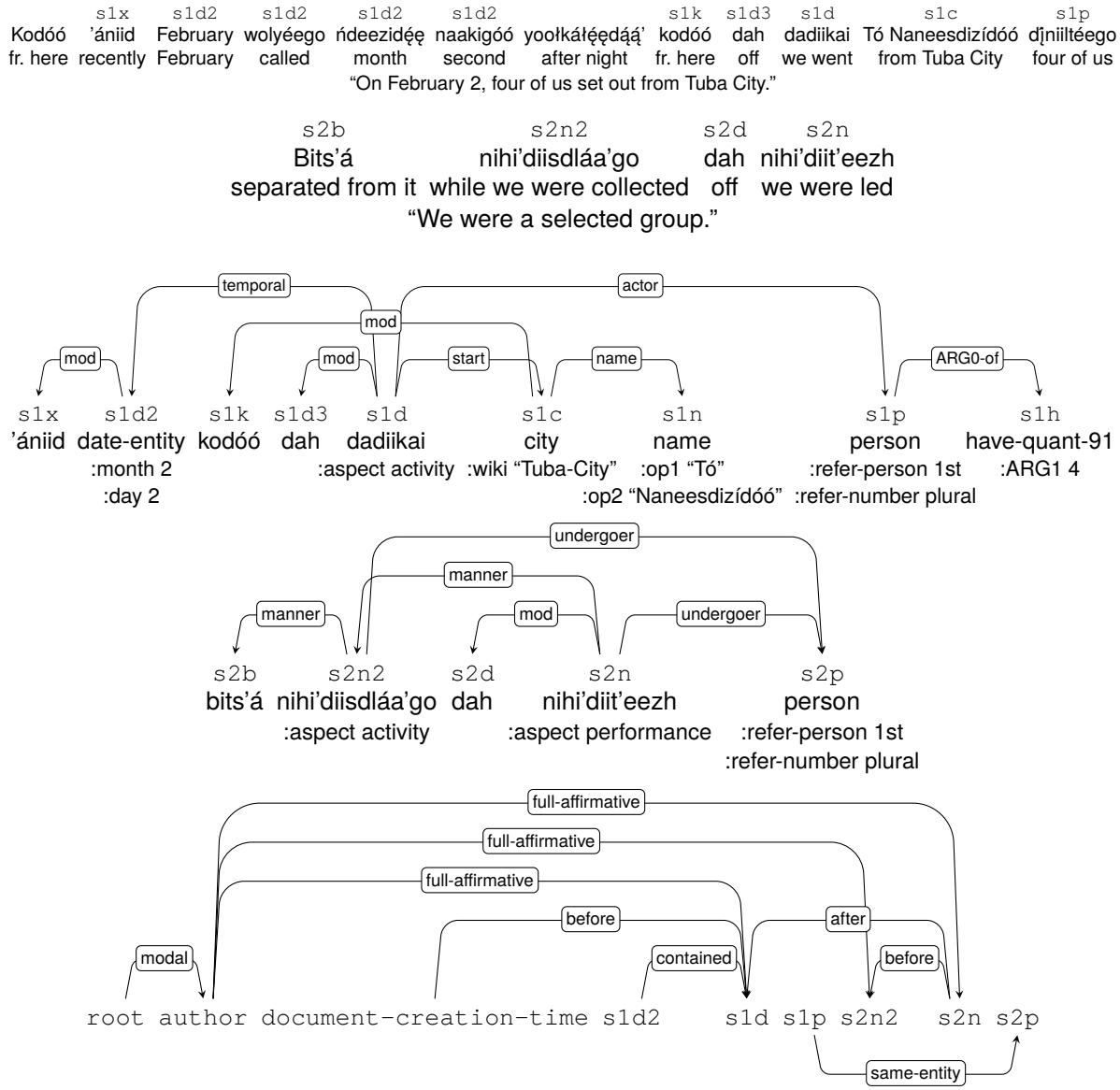

Figure 1: Example of two sentences from the Navajo test data with node–token alignments, sentence-level graphs, and document-level relations.

pendently from the sentence graphs. It is modeled as a set of triples that consist of relevant sentence tokens and their specific relations. The authors developed separately trained models for each of the three tasks: modal relations, temporal relations, and coreference. This approach helps to address the scarcity of available UMR data at that time. Finally, the tokens in the document-level triples are aligned with the corresponding nodes in the sentence graphs.

Inspired by this approach, Markle et al. (2026) present two methods for English text-to-UMR parsing. The first method also utilizes several existing text-to-AMR parsers. These parsers are directly fine-tuned on UMR data. The second method involves first creating UD trees and then converting these into partial UMR structures (Gamba et al., 2025). Subsequently, a T5 Transformer model is trained to expand these partial graphs into complete UMRs.

