UMR Annotation of Multiword Expressions

Julia Bonn[1], **Andrew Cowell**[1], **Jan Hajič**[3]
Alexis Palmer[1], **Martha Palmer**[1], **James Pustejovsky**[2], **Haibo Sun**[2]
Zdenka Urešová[3], **Shira Wein**[4], **Nianwen Xue**[2], **Jin Zhao**[2]
[1]University of Colorado, Boulder, [2]
Brandeis University [3]Charles University, Prague, [4]Georgetown University
Corresponding author: julia.bonn@colorado.edu

Abstract

Rooted in AMR, Uniform Meaning Representation (UMR) is a graph-based formalism with nodes as concepts and edges as relations between them. When used to represent natural language semantics, UMR maps words in a sentence to concepts in the UMR graph. Multiword expressions (MWEs) pose a particular challenge to UMR annotation because they deviate from the default one-to-one mapping between words and concepts. There are different types of MWEs which require different kinds of annotation that must be specified in guidelines. This paper discusses the specific treatment for each type of MWE in UMR.

1 Introduction

Uniform Meaning Representation (UMR) (Gysel et al., 2021) is a graph-based formalism designed to represent natural language semantics. It is based on Abstract Meaning Representation (AMR) (Banarescu et al., 2013), but is enriched and extended in accordance with typological principles to account for linguistic uniformity and variation across a wide range of languages of the world, from languages like Arabic, Chinese, and English that have a large population of speakers, to languages like Arapaho, Kukama, Navajo, and Sanapana with a relatively small number of speakers. Expanding on AMR, UMR also includes a document-level representation that represents linguistic relations that go beyond sentence-boundaries, such as coreferential relations and temporal and modal dependencies.

Like AMR, the basic building blocks of a UMR graph are concepts and relations, with concepts typically mapping to words in a sentence and relations representing how those words are related semantically. UMR concepts are typically lemmas or sense-disambiguated lemmas, but they can also be abstract concepts that do not map to specific word tokens and are instead inferred from the context of the sentence. Common in a UMR graph are subgraphs that represent predicate-argument structures in which the predicate is the parent and its arguments are its children. The relations between a predicate and its arguments are typically the semantic roles that each argument plays with respect to the predicate, but they can also be other types of semantic relations. A UMR example containing a multiword expression (MWE) is provided in (1):

(1) They are willing to throw America under the bus.

```
(w / will-02
    :aspect State
    :modstr FullAff
    :Arg0 (p / person
        :ref-person 3rd
        :ref-number Plural)
    :Arg1 (t2 / throw-under-bus-08
        :Arg0 p
        :Arg1 (c / country
            :name (n / name
                :op1 "America"))))
```

The mapping between UMR concepts and words in a sentence is complex. While most UMR concepts map to single words, there are also UMR concepts as in (1) that map to multiple words, in this case four words. The opposite is true as well where one word can map to multiple UMR concepts, which is often the case in polysynthetic languages like Arapaho (Gysel et al., 2021).

In this paper, we draw from a broad range of

Proceedings of the 4th International Workshop on Designing Meaning Representations, pages 99–109

annotated examples from different languages to discuss the properties of different types of MWEs and how they can be annotated in UMR. Because UMR annotation occurs in two stages– one called stage 0 for languages without rolesets, and one called stage 1 for languages with rolesets– we discuss differences in the strategies used for MWE annotation at both stages. Here we adopt an operational definition of MWE in the context of UMR annotation. When multiple words map to a single concept in UMR, these words form an MWE. Since they are so common, formulating a consistent approach for representing MWEs in UMR is critical to the success of UMR as a representation. This requires, first of all, that we have a good understanding of what types of MWEs exist in languages of the world, and then design a consistent set of guidelines to direct their annotation in UMR.

The MWEs that we consider in this paper include light verb constructions (LVCs) (Section 2), MWEs without a strong figurative interpretation (Section 3), non-consecutive multiword expressions that occur in certain constructions (Section 4), idioms (Section 5), proverbs (Section 6), and two-part allegorical sayings (Section 6.1). These categories distinguish MWEs by how they are handled in UMR, both in terms of how tokens are incorporated into UMR graphs, and whether/how figurative meaning is conveyed through the UMR schema. We expect that this work will be useful for UMR annotation in the future, and is broadly relevant to studies of abstract meaning in formal semantic representations.

