Building a Broad Infrastructure for Uniform Meaning Representations

Julia Bonn[1], Matthew Buchholz[1], Jayeol Chun[3], Andrew Cowell[1], William Croft[2], Lukas Denk[2], Sijia Ge[1], Jan Hajič[4], Kenneth Lai[3], James H. Martin[1], Skatje Myers[1], Alexis Palmer[1], Martha Palmer[1], Claire Benét Post[1], James Pustejovsky,[3] Kristine Stenzel[1], Haibo Sun[3], Zdeňka Urešová, Rosa Vallejos[2], Jens E. L. Van Gysel[2], Meagan Vigus[2], Nianwen Xue[3], Jin Zhao[3]

University of Colorado Boulder [1], University of New Mexico[2],
Brandeis University[3], Charles University[4]

julia.bonn@colorado.edu, matthew.buchholz@colorado.edu, jchun@brandeis.edu,
james.cowell@colorado.edu, wacroft@icloud.com, ldenk@unm.edu, sijia.ge@colorado.edu,
hajic@ufal.mff.cuni.cz klai12@brandeis.edu, james.martin@colorado.edu,
Skatje.Myers@colorado.edu alexis.palmer@colorado.edu martha.palmer@colorado.edu,
benet.post@colorado.edu, jamesp@brandeis.edu, kristine.stenzel@colorado.edu,
hsun@brandeis.edu, uresova@ufal.mff.cuni.cz, rvallejos@unm.edu, mlvigus@gmail.com,
jelvangysel@unm.edu, xuen@brandeis.edu, jinzhao@brandeis.edu

Abstract

This paper reports the first release of the UMR (Uniform Meaning Representation) data set. UMR is a graph-based meaning representation formalism consisting of a sentence-level graph and a document-level graph. The sentence-level graph represents predicate-argument structures, named entities, word senses, and aspectuality of events, as well as person and number information for entities. The document-level graph represents coreferential, temporal, and modal relations that go beyond sentence boundaries. UMR is designed to capture the commonalities and differences across languages; this is done through the use of a common set of abstract concepts, relations, and attributes as well as concrete concepts derived from words from individual languages. This UMR release includes annotations for six languages (Arapaho, Chinese, English, Kukama, Navajo, Sanapaná) that vary greatly in terms of their linguistic properties and resource availability. We also describe on-going efforts to enlarge this data set and extend it to other genres and modalities. We also briefly describe the available infrastructure (UMR annotation guidelines and tools) that others can use to create similar data sets.

Keywords: Uniform Meaning Representation, Graph-based semantic representation, Semantically annotated resources

1. Introduction

This paper reports the first release of the UMR (Uniform Meaning Representation) (Van Gysel et al., 2021) data set consisting of six languages - Chinese, English, Arapaho, Kukama, Navajo, and Sanapaná, with the last four being low resource languages that have quite distinct linguistic properties. UMR is a recent graph-based meaning representation formalism for entire documents that is designed to account for cross-linguistic similarities and differences so that it can be easily applied to languages with diverse typological profiles. Based on Abstract Meaning Representation (AMR) (Banarescu et al., 2013), UMR has a sentence-level representation that not only captures predicate-argument structures, word senses, and named entities as AMR does, but also encodes aspectuality of events and person and number attributes for entities. UMR also has a document-level representation that captures semantic relations that go beyond sentence boundaries. These include entity and event coreference, temporal relations between events and between events and time expressions, and modal dependencies between events and their sources called *conceivers*. Section 2 gives a more detailed description of the Uniform Meaning Representations that are annotated in this release.

In the age of Large Language Models (LLMs) where state-of-the-art systems in NLP are based on black-box neural networks, we believe it is important to continue to develop linguistic resources that can be used to build interpretable and controllable systems for settings where transparency is critical. We believe that the released UMR data set is a step in that direction. We expect that the released UMR data set will be useful for the development of a wide range of applications, including but

Authors are listed alphabetically

LREC-COLING 2024, pages 2537–2547

not limited to Information Extraction, Knowledge-based Question Answering, Human Robot Interaction, and others.

The rest of the paper is organized as follows. Section 2 provides an overview of the UMR formalism. Section 3 describes other semantic annotation resources that are similar to UMR. Section 4 describes each of the six languages included in this release. Section 5 describes the ongoing efforts to enlarge the UMR data set in preparation for future releases. Section 6 describes tools and resources for fellow researchers who are interested in annotating their own UMR data sets. These include the UMR annotation guidelines as well as UMR-Writer, a tool that can be used by researchers to annotate UMR data sets. Finally, we draw our conclusion in Section 7.

2. Overview of UMR

UMR consists of a sentence-level representation that focuses on predicate-argument structures and a document-level representation that captures semantic relations that go beyond sentence boundaries. At the sentence level, in addition to named entities and word senses that are already in AMR, it adds specifications for how to represent aspect, quantification, and scope of events. It also refines how pronouns and multiword expressions (MWEs) (Bonn et al., 2023a) are represented in order to make the representation more cross-linguistically applicable. A sentence-level UMR representation for the sentence "He denied any wrong-doing" is illustrated in 1, where the pronoun "he" is represented in UMR as a *person* concept that has a *ref-person* attribute with value *3rd* and a *ref-number* attribute with value *Singular*.