Finally, Sun et al. (2024) investigate the performance of a GPT-4 model in generating draft sentence-level UMR structures for Chinese. They use few-shot learning and Think-Aloud prompting to guide GPT-4 to generate UMR sentence-level graphs, then compare the results with manually annotated data. The findings are promising, indicating the accuracy of the GPT-generated UMRs is approximately 10-20 % lower than the inter-annotator agreement.

	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

Table 1: Statistics of the data used in the shared task.

2. The Data

The shared task involves processing data for seven languages. For six of them, at least some training data is available; for one language (Italian), no training data is available, leading to a zero-shot scenario. The overall data statistics are presented in Table 1.

The shared task data is now available in the Lindat repository as the UMR release 2.2 (Bonn et al., 2026).[1]

2.1. Training Data

There are six languages for which training data is available: Arapaho, Chinese, Czech, English, Latin, and Navajo. For Arapaho, Chinese, English, and Navajo, the training data is primarily based on UMR 2.0 (Bonn et al., 2025), though not identical: it has been modified to align with the (newly published) UMR format specification (Section 2.2.4) when necessary (for example, typos end errors in bracketing were fixed, and some empty lines, trailing whitespace, or references to non-existent nodes were removed). The training data for Czech and Latin is based on UMR 2.1 (Štěpánek et al., 2025), however, it includes some improvements (the automatic conversion procedure was improved to replace most of the "substitute" lemmata from PDT by corresponding UMR constructs). For Italian, no training data is available at all.

Two types of data available for training are distinguished, "clean" and "dirty":

Clean data is reasonably similar to gold-standard test data, but they are typically very small (if they exits at all). In majority, they contain all annotation parts and should better conform to the annotation guidelines.

Dirty data is much larger, especially for Czech and English. However, it is imperfect or incomplete in various aspects, e.g., it lacks some or all document-level annotation, sometimes also the node–token alignment is missing. There may be additional relations or concepts not defined in UMR.[2]

All training data was freely available during the shared task, without the need to register or sign a contract.

2.2. Test Data

For Arapaho, English and Navajo, the cleaned part of UMR 2.0 data was split approximately to halves and designated as clean training data and test data, respectively.[3] For Latin, the single manually annotated file from UMR 2.0 was used as test set. For Chinese, 25 clean files were available in UMR 2.0. The first 20 were designated as clean training data, the remaining 5 files were combined with 10 previously unpublished files (see below) to become the test data. For Czech, no files released

[1] http://hdl.handle.net/11234/1-6132
(Besides training and test data from the shared task, the release also includes system outputs, as well as two languages that were not part of the shared task but were previously released in UMR: Sanapana and Kukama.)

[2] In English, manual AMR annotation has been partially converted to UMR. Word alignment and document-level relations are missing. In Chinese, one large file lacks document-level relations. Czech and Latin are conversions from the t-layer of Prague Dependency Treebank, resp., Latin Dependency Treebank. In Arapaho and Navajo, the "dirty" files are simply incomplete manual UMR annotations, missing document-level relations and sometimes also word alignment.

[3] Two files in the English test set, `en-0005.umr` and `en-0007.umr`, turned out to contain identical sentences, although their UMR annotation was not identical, with similarity score just below 82%. This is an undocumented feature of UMR 2.0. As we only became aware of it during the test phase of the shared task, we decided to leave the test data as it was.

in UMR 2.0 and 2.1 were used in the shared task test set.

2.2.1. Data Annotated for the Shared Task

New Chinese data. The Chinese UMR test set comprises five documents drawn from the UMR 2.0 release, supplemented by ten newly annotated documents created specifically for this shared task and not previously released. Consistent with the existing Chinese UMR corpus, all ten new documents are sourced from Wikinews, ensuring stylistic and domain continuity. The annotation process follows a semi-automatic pipeline: each document is first processed using an LLM-based UMR parser (Sun et al., 2024), after which trained annotators perform thorough manual correction and validation to ensure high-quality semantic representations. All documents in the test set include comprehensive sentence-level UMR annotations as well as document-level structures, capturing cross-sentence phenomena such as coreference, temporal relations, and modal dependencies. This design ensures that the dataset supports rigorous evaluation of both local semantic parsing and discourse-level understanding.

New Czech data. Four files with manual UMR annotation have been prepared specifically for the shared task:

- general journalistic genre texts, comprising two original Czech documents (newspaper texts from 1992-1994) and one translated document (Czech translations of one Penn Treebank-WSJ text);
- spoken data (a part of a testimony, originally recorded for the Shoah Memory project).

The annotations include sentence-level graphs, node–token alignment, and a partial document-level representation (limited to coreferential relations, both within and across sentences).