2 Light Verb Constructions

Light verb constructions take the form of a semantically-light verb with a nominal predicate as its object. The arguments are selected by the nominal predicate rather than the light verb, although the light verb can contribute proto-roles as well as aspectual interpretations of the construction as a whole. Following AMR, the UMR concept used to represent the LVC in a graph is derived from the nominal predicate, with the light verb contributing to the aspectual annotation for the predicate. In the English example in (2), the light verb "make" pairs with the nominal predicate "break", which has arguments for an entity in motion (literal or abstract) and a destination. In the UMR graph, the light verb is glossed by the appropriate PropBank sense *break-20*, along with its arguments[1].

(2) The children made a break for the playground.

```
(b / break-20
   :Arg0 (c / child
      :refer-number Plural)
   :Arg2 (p / playground)
   :aspect Performance
   :modstr FullAff)
```

LVCs are also common in Chinese In (3), for example, the light verb 获得 ("get") takes a nominal predicate 认可 ("acceptance") as its object and together they form an LVC in which the nominal predicate selects the arguments and the light verb contributes to a *Performance* aspectual value, meaning the acceptance event has been successfully completed.

(3) 这 一 方法 获得认可 。
this one method get acceptance .

"This method got accepted."

```
(x1 / 认可-01
   :aspect Performance
   :modstr FullAff
   :Arg1 (x2 / 方法
      :mod (x3 / 这)))
```

With the Spanish LVC "dar miedo" (scare, lit. give someone fear), UMR annotation omits the light verb and substitutes the whole construction with the roleset for the verb "asustar," which also means *to scare*[2].

(4) Le di miedo.
him I.gave fear

"I scared him."

```
(a / asustar-01
   :Arg0 (p / person
      :refer-person 1st
      :refer-number Singular)
   :Arg1 (p / person
```

[1] From the English PropBank Lexicon: https://github.com/propbank/

[2] In accordance with Spanish roleset conventions (Wein et al., 2022)

:refer-person 3rd
:refer-number Singular)
:aspect Performance
:modstr FullAff)

In Czech, the same phenomenon exists[3]. Example (5) shows the LVC "vznést připomínku" ("to comment" or "to remind," lit. "raise a reminder"), which is similar to the previous Spanish example in that the LVC can be represented in the UMR graph with a roleset for a synonymous verb "připomenout" (lit. "to remind")[4].

(5) Vznesl poté připomínku.
Raised after-that a-comment.

"He then made a comment."

(p / připomenout-01
:Arg0 (p2 / person
:ref-person 3rd
:ref-number Singular)
:temporal (p3 / poté)
:aspect Performance
:modstr FullAff)

There are also cases in Czech, however, where the LVC is complex and cannot be replaced by a single related verbal roleset without some of the meaning being lost. An example is "projít zkouškou ohněm" (lit. "go_through test [by]fire", experience ordeal by fire). Here, the nominal portion is itself idiomatic. The preferred UMR approach in such cases is to use an MWE predicate reflecting the whole construction.

(6) Prošel zkouškou ohněm.
go_through exam by_fire

"He passed the ordeal by fire."

(z / zkoušet-ohněm-01
:Arg1 (i / individual-person
:ref-person 3rd
:ref-number Singular)
:aspect Performance
:modstr FullAff)

[3]The usual Czech linguistic terminology uses the term "compound [verb] phrases."

[4]In Examples 5 and 6 that since Czech is a pro-drop language, subject personal pronouns are usually dropped because all person and number information is provided on the verb thanks to grammatical agreement rules. However, UMR graphs still provide a node for this argument to enable co-reference.

3 MWEs Without Strong Figurative Interpretation

Some MWEs do not have a strong figurative interpretation, i.e., any figurative interpretation can still be derived from the parts. This category includes everything from fully-fixed MWEs like complex function words (Constant et al., 2017) to semi-fixed MWEs (Sag et al., 2002) like Verb Particle Constructions (VPCs) and 'decomposable' idioms (Sag et al., 2002), as well as everything in between. Fixed MWEs are consecutive and do not vary at all, while semi-fixed MWEs can allow a wide array of variations. Where lexical variation is allowed, it ranges from inflection to token addition, alternation, or elision. Some semi-fixed MWEs include a core set of tokens that provide the key semantics and are never altered beyond inflection. These can be combined with additional tokens that can be replaced or modified to contextualize the MWE's semantics. Some semi-fixed MWEs have a fixed word order, and others, such as many VPCs, allow variable word order.