UMR is incomplete with only the sentence-level representation, as some semantic relations cannot be captured without going beyond sentence boundaries. For instance, even if we break down the meaning of the pronoun "he" into a *person* concept with person and number attributes, we won't know to whom it refers until we can identify its antecedent, which can be found in previous sentences in the document. This is illustrated in Figure 2, which is a document-level UMR for a short document of three sentences. The UMR concept *person* for the pronoun "he" in the last sentence refers to another *person* concept in the first sentence with the name "Edmund Pope".

There are other semantic relations that go beyond the boundaries of sentences. For instance, we know that the *convict-01* and *sentence-01* events from the second sentence happened before the *taste-01* event in the first sentence. UMR captures these temporal relations by identifying *temporal dependencies* between an event or time

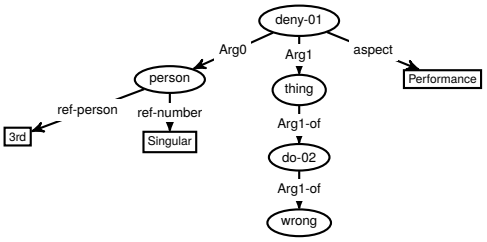

Figure 1: An example sentence-level UMR for the sentence "He denied any wrong-doing." Ovals indicate UMR concepts while rectangles indicate UMR attributes.

expression and its reference event or temporal expression. Furthermore, if we want to assess the factuality of events, we need to represent the level of certainty that their sources (*conceivers* in UMR terminology) assert over these events. These are represented as *modal dependencies* between an event and its source, with a modal strength represented as the relations between them at the UMR document level. For example, in Figure 2, the source of the *deny-01* event is the author *AUTH*, who is affirmative (*:AFF*) that the *deny-01* event occurred, while the source of the *do-02* event is the *person*, who denies (*:NEG*) that the *do-02* event happened.

UMR provides considerable detail about how to represent low-resource languages such as Arapaho and Navajo, which are typologically quite distinct from languages like English. The guidelines address noun incorporation for example (see Example 1), as well as how to handle verbal expressions that are realized as auxiliary verbs in many languages but as affixes in polysynthetic and agglutinating languages. In (1), note that 'horse' is an incorporated noun encoded in the verb stem. The graph predicate includes it, but an additional *animal* node is added to the graph, which will be included in the coreference relations with other instances of the horses in other sentences. The value *animal* comes from UMR's named entity type inventory.

(1) ne'toukutooxebei3i'

ne'-	toukutooxebei -3i'
then-	tie.up.horse -3PL

"Then they tied up their horses."

```
(s53t / toukutooxebei-00
  :actor (s53p / person
       :refer-person 3rd
       :refer-number plural)
  :theme (s53a / animal
       :refer-number plural
       :poss s53p)
  :aspect performance
```

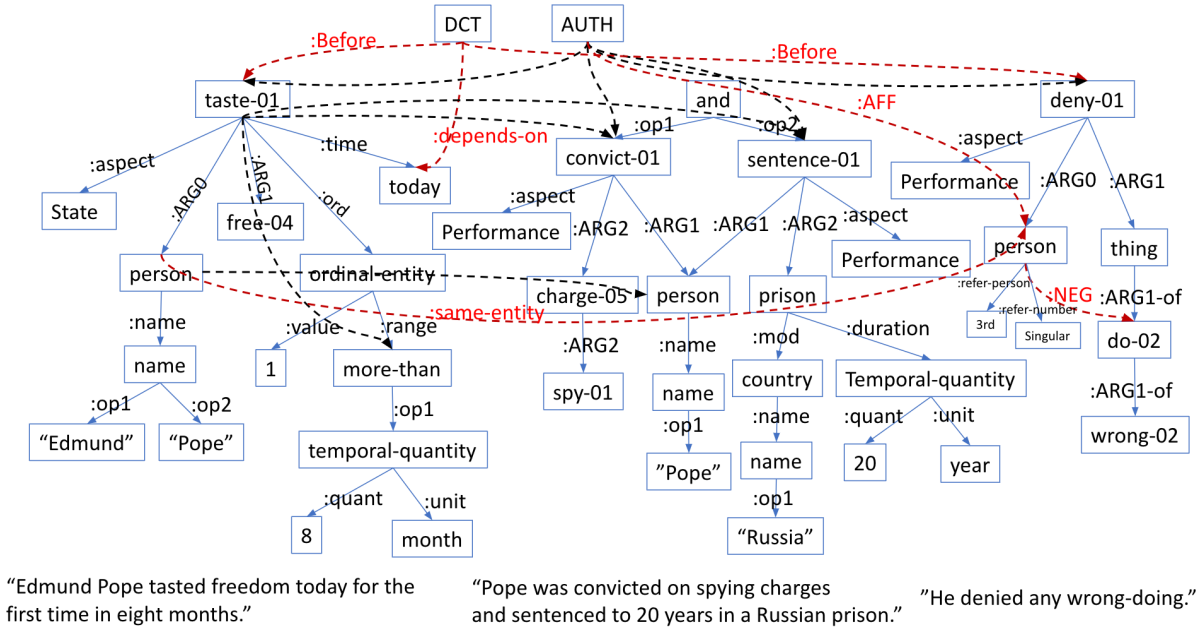

Figure 2: Document-level UMR

:modal-strength full-affirmative)

