Extending Uniform Meaning Representation to Persian: The First Corpus Resource

Minoo Nassajian, Daniel Zeman
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics (ÚFAL)
Prague, Czechia
{nassajian, zeman}@ufal.mff.cuni.cz

Abstract

Uniform Meaning Representation (UMR) has cross-linguistic design principles that make it particularly well-suited as a semantic representation framework for capturing all language-specific phenomena. Despite its growing adoption, no UMR corpus currently exists for Persian. In this paper, we present the first version of a Persian UMR dataset created through a rule-based conversion of existing Persian AMR annotations from The Little Prince corpus, followed by manual mapping of split semantic roles from AMR to their finer-grained UMR counterparts. We report detailed statistics on the conversion, analyze the challenges of mapping Persian AMR structures into UMR, and provide illustrative examples. The resource is freely available and it lays the groundwork for subsequent enrichment of Persian UMR with additional semantic layers, including co-reference, named entities, and discourse relations.

Keywords: Uniform Meaning Representation, UMR, Persian, Abstract Meaning Representation, AMR

1. Introduction

Meaning representation is a fundamental concept in computational linguistics and natural language processing (NLP) that uses formal systems to capture the semantic content of language in structured, machine-processable format, and can improve the performance of different applications such as machine translation (Gao and Vogel, 2011; Song et al., 2019), summarization (Liu et al., 2015), semantic search (Ribeiro et al., 2022), data augmentation (Shou et al., 2022), or dialogue systems (Bonial et al., 2020; Kapanipathi et al., 2021; Pan et al., 2015). Meaning representation frameworks can be broadly categorized into graph-based and non-graph-based structures (Sadeddine et al., 2024). Among the graph-based ones, AMR (Banarescu et al., 2013) has been widely used to represent sentence-level meaning as directed, rooted, acyclic graph in which each node is associated with a specific concept and edges connect concept nodes while showing different semantic relation types. However, despite the widespread usage of AMR, the difficulties of extending it to diverse low-resource languages, coupled with its lack of annotations crucial for logical inference, such as modality, aspect, and scope, necessitated the development of UMR (Van Gysel et al., 2021). The present work is the first step in developing a Persian UMR dataset, with the goal of significantly enhancing cross-lingual semantic understanding, enabling more effective comparison of meaning between Persian and other languages already covered by UMR. In particular, we focus on split roles, following the methodology outlined by (Post et al., 2024).

This paper is structured as follows: Section 2 describes the related work within the context of both AMR and the emerging UMR framework, highlighting the gap in resources for Persian. Section 3 details the origin and characteristics of the initial corpus that serves as the foundation for our conversion effort. Subsequently, Section 4 describes the conversion methodology and the set of rules developed for mapping AMR semantic roles to their UMR counterparts. Section 5 provides a quantitative overview of the resulting Persian UMR corpus, analyzing the frequency and distribution of role mappings to validate our methodology and reveal insightful linguistic patterns. Finally, the conclusion and future work section summarizes our findings and outlines the critical next steps for expanding this corpus into a comprehensive UMR resource.

2. Related Work

UMR research has progressively expanded across diverse languages since its introduction. The foundations of the framework were laid down by Van Gysel et al. (2021), followed by the first data release in 2023 (Bonn et al., 2023b), which included datasets of varying sizes for six languages: English, Chinese (358 sentences from Wikinews), Navajo (Athabascan, USA; 522 sentences from historical narratives), Arapaho (Algonquian, USA; 408 sentences from narrative texts), Kukama (Tupian, Amazon; 105 sentences from traditional sto-

Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, pages 172–182

ries), and Sanapaná (Mascoian, Paraguay; 602 sentences). In the second release (Bonn et al., 2025), the English dataset grew substantially, with 87,038 sentences being converted from existing AMR resources through semi-automated processes. This release also added a dataset for Latin (50 manually annotated sentences from the Sallust treebank) and Czech (175,500 sentences automatically converted from the Prague Dependency Treebank, out of them 91 sentences also annotated manually) (Štěpánek et al., 2025).

There has been relatively limited research on meaning representation for the Persian language. One notable exception is the work by Mirzaei and Moloodi (2016), who introduced the Persian Proposition Bank (PerPB). This resource extends the Persian Dependency Treebank (PerDT) (Rasooli et al., 2013) with a semantic layer of predicate-argument annotations, inspired by PropBank (Kingsbury and Palmer, 2002) and VerbNet (Kipper et al., 2006). Their approach treats not only verbs but also propositional nouns and adjectives as semantic predicates, annotating over 29,000 sentences with detailed semantic roles.