UMR has a similar array of treatments for annotating these MWEs. At any annotation stage, core tokens can be concatenated to form the concept used in the UMR graph. If the MWE is clausal, requires sense-disambiguation, or has a distinct argument structure, a new roleset can be created for it.

3.1 Fixed MWEs

A fixed MWE (Sag et al., 2002) is an unmodifiable, consecutive sequence of word tokens that maps to a single UMR graph concept. Many fixed MWEs are complex function words like *by-and-large* in English (Constant et al., 2017). These do not have argument structures and do not need anything beyond a single concatenated node in the graph (e.g., (b / by-and-large)).

Many predicating prepositional phrases are also fixed (*in_love*, *in_arrears*). Since these are clausal and take at least one argument, they are treated as a predicate in UMR, using the UMR participant roles during stage 0 annotation (7) and being assigned a roleset in stage 1 annotation (8).

(7) "The bank was in arrears."

```
(i / in-arrears
    :theme (b / bank)
    :aspect State
    :modstr FullAff)
```

(8) "John was in love with Mary."

```
(l / love-01
    :Arg0 (p / person
        :name (n / name :op1 "John"))
    :Arg1 (p2 / person
        :name (n2 / name :op1 "Mary")
    :aspect State
    :modstr FullAff)
```

During roleset creation, predicating PPs can either be assigned a unique roleset or included with one that already exists and is semantically/etymologically related. For example, *in-arrears* would be assigned its own roleset since it has no corresponding verbal or nominal roleset (i.e., in-arrears-01, :Arg1-entity owing money, :Arg2-amount, :Arg1-money owed), but, *in-love* can be included as part of love-01, which already exists for verbal/nominal *love*.

3.2 Verb Particle Constructions

Verb Particle Constructions are semi-fixed MWEs that include a specific verb and one or more specific particles. The verb may be inflected, and many VPCs allow the verb and particle to be split up. In UMR, VPCs are represented as a concatenated predicate. In English, they are included as their own rolesets, separate from the base verb.

(9) The sheep ate the flowers up.

```
(e / eat-up-02
    :Arg0 (s / sheep)
    :Arg1 (f / flower
        :refer-number Plural)
    :aspect Performance
    :modstr FullAff)
```

Fixed and semi-fixed MWEs like these are highly language-specific, and different languages may express similar concepts with different types of MWEs. For example, VPCs in English often correspond to verb compounds in Chinese (Sun et al., 2023) as the particle is generally considered to be a verb. However, the UMR annotation is similar, with the verb compound as a whole is treated as a UMR concept.

(10) 小 羊 把 花 吃掉 了 。
little sheep BA flower eat up ASP .

"The little sheep ate up the flower."

```
(x5 / 吃掉-01 ["eat up']
    :aspect Performance
    :modstr FullAff
    :Arg0 (x2 / 羊 ["sheep"]
        :mod (x1 / 小 ["little"]))
    :Arg1 (x4 / 花 ["flower"]))
```

Interestingly, Czech has no VPCs in the proper sense. Instead, some verbs (in one or more of their senses) can require a particular preposition as the only acceptable form of expression of one of its arguments. However, the preposition is not considered to be part of the predicate, even if neither the meaning of the verb and the preposition, nor the preposition and the noun phrase which form a PP, is compositional. An example is "zmínit se o něčem" ("to mention sth", lit. "to mention about sth(.locative-case)"), where the preposition "o" ("about"), requiring locative case, loses its meaning of "aboutness". Such constructions are thus not considered MWEs from the predicate point of view, and they would get the following UMR annotation:

(11) zmínit se o něčem
mention [refl.] about something

"to mention something"

```
(z / zmínit-se-01
    :Arg1 (n / něco)
    :aspect Perfective
    :modstr FullAff)
```

The example also shows that verbs with reflexive particles (such as "se" in this case), having a "frozen" meaning which is required to be used in the sentence simply as a [mostly discontinuous] part of the predicate, are always considered an MWE.