2.2.2. Data from PUD

Finally, the test set was extended with 100 parallel sentences in Czech, English, and Italian, originating in the PUD treebank (Zeman et al., 2017) (genre-wise, they are split 50:50 between online news and Wikipedia). These sentences were manually annotated with UMR in order to evaluate UD-to-UMR conversion (Gamba et al., 2025) and they were not included in previous UMR releases.

2.2.3. Differences in UMR Annotation

Due to the complexity of UMR annotation, not everything is annotated in all the test files to the same extent. Some specifics were already mentioned above; here we summarize them:

- All test sets include node–token alignment.
- The PUD datasets do not have document-level relations. All non-PUD test sets have document-level annotation, but in Czech it is limited to coreference.
- The `:modal-strength` sentence-level relation is annotated in datasets that do not have document-level modal relations, that is in all Czech files and in the PUD files of English and Italian.
- The English PUD file lacks the `:wiki` attribute of named entities. Other English files and other PUD files have this attribute, but it is also missing in Arapaho, Chinese, and non-PUD Czech. English and Navajo data use article titles from English Wikipedia as the value of `:wiki`; Czech, Italian, and Latin use the more portable WikiData identifiers.
- The relations `:actor`, `:undergoer`, `:recipient`, `:experiencer`, `:stimulus`, and `:theme` are only used in so-called Stage 0 annotation of argument roles, i.e., only in Arapaho and Navajo and in the three PUD files. The other test sets use numbered `:ARG`N roles instead. (Note that numbered arguments occur even in Stage 0 annotation because they are used with predefined abstract predicates.)

The shared task evaluation is configured so that systems are not penalized for predicting an attribute or relation that is omitted in the gold data.

2.2.4. UMR File Format

The training data in all languages use the same `.umr` file format, which is a text-based format where each sentence is organized into four annotation blocks, viz. metadata, sentence graph, alignment, and document relations. A similar format was used in previous UMR releases, but without any public formal specification (and more importantly, without validating basic requirements, such as matching brackets in graph encoding). Therefore, we published the format specification,[4] as well as a Python validation script[5] that the participants could use to validate their system output before submitting it.

3. The Task

The participants were given blind test data as the input for their systems. The input text contained no UMR annotation but it was tokenized and segmented to sentences; systems were required to preserve tokenization and segmentation, as this

[4] https://ufal.mff.cuni.cz/umr-parsing/umr-file-format

[5] https://github.com/ufal/umrtools

was necessary for evaluation of the output. Systems were expected to generate sentence-level UMR graphs with nodes aligned to input tokens where appropriate, as well as document-level relations. Omitting some parts of the annotation (e.g., some document-level relations) would be possible but it would be penalized by lower scores. Participants were specifically instructed that node–token alignment is integral part of the annotation and that it plays a crucial role in the evaluation procedure.

Training data was made available at the beginning of the shared task, clearly distinguishing the "clean" and "dirty" subsets (see Section 2.1). Participants had to figure themselves the difference between clean an dirty annotation in each language, and they were given no guidance about what to do with Latin (no clean training data) and Italian (no training data at all). They were not informed which relations and attributes are omitted in individual gold files, only our evaluation procedure was informed not to penalize them for predicting such relations.[6] No restrictions were placed on using external language resources, only the previous UMR releases had to be excluded, as they contain a subset of the data that we now use for testing. We are aware that this subset may have been seen by large language models; nevertheless, given the complexity of the task and scarcity of clean annotated data, we decided to take the risk and allow using pretrained models.

We launched a dedicated virtual machine where system outputs were submitted and immediately evaluated, with no limits on the number of submissions per team. The submission site remains open and can be used to evaluate other systems on the same data, using the same evaluation metric.[7]

The time available for the task was very short, with slightly less than four weeks between releasing the training data and collecting the system outputs; blind test data were made available approximately in the middle of this period.

4. Evaluation

4.1. Evaluation Metrics

Standard approaches to evaluation of semantic graphs consist of two phases:

1. Find mapping between nodes of the corresponding graphs;
2. Using the mapping, compute F_1-score of triples of the following types:
   - parent node – relation – child node[8]
   - node – attribute – value[9]

We use two metrics with different node mapping algorithms. Our main metric is *ju:mætʃ* (Zeman and Gamba, 2026), which takes node–token alignment as the main factor influencing the node-node mapping. For comparison purposes, we also compute *smatch* (Cai and Knight, 2013), although its public implementation[10] can only compare sentence-level graphs. The *ju:mætʃ* evaluation script was publicly available during the shared task.[11]