The most recent research on Persian meaning representation focused on creating a Persian AMR dataset (Takhshid et al., 2022; Tohidi et al., 2024). They developed this corpus by annotating the Persian translation of "The Little Prince," containing 1,562 sentences, and addressed the annotation guidelines[1] for Persian-specific constructions such as light verb constructions, impersonal constructions, and clitics.

Building on this line of research, the present study introduces the first step towards a Persian UMR corpus by converting and extending the existing Persian AMR dataset. Specifically, we apply the AMR-to-UMR conversion framework proposed by Post et al. (2024), which provides guidelines for refining split semantic roles. This effort not only highlights the applicability of cross-linguistic UMR guidelines to Persian but also uncovers language-specific challenges—such as complex predicates, clitic behavior, and pro-drop—that shape the design and annotation of meaning representations. The next section outlines the characteristics of the Persian AMR dataset.

3. Source Data: Persian AMR Corpus

Due to the limited availability of UMR resources for Persian, the present corpus aims to provide an initial annotated dataset that supports linguistic analysis and future development of Persian UMR annotation. In this regard, we use the only available Persian AMR (PAMR) dataset (Takhshid et al., 2021) introduced by Takhshid et al. (2022). This corpus represents the first attempt to adapt AMR to Persian and provides a valuable foundation for cross-linguistic meaning representation. During the construction of PAMR, certain AMR features were adjusted for the Persian language as follows:

- Light verb constructions: Persian uses extensive light verb constructions (LVCs) where a non-verbal element combines with a semantically lighter verb to form a predicate and due to structural integrity they are considered as single lexical verbs (Karimi-Doostan, 1997). So, Persian AMR explicitly preserves them as unified semantic units. For example, in "talâŝ kardan" (to do an effort), the non-verbal element "talâŝ" (effort) cannot be separated from the light verb "kardan" (to do) in semantic representation without losing essential information. In AMR, "talâŝ kardan" is kept as one concept node.

- Pro-drop characteristics: As a pro-drop language, Persian allows null subjects realized only through verb morphology. This creates challenges for AMR's more explicit argument structure. Additionally, inanimate subjects might not follow subject-verb agreement in number, further complicating semantic representation (Karimi, 2008).

- Clitics and possessive constructions: Persian employs various clitics that can serve as subjects, objects, or possessors. Some constructions have no overt subject but use pronominal enclitics to indicate possession or experiencer roles.

As UMR is designed in a way that accommodates languages across diverse typological spectrums, it effectively addresses these linguistic features specific to Persian that were previously challenging for AMR. In particular, UMR allows for the compositional semantic representation of Persian LVCs by aligning the concept, representing the LVC, back to the surface tokens (the light verb and its object). It also systematically represents implicit subjects common in Persian pro-drop contexts by explicitly encoding inferred arguments. Furthermore, UMR is capable of clearly representing semantic roles of Persian clitics, thus resolving the ambiguity that AMR previously faced in encoding possessive and experiencer relationships. It should be also noted that since the corpus is derived from Persian AMR annotations, some upstream annotation errors may propagate. Manual inspection during split-role conversion mitigated this risk by correcting inconsistent semantic structures when detected.

[1] https://github.com/Persian-AMR/Annotation-Guidelines

In the next section, we describe the rule-based mapping approach for converting split roles from AMR to UMR, following the guidelines proposed for English, and discuss how we adapt and apply these rules to Persian.

4. Conversion Methodology

UMR inherits the overall graph-based architecture of AMR[2] but introduces finer-grained semantic roles that better capture cross-linguistic distinctions. Bonn et al. (2023a) provide the most comprehensive cross-mapping to date and show that AMR roles must undergo four distinct types of change when converted to UMR:

1. New roles: UMR introduces roles that have no direct AMR equivalent — such as `:actor`, `:experiencer`, and `:force` — to capture semantic distinctions (e.g., initiator vs. causer) that are crucial across languages.

2. Renamed roles: Certain AMR roles remain conceptually similar but are given more typologically neutral labels to improve clarity, consistency, and cross-linguistic applicability. For example, the `:location` tag becomes `:place`, and `:beneficiary` is changed to `:affectee`.

3. Split roles: Some AMR roles are underspecified and may correspond to multiple UMR roles depending on context. For example, AMR's `:cause` must be resolved to either `:cause` (physical causation) or `:reason` (motivational explanation); `:source` may map to `:source` (animate giver), `:start` (origin of motion), or `:material` (substance).