3. Related work

UMR is the latest in the family of meaning representation formalisms that started with PropBank (Palmer et al., 2005; Pradhan et al., 2022), grew into AMR (Banarescu et al., 2013), and then expanded into a series of special domain- and language-specific AMR adaptations before being unified into the current schema. Whereas PropBank annotations are tied to syntactic structures in a parse tree, AMR abstracts away from syntax by using PropBank rolesets themselves to create a graph structure with labeled nodes and edges. AMR has been quite popular because of the broad semantic coverage afforded by the PropBank rolesets, its ease of use, and its parsability – qualities that also make it a good candidate for expansion and adaptation.

Although it was designed specifically for English, AMR has also been extended to numerous languages. Cross-lingual adaptations have been effective individually, but they diverge more and more from the original schema and from each other, the more the target language differs from English typologically (Wein and Bonn, 2023). These divergences were a major motivation for creating UMR.

One of the most significant extensions to AMR, multisentence AMR (MS-AMR), introduces a layer of annotation of cross-sentence relationships to a document with sentences that have already been annotated with AMR (O'Gorman et al., 2018). These include both identity coreference relations as well as set/member and part/whole bridging relations between entities.

Another extension of AMR, Dial-AMR, allows the annotation of speech acts in dialogue, with a focus on instruction-giving interactions (Bonial et al., 2020). Other extensions include the annotation of multimodal corpora. Spatial AMR (Bonn et al., 2020) adds elements that allow fine-grained, grounded frame of reference tracking, as well as entity grounding, using contextual environmental metadata. Gesture AMR (Brutti et al., 2022) exploits the syntax of AMR to encode both the content and morphology of mainly content-bearing co-speech gesture in multimodal task-oriented dialogues.

Over the years, a number of meaning representation data sets have been developed in other frameworks. For instance, the Groningen Meaning Bank (GMB) (Basile et al., 2012) is a large data set annotated with Discourse Representation Structures (DRS) (Kamp and Reyle, 1993) that makes use of word senses from WordNet (Miller, 1995), semantic roles from VerbNet (Schuler, 2005), and rhetorical relations from SDRT (Asher and Lascarides, 2003). The Parallel Meaning Bank (Abzianidze et al., 2017), built upon the Groningen Meaning Bank, includes annotation of meaning representation for four languages (English, German, Dutch and Italian), but the languages covered are less diverse than those in-

cluded in this first UMR release and do not include any low-resource languages such as Arapaho that have very distinct linguistic properties.

The tectogrammatical layer of the Prague Dependency TreeBank (PDT) (Hajič et al., 2020) covers many of the same semantic distinctions covered by AMR such as argument structure, word senses, coreference, and intra- and inter-sentential discourse relations. Additionally it annotates tense, modalities, and a host of other "semantic" attributes, bridging and textual coreference as well as topic/focus which are not part of UMR annotation. While the PDT uses a multilayered annotation framework where the tectogrammatical layer is explicitly linked to the other layers of linguistic analysis, UMR is an intergrated representation that annotates an entire document as a graph with concepts as nodes and relations as edges between nodes. While the tectogrammatical layer has been used to annotate other languages as in the Prague Czech-English Treebank, like DRS, it has not been extended to a typologically diverse set of languages.

The English Resource Grammar (ERG) (Flickinger, 2000; Flickinger et al., 2016) is a broad coverage, linguistically motivated grammar for English that associates input sentences with semantic representations in the formalism of Minimal Recursion Semantics (MRS) (Copestake et al., 2005). MRS is a sentence-level meaning representation that also focuses on representing predicate-argument structure, sense distinctions where they are grammaticalized, logical semantic phenomena such as quantification and operator-like scopal predicates, and tense, aspect, modality, etc. as determined by morphosyntax. Unlike UMR, ERG only includes sentence-level semantic representations. As MRS emphasizes compositionality, and the semantic representation is typically derived in conjunction with syntactic structures, it cannot be easily extended to other languages, particularly languages that are typologically very different.

Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013) has a foundational layer that focuses on predicate argument structure. The UCCA foundational layer views text as a collection of scenes, and each scene contains a main relation (a state or process) that is the anchor of the scene, as well as participants of the relation. As it currently stands, UCCA does not annotate word senses, named entities, or relations as other meaning representations do, nor does it annotate tense, aspect, modality, or quantification scope. UCCA has also been applied to a number of languages such as English, German, French, Russian, and Hebrew (Abend et al., 2020), but has not been extended to a typologically diverse set of languages.