Smatch will map as many nodes as possible. If one of the graphs has more nodes than the other, remaining nodes will stay unmapped. If the graphs have the same number of nodes, every node will be mapped to a node in the other graph, even if they are clearly unrelated. This may occasionally improve the score when a random attribute occurs in both nodes, but it blurs the interpretation of the score, and any (dis)agreement in attribute values of such nodes is meaningless. Consider the two graphs in Figure 2. Identical concepts and attributes will lead to mapping `x3a-y3a`, `x3i-y3i`, `x3c2-y3c`, `x3s-y3s`. Likewise, `x3p-y3p` will be mapped, despite the mismatch in `:refer-number`. All these mappings are intuitively correct; but we cannot say the same about the remaining nodes. None of `x3u` and `x3u2` will be mapped to `y3u`, which was successfully lemmatized to *utor*, yet the expected concept is utor-03. The values of `:aspect` do not match either, so the only positive point would be the `:ARG0` relation to the person node. This is not enough to enforce the mapping `x3u2-y3u`, as *smatch* also considers whether a node is the 'top node' (root) of the graph, and the 'top' attribute gives a point to `x3c-y3u`. Moreover, the latter mapping allows to earn another point by following the `:ARG1` relation from the top node and mapping `x3u-y3a`, although these nodes are also semantically unrelated.

In contrast, *ju:mætʃ* primarily maps nodes aligned to the same word(s), and for nodes without word alignment (assumed to be a minority in UMR graphs) it requires concept identity. As word alignment is part of the annotation, the metric evaluates

[6] Nevertheless, we should have told the participants that the Czech and English PUD files differed from other Czech resp. English files in the relations they use for predicate-argument structure, and we failed to do so. This is definitely a lesson for future instances of this task.

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

[8] This includes document-level relations in UMR, although they may connect nodes from different sentences, or even special fixed nodes such as `author` or `document-creation-time`.

[9] In the context of UMR, the link between a node and its concept string is a special kind of attribute-value pair.

[10] https://github.com/snowblink14/smatch

[11] Now it is available together with the validator in https://github.com/ufal/umrtools.

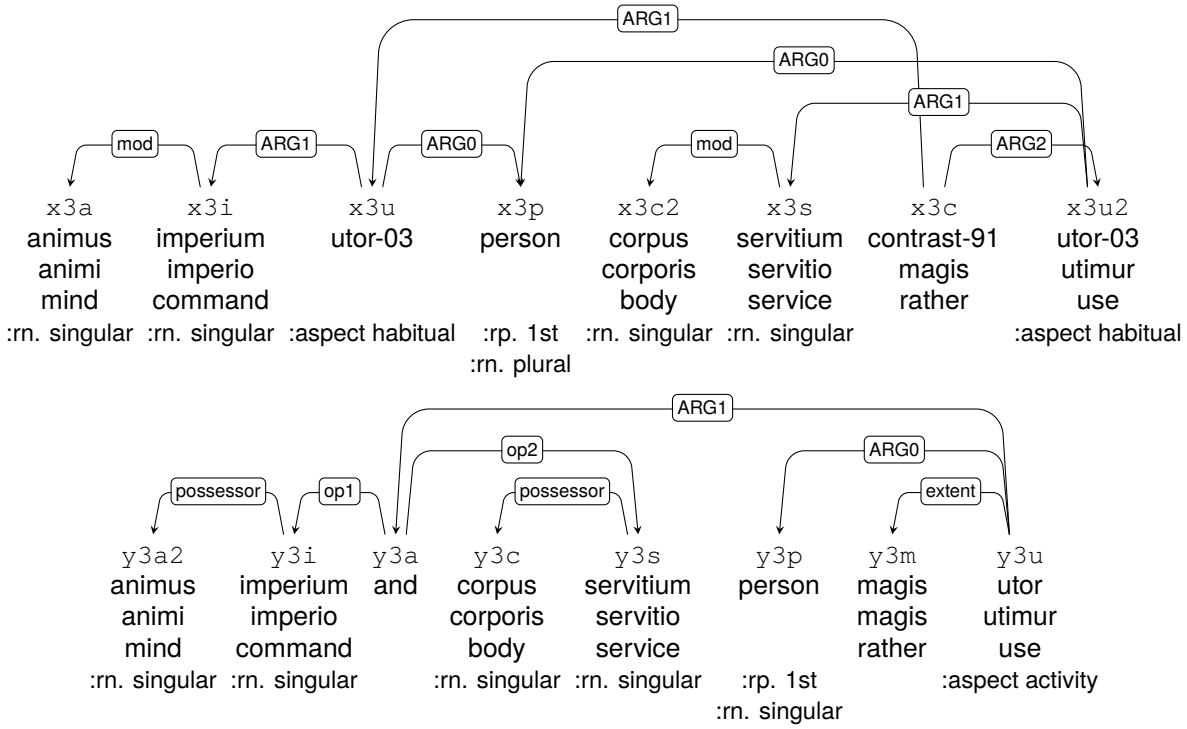

Figure 2: Competing annotations of Latin *animi imperio, corporis servitio magis utimur* "We use the command of the mind rather than the service of the body." In this case, each node in the two graphs is aligned to at most one token and the aligned tokens are shown below node ids and concept strings (lemmas). Abbreviated attributes ":rn." = ":refer-number", ":rp." = ":refer-person".