4. Unchanged roles: A number of roles, such as `:purpose`, `:instrument`, and `:manner`, transfer directly with no modification because they already align well with cross-linguistic semantic needs.

Among these categories, split roles are particularly critical because they require detailed linguistic analysis and often language-specific cues to resolve. Accordingly, this paper focuses on identifying and converting split roles as the essential first step toward a full Persian UMR corpus.

[2]Here we refer specifically to what is known as the sentence-level graph in UMR. In addition, UMR defines document-level relations, which are beyond the scope of the present paper.

4.1. AMR to UMR Conversion Principles

A key challenge in AMR to UMR conversion is the presence of split role cases where a single AMR relation can map to multiple possible UMR roles depending on context such as animacy, event type, or discourse function. Post et al. (2024) employ the animacy feature together with a probability distribution derived from gold-standard frequencies to define conversion rules (a - f) as follows:

(a) `:cause` →

$$\begin{cases} \texttt{:cause} & \text{if animate} \\ \texttt{:cause}, \ \texttt{:reason} & \text{if inanimate} \end{cases} \quad (1)$$

(b) `:destination` →

$$\begin{cases} \texttt{:goal} & \text{if animate} \\ \texttt{:recipient}, \ \texttt{:goal} & \text{inanimate} \end{cases} \quad (2)$$

(c) `:source` →

$$\begin{cases} \texttt{:source} & \text{if animate} \\ \texttt{:source}, \ \texttt{:material}, \ \texttt{:start} & \text{inanimate} \end{cases} \quad (3)$$

(d) `:consist-of` →

$$\begin{cases} \texttt{:group} & \text{if animate} \\ \texttt{:group}, \ \texttt{:part}, \ \texttt{:material} & \text{inanimate} \end{cases} \quad (4)$$

There are two other tags, called `:mod` and `:part`, that are considered "split roles" within the context of this conversion task for two key reasons. First, `:part` is a potential outcome of the genuinely non-deterministic AMR role `:consist-of` (which can also map to `:group` or `:material`). For the model to be complete, it must be able to predict `:part` as a target label, hence its inclusion in the set of "split" outcomes.

Second, for `:mod`, the split is not based on animacy but on semantic vagueness. In AMR, `:mod` denotes a dependent whose function does not fit into more specific relations. In UMR, there are two such vague roles: `:mod` for adjunct-like optional modifiers, and `:other-role` for entities

that seem to be core parts (participants) of the event but do not cleanly fit into any of the specific participant roles. Hence the converting rules for these tags are as follows:

(e) :mod → { :mod ; :other-role, if vague-modifier

(f) :consist-of, :part → {:part

As no robust animacy parser or large animacy-annotated corpus currently exists for Persian, and our objective was to achieve high-precision UMR annotations, we adopted the same role-mapping logic described above and adhered to the official UMR guideline[3] and all split-role decisions in our Persian corpus were performed manually. Moreover, linguistic characteristics such as LVCs, clitics, and pro-drop features create ambiguity when mapping AMR relations to UMR roles, since multiple UMR interpretations may be possible depending on contextual and semantic factors. Consequently, manual analysis was required to ensure consistent role assignment. The next section describes our manual adaptation of these rules for the Persian AMR (PAMR) corpus.

4.2. Manual Split-Role Conversion for Persian

A single expert annotator with a background in computational linguistics and Persian semantics systematically re-labeled every AMR edge that corresponds to a split role. For each occurrence, the annotator examined the full Persian sentence and animacy feature to select the correct UMR label. To assess the reliability of manual split-role annotation, an additional linguist independently annotated a subset of 100 instances containing split-role phenomena. Inter-annotator agreement was measured using Cohen's κ, yielding a score of $\kappa = 0.845$. Disagreements were subsequently resolved through expert adjudication by a third linguist following the official UMR guidelines. The adjudicated labels constitute the final gold annotations. Most disagreements occurred between roles such as :goal vs. :recipient and :source vs. :material, reflecting subtle distinctions encoded in UMR role definitions. The next sections outline the conversion of AMR structures to UMR role representations and we demonstrate the mapping process through concrete examples drawn from the Persian data.