3.3 Semi-fixed MWEs

Many MWEs fall into the semi-fixed category, being semi-compositional, modifiable, and figuratively transparent, while not being entirely literal. Such MWEs are also handled in UMR

by concatenating core tokens in a graph predicate and using either participant roles or numbered arguments associated with a unique (or related) roleset. In some cases, there are a cluster of closely related MWEs that can be grouped together into a single roleset. In (12), a roleset "keep-eye-out-02" is used for instances of *"keep an eye out for", "keep an eye open for",* and *"keep your eyes peeled for"*:

(12) "He was keeping [an eye out]/[an eye open]/[his eyes peeled] for potholes."

```
(k / keep-eye-out-02
    :Arg0 (p / person
        :refer-person 3rd
        :refer-number Singular)
    :Arg1 (p2 / pothole
        :refer-number Plural)
    :aspect Activity
    :modstr FullAff)
```

Inside the roleset file, a slot/token structure shows that the first slot is always *keep*, the third slot is always *eye* (or its plural), and the third slot can take variants *out, open*, and *peeled*. (See Figure (1) for more on slots.)

4 Non-consecutive Constructions

Another challenge in UMR annotation lies in representing constructions that are cued by a non-consecutive (and sometimes inter-clausal) sequence of words. They are also MWEs in the sense that they consist of multiple words, but the words may be predominantly function words, and the meaning may not be derived from any one particular word in the sequence. Following AMR, UMR uses abstract rolesets to represent the established semantics of such constructions. For example, "the more ... the more ..." (*the Xer, the Yer*) is annotated with an abstract roleset called *correlate-91*, which take as arguments the two predicates that are correlated:

(13) The more I studied, the less I understood.

```
(c / correlate-91
    :Arg1 (m / more
        :Arg3-of (h / have-quant-91
            :Arg1 (s / study-01
                :Arg0 (i / i)
                :aspect Activity
                :modstr FullAff)))
    :Arg2 (l / less
        :Arg3-of (h2 / have-quant-91
            :Arg1 (u / understand-01
                :Arg0 i
                :aspect State
                :modstr FullAff))))
```

Chinese has a similar construction that also maps to the abstract concept *correlate-91*:

(14) 时间 越 临近 ，我就 越 感到
time more get close , I then more feel
幸福 。
happy

"The closer the time comes, the happier I will be"

```
(c / correlate-91
    :Arg1 (x3 / 临近-01
        :aspect Performance
        :modstr FullAff
        :Arg0 (x1 / 时间))
    :Arg2 (x8 / 感到-01
        :aspect Performance
        :modstr FullAff
        :Arg1 (x9 / 幸福)
        :Arg0 (i / individual-person
            :ref-person 1st
            :ref-number Singular)))
```

The same example could be given for Czech, with the correlation expressed via a pair of pronouns "čím ..., tím ...", for example "Čím důležitější schůzka, tím jsem nervóznější.", lit. "The more important the meeting [is], the more nervous I am."

5 Idioms - MWEs that Have a Figurative Interpretation

Idioms are MWEs that are ambiguous between a literal meaning and a figurative interpretation, where ambiguity can be resolved in context. Depending on how MWEs are interpreted, they are mapped to UMR concepts in different ways. If the literal meaning is intended given the context, such an expression can be represented compositionally in UMR. However, if the figurative meaning is intended, UMR concepts are created by concatenating the word tokens in the MWE, similar to how fixed or semi-fixed expressions are handled. We illustrate this with the expression "jump

on the bandwagon" in English, with a graph for the literal interpretation in (15) and one for the figurative interpretation (16). With the idiomatic interpretation, an English PropBank roleset jump-on-bandwagon-09 is used, shown in Figure (1).

(15) "He jumped on the bandwagon." *(lit.)*

```
(j / jump-03
    :Arg0 (p / person
        :ref-person 3rd
        :ref-number Singular)
    :Arg1 (b / bandwagon)
    :aspect Performance
    :modstr FullAff)
```

(16) "He jumped on the bitcoin bandwagon." (*'he joined the bitcoin boom'*)

```
(j / jump-on-bandwagon-09
    :Arg0 (p / person
        :ref-person 3rd
        :ref-number Singular)
    :Arg1 (b / bitcoin)
    :aspect Performance
    :modstr FullAff)
```

Rolesets for idiomatic MWEs are able to be quite descriptive about the relationships between the MWE's tokens and between the elements in the literal and figurative frames, as illustrated in Figure (1).