it, too, though only indirectly. Missing or nonsensical alignments will lead to suboptimal node mapping and thus lower scores. On the other hand, using imperfect alignment for node matching is less straightforward than it may seem. A node may be aligned to multiple words;[12] alignments of nodes from two annotations of the same sentence may overlap instead of being identical. Overlapping alignments have to be resolved so that at most one node from either side is retained. We need symmetric one-to-one node mappings—not only because it simplifies subsequent comparison of node properties, but also because it follows the intuition that both nodes were intended to represent the same concept in the meaning structure of the sentence.

The symmetrization works as follows. Whenever a node on either side is mapped to multiple nodes on the other side, mappings are gradually removed until just one target node remains. When removing target nodes, we try to remove those that are least similar to the source node, according to several criteria. The most important criterion is concept string identity. If it does not lead to unique node mapping, we also consider all attributes of the node and their values (outgoing relations are treated analogously). Next, we also do a "weak" comparison of attributes, where we only consider the presence of an attribute, without requiring identical value of the attribute on both sides. This can help distinguish e.g. eventive concepts (having attributes such as `:aspect`) from entities (having e.g. `:refer-number` or `:name`). Finally, we also prioritize alignment to longer words (trying to avoid relying too much on function words, which tend to be shorter).

Nodes without word alignment are paired when they have the same concept; if there are multiple options, we use the same symmetrization approach as described above. The assumption is that many of them are abstract UMR concepts such as `identity-91`, `have-mod-91` or `name`, hence comparing their concept strings will often point at the correct mapping.[13]

When applied to the example in Figure 2, *ju:mætʃ* will trivially discover the mappings that *smatch* got right: `x3a-y3a`, `x3i-y3i`, `x3c2-y3c`, `x3s-y3s`, `x3p-y3p`. It will also map `x3u2-y3u`

[12]For example, in English *He wants to go*, the node representing the going event may be aligned just to the verb *go*, or to the two-word segment *to go*.

[13]Specifically for the `name` concept, we enhance it with the real name from its `:opN` attributes before comparing it to other nodes, e.g., `name["United" "States"]`. This helps in sentences with multiple named entities.

because both are aligned to the verb form *utimur* "we use"; and `x3c-y3m` because both correspond to *magis* "rather". The nodes `x3u` and `y3a` will stay unaligned.

4.2. Summary of Competing Systems

Twelve people expressed interest through our registration form. In the end, the shared task received two submissions, labeled as "orange" (Heinecke and Asadullah, 2026) and "sema" (de Vergnette and Amblard, 2026).

The team "orange" used a two step approach for intra-sentence and inter-sentence information, but merging the document-level annotation with the UMR graph where the relations didn't reach to a different sentence. For the first step, they fine-tuned 3 different models (monolingual Flan-T5 and mT5, and multilingual mT5) and selected the best for each language based on the development data. For the second step, the team used Qwen3 8B with sentence pairs not farther apart than 6, which seemed to cover most of the inter-sentence relations.

The team "sema" used parameter-efficient fine-tuning of a small LLM (Qwen 4B) in three stages:

1. Training on the sentence-level graphs only on dirty data (with a limit of sample per language not to invisibilize less endowed languages);
2. Training on the sentence-level graphs on clean data;
3. Further training the system to output alignments, of the form

```
Word1: node1 (or nothing)
Word2: ...
```

   This way, doing word-to-token and not token-to-word alignment, the format was less ambiguous (avoiding node ordering issues) and more easily predictable.

4.3. Results

Table 2 presents the main evaluation with *ju:mætʃ* scores for each language and system, as well as macro-averages over languages.[14] Since the Stage 0 annotation of the PUD data differs from the rest, we also present separate scores for PUD files (Table 3) and non-PUD files (Table 4). Czech and English are the two languages which occur in both tables; for the Orange system, the PUD files were clearly more difficult, but for Sema the results are mixed: PUD is better than non-PUD in Czech, but the opposite holds in English.

Language	Orange	Sema
Arapaho	0.0815	**0.1115**
Chinese	**0.3651**	0.2585
Czech	0.1587	**0.2652**
English	**0.2219**	0.1919
Italian	0.1394	**0.2137**
Latin	0.1724	**0.1918**
Navajo	**0.2155**	0.1273
Average	0.1935	**0.1943**

Table 2: Main *ju:mætʃ* scores per language and system.

Language	Orange	Sema
Czech	0.1445	**0.2764**
English	0.1590	**0.1641**
Italian	0.1394	**0.2137**
Average	0.1476	**0.2181**

Table 3: Separate *ju:mætʃ* for PUD data.