[3] https://github.com/ufal/umr-guidelines/blob/master/guidelines.md

4.2.1. Mapping Modifier Role

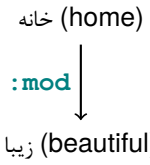
Based on the UMR guidelines, the :mod role (modifier) is a general-purpose semantic label used to indicate a property or attribute of an entity or event. It is applied as a default when the modification does not fit a more specific semantic role, such as :place or :time. The following list shows the categorization of words that can be annotated with this label:

- Adjectives (e.g., *large* in "large house")
- Adverbs denoting manner (e.g., *quickly* in "run quickly")
- Demonstratives (e.g., *this* in "this book")
- Participial modifiers (e.g., *broken* in "broken window")
- Attributive nouns (e.g., *steel* in "steel bridge")

In (1), there is an example of an adjective that has the same :mod label in both AMR and UMR.[4]

(1) زیبا خانه
zibâ xâne-ye
beautiful home
'beautiful home'

The AMR and UMR graphs of the above example are identical:



4.2.2. Mapping Destination Role

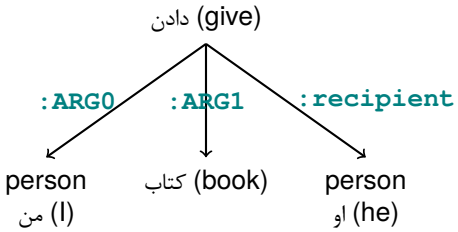
A significant refinement in UMR concerns the mapping of the AMR role :destination. In AMR, :destination is a general role denoting any endpoint of motion or transfer. UMR, informed by cross-linguistic typology, splits this role into two more semantically precise roles based on animacy and the nature of the transfer: :goal for inanimate endpoints and :recipient for animate endpoints that gain possession. This distinction allows UMR to more accurately capture the semantic roles of participants across diverse linguistic constructions. Examples (2) and (3) illustrate how to map AMR :destination role to UMR roles:

[4] Examples are presented with Persian text in the first line, transcription in the second line, English gloss in the third line, and English translation in the fourth line. Note that the words are ordered right-to-left, following the Persian writing system.

(2)

دادم	او	به	کتاب را
dâdam	u	be	ketâb râ
gave	him	to	the book

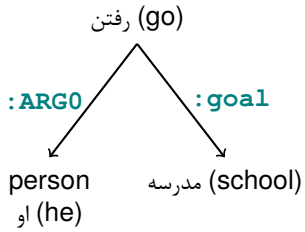
'I gave the book to him'



(3)

رفت	مدرسه	به	او
raft	madrese	be	u
went	school	to	he

'He went to school'



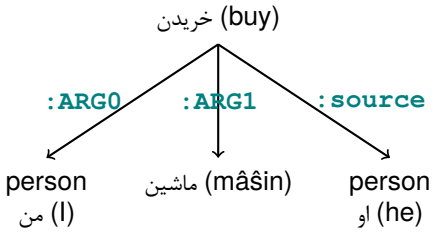
4.2.3. Mapping Source Role

In UMR, the `:source` role specifically marks an animate entity from which a theme separates or originates. It is distinct from the inanimate `:start` (location) and `:material` (composition) roles, reflecting a typologically-motivated split of AMR's general `:source` based on animacy. This role is typically applied to nouns and pronouns denoting people, animals, or organizations. The following examples show this role mapped to UMR roles.

(4)

خریدم	او	از	ماشین را
xaridam	u	az	mâšin râ
bought	him	from	the car

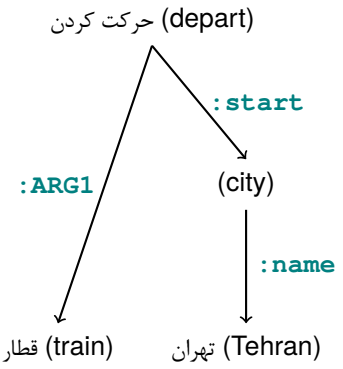
'I bought the car from him'



(5)

حرکت کرد	تهران	از	قطار
harekat kard	Tehrân	az	qatâr
departed	Tehran	from	the train

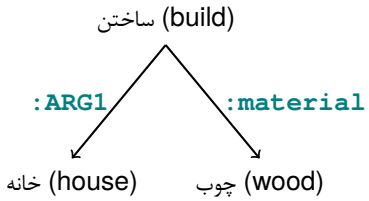
'The train departed from Tehran'



(6)

ساخته شد	چوب	از	خانه
sâxte ŝod	ĉub	az	xâne
was built	wood	from	the house

'The house was built of wood'



4.2.4. Mapping Consist-of Role

In the UMR schema, the AMR role `:consist-of` is decomposed into three semantically more precise roles based on the nature of the part-whole relationship:

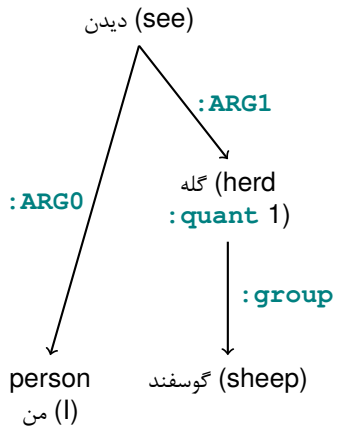
- `:group` for collections of animate entities (e.g., team, committee, herd; the relation is directed from the group concept to the member concept)
- `:part` for integral components of an inanimate whole (e.g., chapter, branch, piece; the relation is directed from the whole to the component)
- `:material` for the constituent substance an entity is composed of (e.g., wood, steel, water; the relation is directed from the entity to the material)

The examples of (7) and (8) show how to map the `:consist-of` tag to the UMR `:group` and `:part` labels (in this case, `:part` is inverted to `:part-of` because the whole is presented as a modifier). Moreover, some instances of `:consist-of` denote raw material, and like with `:source`, we map them to the UMR `:material` role (see the example 6).

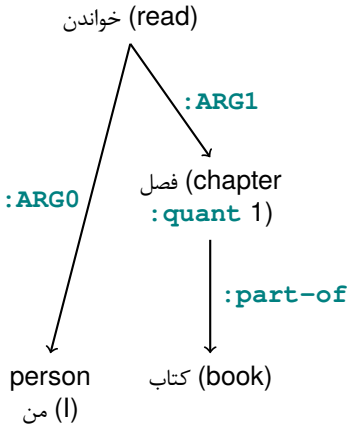
(7)

دیدم	گوسفند	گله	یک
didam	gusfand	galle	yek
saw	sheep	herd	a

'I saw a herd of sheep'



(8) خواندم کتاب را فصل یک
xândam ketâb râ fasl-e yek
read the book chapter a
'I read a chapter of the book'

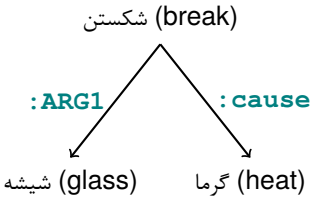


4.2.5. Mapping Cause Role

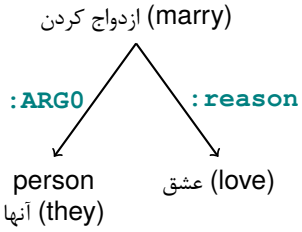
In the UMR schema, the representation of causality is refined through distinct strategies depending on the syntactic and discourse context. The AMR :cause role is split into :cause for inanimate entities that directly bring about an event and :reason for entities that motivate an action. This distinction enhances semantic precision. Examples (9) and (10) show causative constructions. Furthermore, UMR handles inter-sentential causality differently from intra-sentential causality. When a causal relationship links two separate sentences within a corpus, the abstract concept cause-01 is invoked to reify the relation, with :ARG1 pointing to the causing event and :ARG2 to the resulting event (see example 11).

It should be noted that the UMR guidelines also define a :causer role, to be used in causative constructions such as "Grandmother made the kid drink the water." However, that role is defined as Stage 0 role, to be used in languages that have no lexical resources with rolesets for predicates. Persian UMR qualifies as Stage 1 project, as the source AMR annotation already contains numbered PropBank-style arguments. Therefore, causative agents should be annotated as :ARG0 and would not result from conversion of :cause.[5]

(9) شکست شیشه گرما به‌خاطر
šekast šiše garmâ be xâter-e
broke glass heat because of
'The glass broke because of the heat'



(10) ازدواج کردند عشق به خاطر آنها
ezdevâj kardand ?eŝq be xâter-e ânhâ
married love because of they
'They got married because of love'



(11) a. طغیان کرد رودخانه
toqiyân kard rudxâne
flooded river
'The river flooded.'

b. نابود شدند محصولات و
nâbud ŝodand mahsulât va
was destroyed crops and
'And, the crops were destroyed'

4.2.6. Mapping Part Role

Within the UMR schema, the AMR role :part represents a specific case within the broader reclassi-

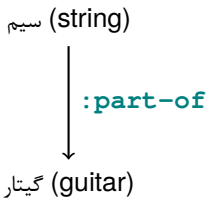
[5]We did not encounter such sentences in the data, though.

fication of part-whole relationships. While its surface mapping is deterministic—AMR :part maps directly to UMR :part—it is conceptually grouped under the umbrella of "split roles." This classification arises because :part is one of three possible outcomes (alongside :group and :material) resulting from the decomposition of the more general AMR role :consist-of. Thus, it is considered a "split role" not due to a change in its label, but because it is a key member of the new, more precise set of roles that collectively replace the underspecified AMR :consist-of. Example (12) illustrates the usage of :part annotation in UMR.