4. The UMR data set

The first UMR data sets for six languages were released in July 2023 and are available through the UMR website[1] and the LINDAT/CLARIAH-CZ Repository[2] The release includes data from four indigenous languages of the Americas representing no- and low-resource availabilities (Van Gysel et al., 2021). Note that in three cases, the original language data on which the UMR annotations have been done was gathered by the current UMR researchers themselves, working with the communities. The same is true for a planned data set of Quechua annotations. In the Navajo case, the UMR researcher worked with the original Navajo compiler of that data. The following subsections describe each data set and some of the challenges as UMR is expanded to typologically diverse languages.

Language	Sent-level	Doc-level
English	209	202
Chinese	358	358
Arapaho	406	109
Navajo	522	168
Sanapaná	602	602
Kukama	105	86
Total	**2202**	**1525**

Table 1: Data sets for all languages

4.1. English

Current English efforts largely focus on the conversion of the LDC English AMR 3.0 release data (Knight et al., 2021) (cf. Section 5.1), but a small corpus of English UMRs was included in the release for users to see what the finished data for English looks like. The English data set consists of five documents totalling 209 sentences, a combination of Lorelei news text (including one dialogue) and a transcription of a speaker describing a silent film. Document-level graphs are included for all but seven sentences.

While the larger UMR schemata for dialogues and multimodal corpora are still in development (cf. Section 5.3), the English data required strategies for handling several instances under these umbrellas. First, in keeping with Sanapaná's convention in Section 4.5, unattributed speech acts in

[1] https://umr4nlp.github.io/web/

[2] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5198

dialogues are captured with an implicit *say-01* roleset. Second, if a speech transcription includes reference to a gesture that fills a syntactic and semantic argument in the sentence, the gesture is captured within a new *gesture-91* roleset. This allows it to be included in the graph but to be distinguished from spoken elements. For example, in "and he goes .. you know [brushing gesture] and then you see three other boys about his age", the brushing gesture is the *:ARG1* of 'go'. Similarly, onomatopoetic elements are captured in a new *emit-sound-91* roleset that allows them to be connected back to an imitated noise-maker. In "one of them has a .. what do you call those little um paddleball?", "chong chong chong chong", the graph for the second sentence is headed with *emit-sound-91*, which includes a slot for coreference to the paddleball in the first sentence.

4.2. Chinese

The Chinese data in the first release contain 358 wikinews sentences with both sentence- and document-level annotation. UMR annotation of Chinese has a number of challenges due to the linguistic properties of the language. The first one has to do with the fact that, in written form, Chinese does not have natural word boundaries like white space as English does; this makes the job of identifying UMR concepts challenging, as UMR concepts in many languages typically correspond to words. In addition, Chinese is morphology poor, meaning that there are not many explicit morpho-syntactic clues for aspectual attributes and temporal relations. Chinese also tends to have implicit pronominal references and discourse relations. These all contribute to the need to identify abstract concepts, which are always harder to identify than concrete concepts.

One of the primary challenges in Chinese UMR is how to annotate the predicate-argument structure of verb compounds, with compounding being a highly productive feature of Chinese. The key issue is whether to annotate the argument structure of the compound as a whole or to annotate the argument structure of the component verbs and the relations between them. Treating each compound as a whole results in a large number of unique predicates for which rolesets must be defined, whereas annotation of the argument structures of the component verbs significantly reduces the number of rolesets required, but it is not appropriate when a verb compound is idiomatic and should thus be treated as a single predicate. We analyze Chinese verb compounds based on the principle of compositionality, level of grammaticalization, and productivity of their component verbs and classify them into subtypes that require different annotation strategies. For compositional verb compounds, the argument structure of the component verbs as well as the semantic relations between them are annotated. More grammaticalized verb components of compounds are annotated as either attributes of the primary verb or as relations. Some non-compositional verb compounds are annotated as a whole and have new rolesets defined for them (Sun et al., 2023).

4.3. Arapaho

The first release of Arapaho (Algonquian, US) data includes five narrative documents from the Arapaho Text Database (Cowell, 2010) with 408 total sentences. Three of these documents have both sentence and document level graphs (109 sentences), while two have only sentence level graphs. The data are annotated according to UMR stage 0 guidelines, meaning that no formalized rolesets have been created yet. However, some work has been done to conventionalize Arapaho predicates beyond what would normally occur at stage 0. Arapaho is a highly polysynthetic and agglutinating language that uses affixes to express much information that is expressed as separate lexical items in many languages, including tense, aspect, modality, auxiliary-verb-like concepts, person and number, and inclusion of event participants through noun incorporation. While UMR's general suggestion for stage 0 languages is to include the entire inflected/derived form of a verb as the graph predicate, doing so with Arapaho would eventually lead to thousands of complex predicates, most of which would be very unlikely to recur. The challenge of this dataset was determining which semantic information should be encoded as part of a graph predicate and which should not. As of this release, affixes are dropped from the predicate node as long as they can be represented adequately via some other graph structure (arguments, attributes, etc.). This convention will be refined, as developing a roleset lexicon for Arapaho is underway. While there are roleset lexicons for many languages beyond English, the Arapaho lexicon is likely the first for a polysynthetic or agglutinating language, and will require additional guidelines for handling morphologically complex predicates.