Language	Orange	Sema
Arapaho	0.0815	**0.1115**
Chinese	**0.3651**	0.2585
Czech	0.1685	**0.2573**
English	**0.2734**	0.2146
Latin	0.1724	**0.1918**
Navajo	**0.2155**	0.1273
Average	**0.2127**	0.1935

Table 4: Separate *ju:mætʃ* for non-PUD data.

Table 5 shows evaluation using the *smatch* score. Here the overall numbers are higher for two main reasons: 1. unlike *ju:mætʃ*, *smatch* does not use word alignment to restrict node mapping between the system output and gold standard; 2. *smatch* does not evaluate document-level relations, which proved difficult and were only partially predicted by the participating systems (see also Table 13). At the same time, *smatch* considers a special relation marking the "top node" (root) of the graph, which is typically easy to predict and which was not included in *ju:mætʃ* by default. To compensate for the second point, Table 6 shows modified *ju:mætʃ* for sentence graphs with top nodes.

Since *ju:mætʃ* heavily depends on the system's ability to predict node-word alignment, we also tried to shed some light on that factor. In Table 7, we evaluate word alignment independently of node

[14] The outputs of the two systems are released together with the shared task training and gold standard test data in UMR 2.2 (Bonn et al., 2026).

Language	Orange	Sema
Arapaho	**0.4308**	0.3447
Chinese	**0.5194**	0.4363
Czech	**0.4921**	0.4456
English	**0.4625**	0.4169
Italian	0.3567	**0.3796**
Latin	**0.4446**	0.3879
Navajo	**0.3754**	0.2568
Average	**0.4402**	0.3811

Table 5: *Smatch* scores per language and system.

Language	Orange	Sema
Arapaho	0.1008	**0.1272**
Chinese	**0.3777**	0.2471
Czech	0.1582	**0.2724**
English	**0.2509**	0.2057
Italian	0.1368	**0.2155**
Latin	**0.1950**	0.1846
Navajo	**0.2551**	0.1267
Average	**0.2107**	0.1970

Table 6: *Ju:mætʃ* scores modified to be more comparable with *smatch*, i.e., disregarding document-level relations but counting a special relation for top nodes.

Language	Orange	Sema
Arapaho	0.4078	**0.7368**
Chinese	**0.7074**	0.6122
Czech	0.4308	**0.4957**
English	**0.7148**	0.6457
Italian	0.1873	**0.2539**
Latin	0.5114	**0.5493**
Navajo	**0.4818**	0.2611
Average	0.4916	**0.5078**

Table 7: F_1 score of tokens aligned to a node.

mapping. We consider sets of tokens corresponding to a node in each file, regardless of whether the node mapping algorithm actually mapped such nodes to each other (although it is likely that it did). For example, if there is the phrase *to the city* in the English data, and all three words are included in the gold standard alignment of the same node, we expect the system output to also align all three words to one node. If the system predicts a node representing the `city` concept and decides to align it only to *city* (leaving the preposition and the article unaligned), Table 7 will not count it as matching alignment, although it is still

Language	Orange	Sema
Arapaho	0.2517	**0.5372**
Chinese	**0.6046**	0.4521
Czech	0.3933	**0.5488**
English	**0.5643**	0.4856
Italian	0.3766	**0.5194**
Latin	0.4289	**0.4700**
Navajo	**0.5309**	0.3514
Average	0.4500	**0.4806**

Table 8: Proportion of mapped nodes 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.

Language	Orange	Sema
Arapaho	**0.5570**	0.1742
Chinese	**0.5965**	0.5208
Czech	**0.5210**	0.4872
English	**0.4198**	0.4071
Italian	0.4044	**0.4116**
Latin	**0.4545**	0.4289
Navajo	**0.4641**	0.3568
Average	**0.4882**	0.3981

Table 9: *Ju:mætʃ* scores ignoring triples governed by unmapped nodes.

possible that the node mapping algorithm will map the concept nodes correctly. On the other hand, Table 8 gives the proportion of nodes for which a corresponding node was found in the other file. Note that these scores indicate success in finding *some* mapping; they do not say anything about quality of the mapping. Finally, Table 9 shows the *ju:mætʃ* scores computed solely over triples where the governing nodes were successfully mapped. When contrasted with Table 2, Orange now looks much better than Sema; however, one has to remember that Orange was less successful in reproducing word alignment, thus excluding more potentially difficult nodes from the comparison.