(12)

گیتار	سیم
gitâr	sim-e
guitar	string

'Guitar string.'



The next section will discuss a quantitative profile of the final Persian-UMR dataset and highlight the key role-mapping results.

5. Corpus Statistics and Analysis

PAMR contains 1,562 manually annotated sentences of the Persian translation of "The Little Prince" story. Table 1 summarizes the main statistics of the dataset.

Feature	Value
Number of sentences	1,562
Number of unique words	3,520
Number of tokens	14,427
The shortest sentence length	1
The longest sentence length	65

Table 1: Corpus statistics for Persian AMR.

The quantitative analysis of the annotated corpus reveals critical insights into the complexity of converting AMR roles to UMR. Table 2 highlights the varying degrees of determinism and ambiguity inherent in the mapping process.

According to Table 2, the AMR role :mod maps deterministically to UMR :mod in all 1,236 instances (100% of cases), meaning that the annotator did not identify any of the instances as being integral event participants. This is not too surprising given the nature of :other-role: it is intended as a last resort, but at present it is not used (needed) in any of the 8 languages in UMR 2.0.

AMR Role	#AMR	UMR Role	#UMR
:mod	1,236	:mod	1,236
:destination	5	:goal	5
		:recipient	0
:cause	2	:cause	75
cause-01	126	**cause-01**	38
		:reason	11
		:purpose	2
		:result	1
		:condition	1
:source	27	:source	17
		:start	7
		:cause	2
		:manner	1
:part	76	:part	76
:consist-of	13	:part	3
		:material	5
		:unit	3
		:group	2
Total	1485	**Total**	1485

Table 2: Mapping of Persian AMR to UMR roles. Note that **cause-01** is an abstract concept rather than a role; switching between it and a role involves changes in the graph structure.

Moreover, as explained before, the conversion of the :part role is deterministic and all 76 AMR :part instances are unchanged in UMR.

Conversely, other roles demonstrate significant splitting, necessitating disambiguation. The limited sample of :destination (5 instances) maps entirely to :goal, suggesting a potential bias in the data sample towards inanimate endpoints. A larger corpus would be required to observe the expected split with :recipient for animate entities.

The :source role is observed as a complex non-deterministic role, as it mapped to :source and :start based on Section 4.2.3. However, while analyzing the data, we encountered unexpected[6] mapping of the :source role to :cause, :location, and :manner. For example, in sentence (13) from the Persian AMR corpus, the word 'همین' (this) is annotated as :source but it was converted to :cause in UMR.

(13)

است	همین	از	غصه‌م
ast	hamin	az	qosse-am
is	this	from	my sorrow

'My sorrow is from this'

In the case of the :cause role, we can see that the majority of mappings are to :cause and

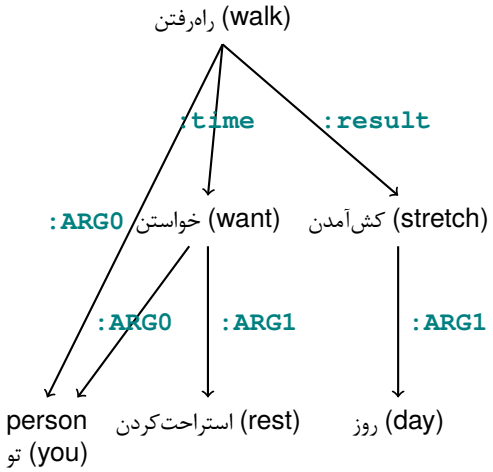
[6]By 'unexpected' we mean that it was not observed by Post et al. (2024) as an option in English, hence we did not list it in Section 4.1.

then cause-01. Moreover, contrary to the English conversion rules, there were some `:cause` annotations in the Persian AMR corpus that had to be mapped to `:result` (example (14)), `:condition` (example (15)), and `:purpose` (example (16)) in UMR.