First, numbered roles are provided for participants in the idiomatic frame as well as any modifiers the expression can take. Here, the Arg0 of jump-on-bandwagon-09 corresponds to the same agent that appears in (15). Additionally, the expression *'jump on the bandwagon'* is modifiable, so the roleset provides an Arg1 for any phrase that might be used to modify 'bandwagon'. The idiomatic meaning of the expression is *'to join in with others who are following a certain fad'*, and it conveys this by evoking historical imagery of political parade-goers jumping onto the wagon that carried the band at the front of the parade. In current use, speakers identify the 'bandwagon' in the expression with a fad, and tell us what the fad is by modifying that token syntactically.

Next in the roleset, the tokens are identified and labeled with slot position, part of speech, and syntactic head. Then, two parallel graphs are given that use the slot labels (A-D), the numbered arguments (N-ARG0 and N-ARG1), and token values to map between the literal frame and the metaphorical frame. The token 'jump' in slot A is equated with the 'jump-03' roleset (physical jumping) in the literal interpretation and with the 'join-in-05' roleset (joining a group) in the idiomatic interpretation. The token 'bandwagon' in slot D appears as the destination argument of jump-03 in the literal frame and is equated with 'people following a fad' in the figurative frame. N-ARG0 is equated with the literal jumper and the figurative fad-joiner. Lastly, N-ARG1 tells us what kind of fad is being discussed in the figurative frame (as in *He joined the bitcoin boom*).

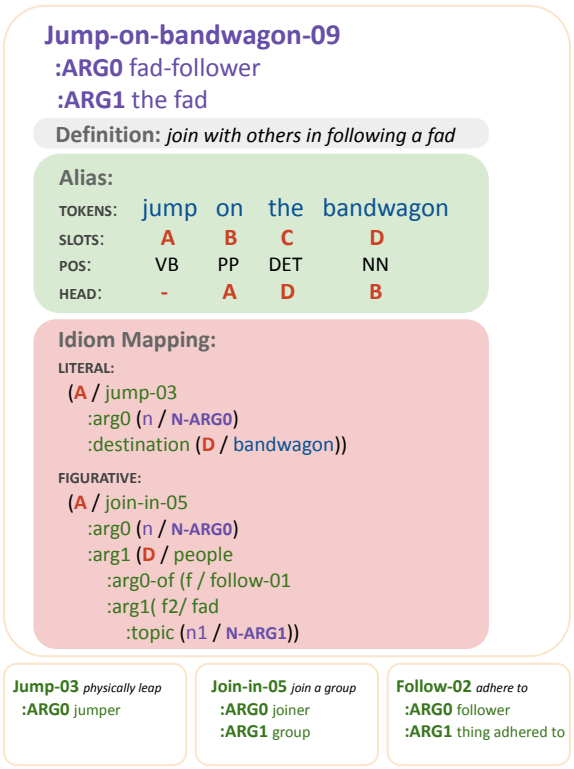

Figure 1: Roleset for jump-on-bandwagon-09 with token breakdowns and mappings between literal and figurative frames.

5.1 Chinese Idioms

Chinese idioms, known in Chinese as *xiyu*, can also have literal or figurative interpretations and are annotated in a similar manner to English. For example, the Chinese expression 炒鱿鱼 ('stir fry squid') can have a literal or figurative meaning depending on the context. In (17), 炒鱿鱼 should be interpreted literally and

compositionally, mapping to two UMR concepts, one for 炒 (stir-fry) and one for 鱿鱼 (squid):

(17) 他 在 厨房 里 炒 鱿鱼 。
he at kitchen inside stir-fry squid .

"He was stir-frying squid in the kitchen."

```
(x5 / 炒-01
    :Arg0 (i / individual-person
        :ref-person 1st
        :ref-number Singular)
    :Arg1 (x6 / 鱿鱼)
    :place (x4 / 厨房)
    :aspect Activity
    :modstr FullAff)
```

More often, however, the expression has a figurative interpretation of 'fire from a job', as in (18). In this case, it is treated as an MWE that maps to a single UMR concept.

(18) 他 被 那 个 公司 炒 了
he BEI that CL company stir-fry ASP
鱿鱼 .
squid .