We also offer separate evaluation of a few selected attributes or relations. These are partial *ju:mætʃ* F_1 scores; for nodes that do not have a counterpart in the cross-file node mapping, values of all attributes are automatically wrong. Table 10 shows evaluation of concept prediction, Table 11 evaluates all numbered argument relations (`:ARG`N and `:ARG`N`-of`). It should be noted again that some of the test documents contain Stage 0 UMR annotation, hence they do not use

Language	Orange	Sema
Arapaho	**0.2119**	0.1354
Chinese	**0.5287**	0.3503
Czech	0.2724	**0.3091**
English	**0.3866**	0.2957
Italian	0.2110	**0.2653**
Latin	**0.2541**	0.2337
Navajo	**0.4287**	0.2099
Average	**0.3276**	0.2571

Table 10: Concept *ju:mæʧ*.

Language	Orange	Sema
Arapaho	0.0000	0.0000
Chinese	**0.2270**	0.1369
Czech	0.0231	**0.1010**
English	**0.1490**	0.0777
Italian	0.0089	**0.0113**
Latin	0.0538	**0.0581**
Navajo	0.0000	0.0000
Average	**0.0660**	0.0550

Table 11: *Ju:mæʧ* of `:ARG`N and `:ARG`N`-of` relations.

Language	Orange	Sema
Arapaho	0.0764	**0.2192**
Chinese	0.2828	**0.3153**
Czech	0.0047	**0.3554**
English	0.0031	**0.3072**
Italian	0.0645	**0.2741**
Latin	0.1576	**0.2042**
Navajo	**0.1002**	0.0470
Average	0.0985	**0.2460**

Table 12: *Ju:mæʧ* of `:aspect`.

Language	Orange	Sema
Arapaho	0.0000	**0.3596**
Chinese	0.4261	**0.4634**
English	0.0000	**0.1386**
Latin	0.0000	**0.3236**
Navajo	0.0000	**0.2062**
Average	0.0852	**0.2983**

Table 13: *Ju:mæʧ* of document-level modal relations (not available in Czech and Italian).

such relations with normal verbs; however, all languages may have such relations under abstract

Language	Orange	Sema
Czech	0.0000	**0.3864**
English	0.0000	**0.0195**
Italian	0.0000	**0.0653**
Navajo	**0.0316**	0.0000
Average	0.0079	**0.1178**

Table 14: *Ju:mæʧ* of `:modal-strength` (sentence-level modal annotation, not available in Arapaho, Chinese, and Latin).

predicates such as `have-quant-91`. Table 12 evaluates the `:aspect` attribute of events.

Document-level relations are split to three categories: modal, temporal, and coreferential. None of the two systems scored in coreference, and the only non-zero score for temporal relations was achieved by Orange on the Chinese data ($F_1 = 0.0911$). For modal relations, the scores are presented in Table 13. They do not include Czech and Italian because the gold data in these languages use Stage 0 approach to modal annotation (although Czech is Stage 1 in other aspects, such as argument roles). Instead, these two languages employ the `:modal-strength` attribute in sentence-level graphs (Table 14). The same approach is also taken in one English file (the PUD data, parallel in Czech, English, and Italian). The Navajo dataset sticks out, as it contains both document-level and sentence-level modal relations (the latter partially recovered by the Orange system).

5. Conclusion

We have introduced the first shared task on UMR parsing, establishing a benchmark for evaluating multilingual semantic graph generation. The goal of the shared task is to assess how effectively current systems can construct structured meaning representations that capture the underlying semantics of sentences across diverse languages. The evaluation results demonstrate that UMR parsing remains a challenging problem, particularly with respect to node–token alignment and document-level relations such as modality and temporal structure. While participating systems achieved partial success with sentence-level representations, the overall performance scores indicate significant room for improvement, especially in low-resource and zero-shot settings.

The findings further emphasize the importance of data quality and the completeness of annotations. Performance varied notably across different languages and datasets, suggesting that inconsistencies, incomplete annotations, or language-

specific phenomena can dramatically affect parser effectiveness. These observations highlight the need for more robust modeling approaches and richer annotated resources.

Overall, these findings reveal that significant obstacles remain in UMR parsing, both in terms of model architecture and data availability. Future research should prioritize improving alignment prediction mechanisms, advancing systems' treatment of document-level structures, and leveraging multilingual transfer learning to enhance performance in low-resource contexts.

By establishing this shared task, we hope to stimulate further research in UMR parsing and support the broader goal of building scalable, cross-linguistic meaning representation systems that can facilitate a deeper understanding of natural language semantics across diverse languages and domains.

6. 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, and GAUK No. 104924 of the Charles University.

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).

This work is also supported in part by grants from the CNS Division of National Science Foundation (Awards no: NSF_2213804) entitled "Building a Broad Infrastructure for Uniform Meaning Representations". Any opinions, findings, conclusions or recommendations expressed in this material do not necessarily reflect the views of NSF.

7. Bibliographical References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, Jens E. L. Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Jayeol Chun and Nianwen Xue. 2024. Uniform Meaning Representation Parsing as a Pipelined Approach. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 40–52, Bangkok, Thailand. Association for Computational Linguistics.