(14)

کش‌خواهدآمد	روز	راه‌برو	می‌خواهی	استراحت	هروقت
keŝ xâhad âmad	ruz	râh boro	mixâhi	esterâhat	har vaqt
will stretch	day	walk	want	rest	when ever

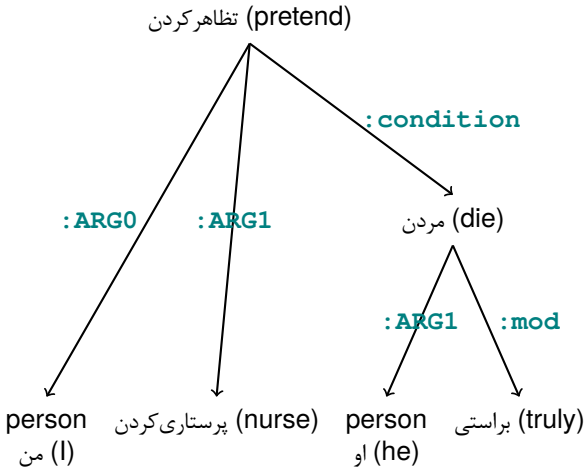
‘Whenever you want to rest, walk.
Then the day will stretch.’



(15)

می‌مرد	براستی	وگرنه	تظاهر می‌کنم	به پرستاری
mimord	be râsti	vagar na	tazâhor mikonam	be parastâri
die	truly	other wise	pretend	to nurse

‘I pretend to nurse.
Otherwise, He would truly died.’



(16) a.

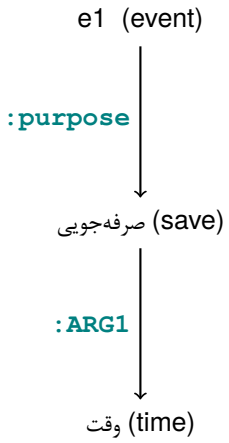
می‌فروشی؟	قرص	چرا
miforuŝi	qors	ĉerâ
sell	pills	why

‘Why do you sell pills?’

b.

در وقت	صرفه‌جویی	برای
dar vaqt	sarfe juyi	barây-e
time	save	to

‘To save time.’



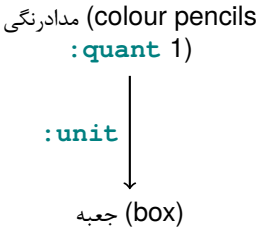
Furthermore, in cases where there is insufficient evidence to determine if the causer is animate or motivational (e.g., the `:ARG0` is unknown in the AMR graph), the annotator often defaulted to the more general role `:cause`. This conservative strategy is employed because `:cause` is semantically broader, it does not assume intention (which would be required for `:reason`), and it can accommodate both physical forces and abstract, non-motivational causes. For instance, in a sentence like “The event was caused by [an unknown factor],” where the factor’s nature is unclear, `:cause` serves as a safe, neutral default.

For the `:consist-of` role, we faced a similar issue as well and some AMR data were mapped to the `:unit` role. This occurred in contexts like “a box of colour pencils,” where the annotator interpreted the relationship not as material composition (`:material`) or integral parthood (`:part`), but primarily as quantification and containment—the box serves as a quantitative unit for the pencils.

(17)

مدادرنگی	جعبه	یک
medâd rangi	ja?bey-e	yek
colour pencils	box of	a

‘A box of colour pencil’



6. Conclusion and Future Work

This paper presented the first Persian UMR corpus created by systematically converting an existing Persian AMR resource using rule-based mapping combined with manual annotation. Like its AMR source, the UMR corpus is freely available for research.[7] We focused on split semantic roles, which present the greatest challenge for AMR to UMR conversion because a single AMR role can correspond to multiple UMR roles depending on context. This study provides the critical foundation for a comprehensive Persian UMR infrastructure. The immediate next step can be the systematic expansion of this initial corpus. This entails annotating the full range of UMR tags not covered in this study, particularly those for document-level semantics such as coreference chains, event temporality, and cross-sentence modality; also, the alignment between surface tokens and UMR nodes cannot be extracted from AMR data and will have to be obtained using other methods.

7. Ethics Statement

We are not aware of any ethical concerns related to this work. The corpus was manually annotated as part of academic research. In addition to the primary annotator, independent linguistically trained annotators contributed to annotation validation and agreement assessment. All annotation work was conducted voluntarily for research purposes, and no sensitive or personal data were involved, as the corpus is based on a publicly available literary text.

8. Limitations

At the present stage, the resource does not provide the full range of annotations specified in UMR guidelines; we list the main omissions in Future Work. This is a limitation, but given the complexity of UMR, the same limitation currently applies even to some datasets in the official UMR 2.0 release.

9. Acknowledgements

The work described herein was supported by the Charles University, project GAUK No. 394625. The second author was also funded by *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic. This research was also partially supported by SVV project number 260 821.

[7] http://hdl.handle.net/11234/1-6135

10. 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.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. 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, Marseille, France. European Language Resources Association.