"He was fired from that company."

```
(x6 / 炒鱿鱼-00 ["fire"]
    :Arg1 (i / individual-person
        :ref-person 3rd
        :ref-number Singular)
    :Arg0 (x5 / 公司 ["company"]
        :mod (x3 / 那 ["that"]))
    :aspect Performance
    :modstr FullAff)
```

Chinese Chengyu Chengyu (成语) are idioms that obey the grammar of ancient Chinese but have become fixed expressions in modern Chinese. The literal meaning of such expressions often describes a scenario in the past that no longer applies today. This is illustrated in (19), where the underlined expression is an idiom meaning *step by step*:

(19) 他 因 初次 创业 ,
he because first time start business ,
所以 凡事 都 步步为营 。
therefore everything all step by step .

"Because he started his own business for the first time, he took everything step by step."

```
(x9 / 步步为营-00
    :Arg0 (i / individual-person
        :ref-person 3rd
        :ref-number Singular)
    :mod (x7 / 凡事 ["everything"])
    :mod (x8 / 都 ["all"]))
    :cause (x4 / 创业-01
        :Arg0 i
        :mod (x3 / 初次 ["first time"])
        :aspect Performance
        :modstr FullAFF)
    :aspect Activity
    :modstr FullAFF)
```

5.2 Spanish Idioms

Idiomatic phrases in Spanish can also be taken literally or figuratively. For example, the Spanish phrase "con las manos en la masa" (meaning caught red-handed or in the act, lit. "with hands in the dough") could refer to an actual dough thief caught because of their messy fingers, or some other crime a person is caught committing.

(20) él fue atrapado con las manos en la
he was trapped with the hands in the
masa.
dough.

"he was caught red-handed."

```
(a / atrapar-01
    :Arg0 (p / person
        :ref-person 3rd
        :ref-number Singular)
    :manner (c / con-manos-en-masa-01
        :Arg0 p))
```

The UMR graph captures the idiomatic meaning by modifying a 'caught' verb with a roleset for the 'caught in the act' sense of 'with hands in the dough'.

The Spanish idiom "el que corta el bacalao" which denotes the person in charge (literally *he who cuts the cod*) is handled similarly, with a roleset for the idiomatic interpretation of 'cut cod':

(21) el que corta el bacalao
he who cuts the cod

"person in charge."

```
(c / cortar-bacalao-01
    :Arg0 (p / person
```

:ref-person 3rd
:ref-number Singular))

5.3 Czech Idioms

In Czech, the same literal vs. figurative interpretations exist for many expressions, with most being interpreted idiosyncratically or figuratively (i.e., non-compositionally). In such cases, using a single synonymous predicate is generally preferred, although the choice of predicate can be context-dependent. (22) shows two different solutions used for the idiom "jít z kopce" (lit. "go downhill"): the first one uses a predicate corresponding to the literal meaning of the idiom (but which could also be interpreted figuratively), whereas the second solution (with the predicate "chudnout-01") assumes that the context implies that it means "he was getting poorer."

(22) šlo to s ním z kopce.
went it with him down [the]hill.

He deteriorated/became poorer/became more sick/became asocial/was getting worse results at work/...

(j / jít-z-kopce-01
:Arg0 (p / person
:refer-person 3rd
:refer-number Singular))

(c / chudnout-01
:Arg0 (p / person
:refer-person 3rd
:refer-number Singular))

5.4 Arapaho Idioms

Arapaho is an agglutinating polysynthetic language and as such has very few constructions that might qualify as multiword expressions. Still, Arapaho has idiomatic constructions that need to be treated in a way that allows literal interpretations to be separated from figurative ones. In (23), *'nih3iikoncebeit'* is a word/phrase (lit. 'a ghost shot him [with an arrow]') that means that someone gave the person in question a disease. In the morphological breakdown, */3iikon-/* is a noun-incorporating preverb that refers to the ghost. While Arapaho might not normally include such preverbs as part of the predicate in a UMR graph (instead, using just */ceb/*, 'shoot', as the predicate), in the case of idiomatic expressions like this, the graph predicate includes it. A roleset for this phrase would include numbered arguments for the shooter (disease-giver) and victim as well as the disease. A source/target mapping in the roleset file would link */3iikon-/* (ghost) to the numbered argument for disease. This roleset is separate from the roleset for literal shooting (ceb-01), but is included in the same file.

(23) nih- 3iikon- ceb -eit
PAST- ghost- shoot -4/3

"Someone gave him a disease."