4.4. Navajo

Navajo (Athabaskan, USA) is a polysynthetic language, but it can also be classified as fusional. In total, 522 sentences were annotated at the sentence level, with 168 of them also annotated at the document level. These sentences originate from historical narratives from the 1940s and 1950s (Young and Morgan, 1952, 1954). Despite its morphological richness, it was chosen to treat entire

words as heads, rather than morphemes, due to the many verbs where lexical information is discontinuous and fused with inflectional information. One challenge posed by the semantic annotation of Navajo is its frequent use of constructions for proper name references, which make annotators hesitate whether to use the shortcut role for name or annotate the entire 'be called' event with its roles. The annotation also led to the decision to include 'clans' as a category under Named Entity Types.

Although this phenomenon is not specific to Navajo, it was demonstrated that there is also a need to account for purely vocative roles, like *sha'ałchíní* meaning 'my children'. Furthermore, complex descriptions of locomotion, such as *ałchíní bił íbąąs* 'The children arrived (in a wagon)' (the children with them it arrived rolling), sometimes make it challenging to discern between actor, undergoer, and theme. Finally, the existence of postpositions marked for possessors, such as *yas biyi'*, meaning 'inside the snow' (snow its-inside), raises the question of whether a more literal annotation of part-whole or possessor-possessed relationships needs to be considered.

4.5. Sanapaná

The Sanapaná data set (Enlhet-Enenlhet, Paraguay) consists of 602 sentences with annotation at the sentence and document level. The rich agglutinating morphology of this language includes discourse-level patterns used for the expression of grammatical functions in ways that were hard to annotate in current UMR. Among these are constructions expressing "discourse deixis", including frequent anaphoric constructions with an elided argument that is co-referential with a whole section of prior discourse. Others include the frequent absence of reporting verbs in narrations, with speakers simply switching back and forth between stretches of reported speech from different actors in their story. This required the annotation of many 'say'-events and their arguments in the UMR that were not explicitly present in the Sanapaná text – a practice not common previously in English AMR but which has now been adopted for UMR as part of this release.

Further aspects of Sanapaná morphosyntax that prompted reflection about and additions to the UMR guidelines include its fairly rich set of associated motion morphology expressing that an event takes place during or after motion towards or away from the deictic center. An ongoing process of grammaticalization of nominalized verbs into grammatical markers of, amongst others, participant roles and interclausal relations, poses new questions about the event identification section of the UMR guidelines.

4.6. Kukama

A total of 105 sentences in two documents were annotated for Kukama (Tupian), and one of the two documents (86 sentences) has sentence- and document-level graphs. The data come from two traditional stories collected in Kukama territory (Vallejos, 2018). Kukama displays several features that proved challenging to UMR annotation schemas. Such features included the optionality of a variety of major grammatical categories including number and tense-aspect-mood (TAM) marking. These are generally conveyed by positionally fixed clitics, and tense and aspect marking license different word order patterns. However, TAM marking is not obligatory: once the temporal frame of a story is established, they do not appear again, unless it is to manipulate the temporal frame. Similarly, not all nominals which refer to plural entities are grammatically marked as plural. Assessing when plural number needed to be annotated was challenging for the annotators – initially, two UMR experts unfamiliar with the structure of Kukama. Correctly interpreting the multiple types of subordinate clauses that differ in terms of co-reference between arguments was also a challenge, including disambiguating between predicated main verbs and gerunds functioning as e.g. manner adverbials, as well as between main clauses and clausal nominalizations. Although the annotated stories had word-level glosses and sentence-level free translations into English and Spanish, initial annotations by UMR experts had to be reviewed with the language expert in order to capture the semantic details conveyed in the story.

Annotation of Kukama also caused fine-tuning of the UMR guidelines in various areas. The frequent use of causative derivations where English would use unrelated lexemes (e.g. *era* 'be okay' vs. *erata* 'fix', *ikua* 'know' vs. *ikuata* 'notify') prompted further reflection on the annotation conventions for argument structure. This led to the decision that causers of most causativized transitives would be annotated with the :causer participant role, while causatives of mental or cognitive events (e.g. 'cause to see', 'cause to know') would be annotated with the same :actor, :theme, and :recipient participant roles as other transfer events. Similarly, UMR did not have conventions for annotating vocatives which were quite frequent in the Kukama data (see also Section 4.4). UMR has now adopted a *:vocative* role.

4.7. Data distribution and preservation

Open access to language resources and their long-term preservation is a must, especially for valuable resources such as manually annotated data.