Rémi de Vergnette and Maxime Amblard. 2026. Sema system for the DMR 2026 shared task: Multistage UMR parsing with Qwen3-4B. In *Proceedings of the Seventh International Workshop on Designing Meaning Representations*, Palma, Spain. ELRA.

Federica Gamba, Alexis Palmer, and Daniel Zeman. 2025. Bootstrapping UMRs from Universal Dependencies for Scalable Multilingual Annotation. In *Proceedings of the 19th Linguistic Annotation Workshop ((LAW-XIX-2025)*, pages 126–136, Wien, Austria. Association for Computational Linguistics.

Johannes Heinecke and Munshi Asadullah. 2026. Orange @ UMR parsing shared task. In *Proceedings of the Seventh International Workshop on Designing Meaning Representations*, Palma, Spain. ELRA.

Emma Markle, Javier Gutierrez Bach, and Shira Wein. 2026. SETUP: Sentence-level English-To-Uniform Meaning Representation Parser. In *Proceedings of the 2026 International Conference on Language Resources and Evaluation (LREC 2026)*, Palma, Spain. ELRA.

Haibo Sun, Nianwen Xue, Jin Zhao, Liulu Yue, Yao Sun, Keer Xu, and Jiawei Wu. 2024. Chinese UMR annotation: Can LLMs help? In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 131–139, Torino, Italia. ELRA and ICCL.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *Künstliche Intelligenz*, 35(3):343–360.

Daniel Zeman and Federica Gamba. 2026. Word alignment-based evaluation of Uniform Meaning Representations. In *arXiv:2603.26401 [cs.CL]*. arXiv.org.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

8. Language Resource References

Julia Bonn and Claire Bonial and Matt Buchholz and Hsiao-Jung Cheng and Alvin Chen and Ching-wen Chen and Andrew Cowell and William Croft and Lukas Denk and Ahmed Elsayed and Eva Fučíková and Federica Gamba and Carlos Gomez and Jan Hajič and Eva Hajičová and Jiří Havelka and Loden Havenmeier and Ath Kilgore and Veronika Kolářová and Lucie Kučová and Kenneth Lai and Bin Li and Jingyi Li and Markéta Lopatková and Marie MacGregor and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Skatje Myers and Michal Novák and Tim O'Gorman and Petr Pajas and Alexis Palmer and Nartha Palmer and Jarmila Panevová and Benét Post and James Pustejovsky and Petr Sgall and Jialin Song and Li Song and Magda Ševčíková and Jan Štěpánek and Zdeňka Urešová and Haibo Sun and Yao Sun and Rosa Vallejos Yopán and Jens Van Gysel and Meagan Vigus and Kristin Wright-Bettner and Jiawei Wu and Nianwen Xue and Dan Xing and Keer Xu and Zhixing Xu and Liulu Yue and Daniel Zeman and Jin Zhao and Šárka Zikánová and Zdeněk Žabokrtský. 2025. *Uniform Meaning Representation 2.0*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Julia Bonn and Claire Bonial and Matt Buchholz and Hsiao-Jung Cheng and Alvin Chen and Ching-wen Chen and Andrew Cowell and William Croft and Lukas Denk and Ahmed Elsayed and Eva Fučíková and Federica Gamba and Carlos Gomez and Jan Hajič and Eva Hajičová and Jiří Havelka and Loden Havenmeier and Hana Hledíková and Ath Kilgore and Veronika Kolářová and Lucie Kučová and Kenneth Lai and Bin Li and Jingyi Li and Markéta Lopatková and Marie MacGregor and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Skatje Myers and Michal Novák and Tim O'Gorman and Petr Pajas and Alexis Palmer and Martha Palmer and Jarmila Panevová and Benét Post and James Pustejovsky and Petr Sgall and Jialin Song and Li Song and Magda Ševčíková and Jan Štěpánek and Zdeňka Urešová and Haibo Sun and Yao Sun and Rosa Vallejos Yopán and Jens Van Gysel and Meagan Vigus and Kristin Wright-Bettner and Jiawei Wu and Nianwen Xue and Dan Xing and Keer Xu and Zhixing Xu and Liulu Yue and Daniel Zeman and Jin Zhao and Šárka Zikánová and Zdeněk Žabokrtský. 2026. *Uniform Meaning Representation 2.2*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jan Štěpánek and Markéta Lopatková and Daniel Zeman and Federica Gamba and Hana Hledíková and Eva Fučíková and Michal Novák and Šárka Zikánová and Eva Hajičová and Jiří Havelka and Veronika Kolářová and Lucie Kučová and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Petr Pajas and

Jarmila Panevová and Petr Sgall and Magda Ševčíková and Zdeňka Urešová and Zdeněk Žabokrtský and Jan Hajič. 2025. *Uniform Meaning Representation 2.1 (Czech and Latin)*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.