(x / 3iikonceb-01
:Arg0 (p / person
:refer-person 3rd
:refer-number Singular)
:Arg1 (p2 / person
:refer-person 3rd
:refer-number Singular)
:Arg2 [implicit-for-coref]
:aspect Performance
:modstr FullAff)

6 Proverbs

Like idioms, proverbs also have a literal and figurative interpretation. Unlike idioms, however, proverbs are often self-contained sentences with all participants of the predicates filled, and it is hard to construct alternative contexts in which the proverbs can be interpreted literally. Since they tend to be longer than idioms and their literal meaning can be constructed compositionally in UMR, we annotate proverbs with an abstract roleset called *proverb-91*, which takes two arguments. The first argument is required and is annotated compositionally; the second argument will be described in the next section. We illustrate a standard proverb that uses the Arg1 with a Chinese proverb in (24).

(24) 山 高 皇帝 远
mountain high emperor far away

"The mountains are high and the emperor is far away."

(p / proverb-91
:Arg1 (a / and
:op1 (x2 / 高-01 ["high"]
:Arg0 (x1 / 山 ["mountain"])

:aspect State
:modstr FullAFF)
:op2 (x4 / 远-01
:Arg0 (x3 / 皇帝 [“emperor”])
:aspect State
:modstr FullAFF)))

6.1 Xiehouyu, or two-part allegorical sayings

Xiehouyu, also known as a two-part allegorical saying (Lai, 2008), is common in Chinese and other Asian languages; similar forms can be found in other languages as well. Xiehouyu consists of two parts–an antecedent that is a highly allegorical and figurative expression, and a consequent that provides an explanation for the antecedent. We represent such sayings with *proverb-91* as well, using Arg1 for the antecedent and Arg2 for the consequent:

(25) 你 这 是 大炮 打 蚊子 ——
you this be cannon shoot mosquito -
小题大做
solving small problem with big action

“(By doing this), you are shooting cannon at mosquitoes - making too much out of something small.”

(c / proverb-91
:Arg1 (x5 / 打-02 [“shoot”]
:Arg0 (i / individual-person
:ref-person 2nd
:ref-number Singular)
:Arg1 (x6 / 蚊子 [“mosquito”])
:instrument (x4 / 大炮 [“cannon”])
:aspect Habitual
:modstr FullAff)
:Arg2 (x8 / 小题大做-01 [”make too much out of something small”]
:aspect Habitual
:modstr FullAff))

The English saying *“Life is like a box of chocolates– you never know what you’re going to get.”* follows the same format, and two-part proverbs (where the second part is optional) are also present in Russian (Dahl, 2000).

7 Related & Future Work

MWEs have always been a thorny issue for computational linguistics (Sag et al., 2002) and have been studied from various perspectives, from linguistic modeling (annotation) to automatic identification. The existence of a series of workshops focusing on MWEs (Markantonatou et al., 2020; Cook et al., 2021; Bhatia et al., 2022) attests to the interest of a large community of researchers. For many European languages, multiword expressions including verbal ones have been tackled in the PARSEME project[5] (Rosén et al., 2015; Savary et al., 2017). For Czech, previous work on identification and extraction of verbal (and other) MWEs from treebanks is described in (Uresova et al., 2013; Urešová et al., 2016; Bejček et al., 2017).

This work addresses the representation of MWEs in UMR, resolving many of the issues surrounnding many-to-one word-to-concept mappings. It builds on previous work respecting light verb constructions in multilingual PropBanks (Hwang et al., 2010) and AMR annotation of certain English constructions (Bonial et al., 2018), and expands it to include additional MWEs in multiple languages. We have discussed different types of MWEs that present challenges to UMR annotation and presented solutions for their treatment. Users with an existing valency lexicon or PropBank may wish to undertake creating new MWE rolesets based on recommendations we have outlined. Forthcoming work will build on the strategies outlined here as we tackle one-to-many word-to-concept mappings.

Acknowledgments

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. The work on Spanish is supported by a Clare Boothe Luce Scholarship.

[5] https://typo.uni-konstanz.de/parseme

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.

Eduard Bejček, Jan Hajič, Pavel Straňák, and Zdeňka Urešová. 2017. Extracting verbal multiword data from rich treebank annotation. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT 15)*, pages 13–24, Bloomington, IN, USA. Indiana University, Bloomington, Indiana University, Bloomington.