The UMR project uses the LINDAT/CLARIAH-CZ[3] repository as the official preservation and distribution channel on top of the current UMR GitHub repository. Data snapshots are packaged to contain all the necessary files, licenses, authorship information, and links to documentation, in order to allow a single point of reference for building tools over the data or expanding them, and to enable reproducibility and verification of any experiments made on the data. A single data item in the repository is a tarball which contains all the languages in which the UMR annotation is available, bearing a single combined license given that each language might be, for various reasons, distributed under a different one. At the moment, UMR 1.0 combines two licenses, one that allows commercial use (for English, CC BY-SA 4.0) and a non-commercial, no-derivatives-allowed license for the other languages (CC BY-NC-ND 4.0). The data itself are in an extended "penman" bracketed format, as known from the AMR annotation scheme and the data. The metadata also contains information about authorship, a short description, information about the size of the dataset in various units, and funding information. The item has received a persistent ID for data citation[4] and it is indexed by Google Data Search, OpenAire, and other scientific and research data-oriented indexes.

5. On-going effort to extend the UMR data set

5.1. English conversion

Conversion of English AMR to UMR data is well underway, with conversion of the AMR data shown in Table 2 to UMR. They consist of the AMR 3.0 release (Knight et al., 2021), the Little Prince Corpus[5], and the Spatial AMR-annotated Minecraft Dialogue Corpus[6] (Bonn et al., 2020; Narayan-Chen et al., 2019). Both AMR 3.0 and Minecraft have a limited number of annotations in the multi-sentence AMR format, which may be able to be converted to UMR document-level graphs.

The conversion process is partially automated. Some tasks, such as one-to-one changes in the named entity hierarchy, can be completely automated. To manipulate the graphs, we use Penman, an open-source Python library (Goodman, 2020).

Other changes in converting AMR to UMR require human judgment. For those, scripts extract the relevant sentences and metadata for manual review by two to three annotators to decide if and how the data is to be changed. Thus, developing infrastructure to expedite conversion has been a concerted focus. Improvements include a scripting framework for the dataset, allowing quick and easy searches and the application of automatic changes, along with the development of tools to aid human annotators in manual review and editing work. To further reduce the need for human annotation, we are also exploring the use of trained neuro-symbolic models for making judgments which are not one-to-one changes. An example of such a change is the splitting of the AMR role `destination` to one of two possible UMR roles, `goal` or `recipient`.

For a more complete description of the changes required to convert AMR graphs to UMR, we refer to previous work outlining this task (Bonn et al., 2023b).

Source	Sentences
AMR 3.0	59,255
Little Prince	1,562
Minecraft	26,221
Total	**87,038**

Table 2: English conversion data set

5.2. Bootstrapping Arapaho UMRs

A second important ongoing effort is the development of tools for bootstrapping from existing resources for low-resource languages, such as interlinearized glossed text (IGT) and lexicons, to build UMR graphs semi-automatically. Pathways to build a supporting valency lexicon (i.e., frame files) semi-automatically are being tested with Arapaho data. This process involves defining classes of similar Arapaho verbs (like English VerbNet), copying the frame for an English verb with the desired argument structure, and then using the lexicon and IGT to find the various surface forms of each Arapaho verb stem.

Our initial work (Buchholz et al., 2024) shows that we can successfully identify basic predicate argument structure and some participant information for six different classes of verbs in Arapaho.

The Arapaho data are also being used as a proof-of-concept to explore ways to build UMR graphs automatically. This will allow specification of which IGT glosses correspond to which nodes in a UMR graph. A script will automatically extract and build UMR graphs based on the IGT supplied. These generated graphs will then be importable into UMR-Writer for easy additional refinement. Further testing will include several thousand

[3] https://lindat.cz

[4] http://hdl.handle.net/11234/1-5198

[5] The Little Prince AMR corpus: https://amr.isi.edu/download.html

[6] https://github.com/cu-clear/Spatial-AMR

words of existing data from Quechua texts, some of which have been previously annotated with an earlier version of UMR guidelines. The long-term goal is to abstract away from the specifics of Arapaho or Quechua to develop a system that can be broadly applied to any low-resource language to convert existing interlinearized data in other formats to UMR data. This will lower the barrier to entry for many groups who may be interested in UMR annotation of their data but lack the time or resources to (re)do annotation manually.

5.3. Multimodal extension

We are currently extending the existing Gesture AMR guidelines (Brutti et al., 2022) for multimodal annotation in UMR. Two very different languages and cultures, American English and Arapaho, are used to develop the extension, with Arapaho data coming from the Arapaho Conversational Database[7] in video format (Cowell, 2010). Currently 20 minutes of Arapaho single-speaker data and 15 minutes of multi-speaker interaction have received first-pass annotation in ELAN. Additions in UMR multi-modal annotation will include richer annotation of metaphorical and metonymic gesture, attention to aspect (held or iterative gesture), a need to allow for citation of gesture by a narrator, parallel to quoted speech, and attention to metapragmatic gestures. Arapaho speakers also use locally-based geography extensively in spatial reference ('upstream this way' vs 'upstream that way' for example, rather than 'east' or 'west') which provides a gestural typology notably different from standard American English.

6. The UMR infrastructure

The UMR data set included in this release is annotated by following the UMR guidelines (Van Gysel et al., 2021) and using UMR-Writer (Zhao et al., 2021; Ge et al., 2023) as the annotation tool. Others who are interested can get access to the publicly available infrastructure. In this section we briefly describe the UMR guidelines and UMR-Writer for researchers who might also be interested in perform UMR annotation for additional languages.