Archna Bhatia, Paul Cook, Shiva Taslimipoor, Marcos Garcia, and Carlos Ramisch. 2022. Proceedings of the 18th workshop on multiword expressions@ lrec2022. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC2022*.

Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O' Gorman, Martha Palmer, and Nathan Schneider. 2018. Abstract meaning representation of constructions: The more we include, the better the representation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Paul Cook, Jelena Mitrović, Carla Parra Escartín, Ashwini Vaidya, Petya Osenova, Shiva Taslimipoor, and Carlos Ramisch. 2021. Proceedings of the 17th workshop on multiword expressions (mwe 2021).

VI Dahl. 2000. Proverbs of the russian people. moscow: Russian language media.

Jens Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Timothy J. O'Gorman, Andrew Cowell, W. Bruce 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 Intell.*, 35:343–360.

Jena D Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 82–90.

Huei-ling Lai. 2008. Understanding and classifying two-part allegorical sayings: Metonymy, metaphor, and cultural constraints. *Journal of Pragmatics*, 40(3):454–474.

Stella Markantonatou, John Philip McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova, and Agata Savary. 2020. Proceedings of the joint workshop on multiword expressions and electronic lexicons. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*.

Victoria Rosén, Gyri Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Mititelu. 2015. A survey of multiword expressions in treebanks. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 179–193, Warszawa, Poland. IPI-PAN, IPIPAN.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, Antoine Doucet, Kübra Adalı, Verginica Mititelu, Eduard Bejček, Ismail El Maarouf, Gülşen Cebiroğlu Eryiğit, Luke Galea, Yaakov Kerner, Chaya Liebeskind, Johanna Monti, Carla Escartín, Jolanta Kovalevskaitė, Simon Krek, Lonneke Plas, Cristina Aceta, Itziar Aduriz, Jean-Yves Antoine, Greta Attard, Kirsty Azzopardi, Loic Boizou, Janice Bonnici, Mert Boz, Ieva Bumbulienė, Jael Busuttil, Valeria Caruso, Manuela Cherchi, Matthieu Constant, Monika Czerepowicka, Anna Santis, Tsvetana Dimitrova, Tutkum Dinç, Hevi Elyovich, Ray Fabri, Alison Farrugia, Jamie Findlay, Aggeliki Fotopoulou, Vassiliki Foufi, Sara Galea, Polona Gantar, Albert Gatt, Anabelle Gatt, Carlos Herrero, Uxoa Inurrieta, Glorianna Jagfeld, Milena Hnátková, Mihaela Ionescu, Natalia Klyueva, Svetla Koeva, Viktória Kovács, Taja Kuzman, Svetlozara Leseva, Sevi Louisou, Teresa Lynn, Ruth Malka, Héctor Martínez Alonso, John McCrae, Helena Caseli, Ayşenur Miral, Amanda Muscat, Joakim Nivre, Michael Oakes, Mihaela Onofrei, Yannick Parmentier, Caroline Pasquer, Maria Buono, Belem Sanchez, Annalisa Raffone, Renata Ramisch, Erika Rimkutė, Monica-Mihaela Rizea, Katalin Simkó, Michael Spagnol, Valentina Stefanova, Sara Stymne, Umut Sulubacak, Nicole Tabone, Marc Tanti, Maria

Todorova, Zdeňka Urešová, Aline Villavicencio, and Leonardo Zilio. 2017. Annotated corpora and tools of the PARSEME shared task on automatic identification of verbal multiword expressions (edition 1.0).

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.

Zdeňka Urešová, Eduard Bejček, and Jan Hajič. 2016. Inherently pronominal verbs in czech: Description and conversion based on treebank annotation. In *Proceedings of the 12th Workshop on Multiword Expressions (ACL 2016)*, pages 78–83, Stroudsburg, PA, USA. Association for Computational Linguistics (ACL), Association for Computational Linguistics (ACL).

Zdenka Uresova, Jan Hajic, Eva Fucikova, and Jana Sindlerova. 2013. An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 58–63, Atlanta, Georgia, USA. Association for Computational Linguistics.

Shira Wein, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, Leonie Harter, and Nathan Schneider. 2022. Spanish Abstract Meaning Representation: Annotation of a general corpus. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.