6.1. The UMR guidelines

The UMR guidelines start with an overview section that is intended to help the user get an overall grasp of UMR as a representation, using a short document as a running example. It then has a section that provides a detailed description of a sentence-level representation that includes UMR *concepts*, *relations*, and *attributes*. UMR concepts include eventive concepts, named entity types, word senses, quantification and negation scope, as well as discourse relations. It is worth noting that UMR concepts such as discourse relations are *reified* relations that may be represented as UMR relations as well. UMR relations include *participant roles* that are also known as semantic roles elsewhere, and non-participant-role relations that include typifying relations and referring expressions. UMR attributes include aspect, mode, polarity, person and number, and degree. While mode and polarity are inherited from AMR, the others are new in UMR. The final section of the UMR guidelines provides specifics on how to annotate semantic relations that go beyond sentence boundaries, and it includes subsections on how to annotate entity and event coreference, temporal relations, and modal dependencies.

The UMR guidelines go into considerable length in describing concept-word mismatches that are observed in languages across the world to help users consistently handle such cases. These mismatches include cases where the predicate and its argument are in one word, a linguistic phenomenon that is commonly observed in polysynthetic languages like Arapaho. In this case, a single word will map to multiple UMR concepts that form a sub-UMR graph. It is also common to find the opposite case where multiple word tokens map to a single UMR concept. For instance, the English expression "jump on the bandwagon" forms a single UMR concept that is a concatenation of the words called *jump-on-band-wagon*. This is commonly known as a mutliword expression (MWE), and UMR has detailed guidelines on how to map different types of MWEs into UMR concepts (Bonn et al., 2023a).

For users who are already familiar with AMR, the UMR guidelines also have sections that discuss the differences between AMR and UMR annotation (Bonn et al., 2023b; Wein and Bonn, 2023) as well as appendices with useful mappings for AMR→UMR corpus conversion. In some cases the UMR guidelines expand the semantic inventory inherited from AMR to account for new concepts found in new languages and cultures. For instance, the UMR named entity hierarchy has been upgraded to address some neglected semantic areas identified by AMR annotators. These include insights regarding named entity categories not found in large industrialized societies but that may be relevant to speakers of languages in other cultural contexts, such as Navajo entities referring to clans, Arapaho age-grade societies, and Kukama supernatural beings (Van Gysel et al., 2021). These is a handy visual training resource for AMR-trained users to assist in quickly boot-

[7] https://www.elararchive.org/dk0194/

strapping from AMR to UMR annotation.

As languages may not have equal levels of availability of resources, the UMR guidelines allow the flexibility of either using PropBank-style predicate-specific roles that have become familiar if a valency lexicon that defines such predicate-specific roles exists, or generic roles that do not require such roles for languages that do not have such resources. In spite of the flexibility that the UMR guidelines afford, when using them, it is essential to realize that they do not capture all the nuances of each individual language. When applying the UMR guidelines to a specific language, it may still be necessary to further specify how certain peculiarities need to be represented. As has been discussed in Section 4, in UMR annotation of Arapaho, annotators need to decide how to handle the proliferation of complex predicates when mapping them to UMR concepts. For UMR annotation of Chinese, we need to specify how to annotate the predicate-argument structure of Chinese verb compounds. When annotating Sanapaná, there is the issue of elided report verbs that need to be recovered and represented.

6.2. UMR-Writer

Having a suitable annotation interface is critical for a complicated annotation framework like UMR that has a fairly large inventory of abstract concepts, relations, and attributes that span many dimensions of meaning. This cannot be done without an appropriate annotation tool to support the annotation. Fortunately, such a tool already exists and is publicly available. UMR-Writer[8] is a robust annotation tool that can support UMR annotation across languages. As UMR-Writer is a web-based tool, adding a new language is relatively straightforward.

UMR-Writer includes an intuitive click-based interface that is more suitable for novice users and a keyboard-based tool for trained users who prefer a faster annotation speed. The click-based user interface (Zhao et al., 2021) allows annotators to effortlessly select and click concepts and roles from text and menus. For more experienced users, the keyboard based tool enables the input of editing commands to construct UMR graphs, similar to the AMR editor. With a single command, users can add nodes, and the interface provides dynamic displays of roleset frame information with a clear and intuitive user interface (Ge et al., 2023).

In addition, UMR-Writer offers various functionalities to enhance flexibility and efficiency during annotation: building a lexicon during annotation; direct importation of files annotated with the AMR editor; the ability to copy and paste specific graph segments from one graph to another; and a comprehensive search feature, allowing for node, string, and triple searches. Users can further refine searches using a username filter. UMR-Writer also provides an annotation workflow designed for effective project management. Additionally, its administrative permission hierarchy system ensures that different user groups can collaborate efficiently, balancing speed and quality in annotation tasks.

Current testing for UMR-Writer extends to multilingual data, encompassing languages such as Czech and Arabic, as well as IGT formats for under-resourced languages.

[8] https://umr-tool.cs.brandeis.edu/

7. Conclusion

In this paper, we report the first release of the UMR data set, which consists of sentence-level and document-level UMR annotation of six languages, which span high-resource languages like English and Chinese, as well as low-resource languages like Arapaho, Navajo, Kukama, and Sanapaná. We also discuss our ongoing efforts to expand the data set and extend it to additional genres and modalities. For fellow researchers who wish to perform their own UMR annotation, we also describe resources that are available to support such annotation efforts.

Acknowledgements

This work is supported by grants from the CNS Division of National Science Foundation (Awards no: NSF_2213805, NSF_2213804, NSF_IIS 1764048, NSF_1763926 RI) entitled “Building a Broad Infrastructure for Uniform Meaning Representations” and “Developing a Uniform Meaning Representation for Natural Language Processing”, respectively. The work on Czech has been supported by the UMR project No. LUAUS23283 supported by the Czech Ministry of Education, Youth and Sports (MSMT CR). It has used data provided by the LRI LINDAT/CLARIAH-CZ, Projects No. LM2018101 and LM2023062, supported by the MSMT CR.

8. Bibliographical References

Omri Abend, Dotan Dvir, Daniel Hershcovich, Jakob Prange, and Nathan Schneider. 2020. Cross-lingual semantic representation for nlp with ucca. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 1–9.

Omri Abend and Ari Rappoport. 2013. Ucca: A semantics-based grammatical annotation scheme. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 1–12.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

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.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *LREC*, volume 12, pages 3196–3200.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: abstract meaning representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695.

Julia Bonn, Andrew Cowell, Jan Hajic, Alexis Palmer, Martha Palmer, James Pustejovsky, Haibo Sun, Zdenka Uresova, Shira Wein, Nianwen Xue, et al. 2023a. UMR annotation of multiword expressions. In *Proceedings of the 4th International Workshop on Designing Meaning Representations*.

Julia Bonn, Skatje Myers, Jens EL Van Gysel, Lukas Denk, Meagan Vigus, Jin Zhao, Andrew Cowell, William Croft, Jan Hajic, James H Martin, et al. 2023b. Mapping AMR to UMR: Resources for adapting existing corpora for cross-lingual compatibility. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 74–95.

Julia Bonn, Martha Palmer, Jon Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded minecraft corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),*.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract meaning representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583.

Matt Buchholz, Julia Bonn, Claire Benét Post, Andrew Cowell, and Alexis Palmer. 2024. Bootstrapping UMR Annotations for Arapaho from Language Documentation Resources. In *Proceedings of LREC-COLING 2024*.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Dan Flickinger. 2000. On building a more effcient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Dan Flickinger, Emily M Bender, and Woodley Packard. 2016. English resource semantics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5.

Sijia Ge, Jin Zhao, Kristin Wright-Bettner, Skatje Myers, Nianwen Xue, and Martha Palmer. 2023. UMR-Writer 2.0: Incorporating a new keyboard interface and workflow into UMR-Writer. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 211–219.

Michael Wayne Goodman. 2020. Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319, Online. Association for Computational Linguistics.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Hans Kamp and Uwe Reyle. 1993. *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and*

discourse representation theory. Kluwer, Dordrecht.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christopher D Manning. 2004. Lingo redwoods. *Research on Language and Computation*, 2(4):575–596.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th international conference on computational linguistics*, pages 3693–3702.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-Bettner, and Martha Palmer. 2022. Propbank comes of age—larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Haibo Sun, Yifan Zhu, Jin Zhao, and Nianwen Xue. 2023. UMR annotation of chinese verb compounds and related constructions. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+ NLP, GURT/SyntaxFest 2023)*, pages 75–84.

Rosa Vallejos. 2018. Kukama–Kukamiria. *International Journal of American Linguistics*, 84(S1):S129–S147.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Timothy J. O'Gorman, Andrew Cowell, William Croft, Chu Ren Huang, Jan Hajic, 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*, pages 1–18.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the 4th International Workshop on Designing Meaning Representations*.

Jin Zhao, Nianwen Xue, Jens Van Gysel, and Jinho D Choi. 2021. UMR-Writer: A web application for annotating uniform meaning representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 160–167.

9. Language Resource References

Cowell, Andrew. 2010. *A Conversational Database of the Arapaho Language in Video Format*. Endangered Languages Archive. [link].

Knight, Kevin and Badarau, Bianca and Baranescu, Laura and Bonial, Claire and Bardocz, Madalina and Griffitt, Kira and Hermjakob, Ulf and Marcu, Daniel and Palmer, Martha and O'Gorman, Tim and others. 2021. *Abstract meaning representation (AMR) annotation release 3.0*. Abacus Data Network.

Young, Robert W and Morgan, William. 1952. *The Trouble at Round Rock*. US Department of the Interior, Bureau of Indian Affairs, Division of Education.

Young, Robert W and Morgan, William. 1954. *Navajo Historical Selections: Selected, Edited and Translated from the Navajo*. Department of the Interior, Bureau of Indian Affairs, Branch of Education, 3.