Julia Bonn, Skatje Myers, Jens E. L. Van Gysel, Lukas Denk, Meagan Vigus, Jin Zhao, Andrew Cowell, William Croft, Jan Hajič, James H. Martin, Alexis Palmer, Martha Palmer, James Pustejovsky, Zdenka Urešová, Rosa Vallejos, and Nianwen Xue. 2023a. 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, Washington, D.C. Association for Computational Linguistics.

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

Qin Gao and Stephan Vogel. 2011. Corpus expansion for statistical machine translation with semantic role label substitution rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 294–298, Portland, Oregon, USA. Association for Computational Linguistics.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, Alfio Gliozzo, Sairam Gurajada, Hima Karanam, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yunyao Li, Francois

Luus, Ndivhuwo Makondo, Nandana Mihindukulasooriya, Tahira Naseem, Sumit Neelam, Lucian Popa, Revanth Gangi Reddy, Ryan Riegel, Gaetano Rossiello, Udit Sharma, G P Shrivatsa Bhargav, and Mo Yu. 2021. Leveraging Abstract Meaning Representation for knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3884–3894, Online. Association for Computational Linguistics.

Simin Karimi. 2008. *A minimalist approach to scrambling: Evidence from Persian*, volume 76. Walter de Gruyter, Berlin, New York.

Gh. Karimi-Doostan. 1997. *Light Verb Constructions in Persian*. Ph.D. thesis, Essex University, England.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to PropBank. In *LREC*, pages 1989–1993.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *LREC*, pages 1027–1032. Genoa.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Azadeh Mirzaei and Amirsaeid Moloodi. 2016. Persian proposition bank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3828–3835.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 1130–1139.

Claire Benet Post, Marie C McGregor, Maria Leonor Pacheco, and Alexis Palmer. 2024. Accelerating UMR adoption: Neuro-symbolic conversion from AMR-to-UMR with low supervision. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations@ LREC-COLING 2024*, pages 140–150.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 306–314.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Zacchary Sadeddine, Juri Opitz, and Fabian Suchanek. 2024. A survey of meaning representations – from theory to practical utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2877–2892, Mexico City, Mexico. Association for Computational Linguistics.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Jan Štěpánek, Daniel Zeman, Markéta Lopatková, Federica Gamba, and Hana Hledíková. 2025. Comparing manual and automatic UMRs for Czech and Latin. In *Proceedings of the Sixth International Workshop on Designing Meaning Representations*, pages 1–12, Prague, Czechia. Association for Computational Lingustics.

Reza Takhshid, Razieh Shojaei, Zahra Azin, and Mohammad Bahrani. 2022. Persian abstract meaning representation. *arXiv preprint arXiv:2205.07712*.

Nasim Tohidi, Chitra Dadkhah, Reza Nouralizadeh Ganji, Ehsan Ghaffari Sadr, and Hoda Elmi. 2024. Pamr: Persian abstract meaning representation corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–20.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

11. Language Resource References

Bonn, Julia and Bonial, Claire and Buchholz, Matt and Cheng, Hsiao-Jung and Chen, Alvin and Chen, Ching-wen and Cowell, Andrew and Croft, William and Denk, Lukas and Elsayed, Ahmed and Fučíková, Eva and Gamba, Federica and Gomez, Carlos and Hajič, Jan and Hajičová, Eva and Havelka, Jiří and Havenmeier, Loden and Kilgore, Ath and Kolářová, Veronika and Kučová, Lucie and Lai, Kenneth and Li, Bin and Li, Jingyi and Lopatková, Markéta and MacGregor, Marie and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Myers, Skatje and Novák, Michal and O'Gorman, Tim and Pajas, Petr and Palmer, Alexis and Palmer, Martha and Panevová, Jarmila and Post, Benét and Pustejovsky, James and Sgall, Petr and Song, Jialin and Song, Li and Ševčíková, Magda and Štěpánek, Jan and Urešová, Zdeňka and Sun, Haibo and Sun, Yao and Vallejos Yopán, Rosa and Van Gysel, Jens and Vigus, Meagan and Wright-Bettner, Kristin and Wu, Jiawei and Xue, Nianwen and Xing, Dan and Xu, Keer and Xu, Zhixing and Yue, Liulu and Zeman, Daniel and Zhao, Jin and Zikánová, Šárka and Žabokrtský, Zdeněk. 2025. *Uniform Meaning Representation 2.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). PID http://hdl.handle.net/11234/1-5902.

Takhshid, Reza and Azin, Tara and Shojaei, Razieh and Bahrani, Mohammad. 2021. *Persian AMR Dataset*. GitHub. PID https://github.com/Persian-AMR/Dataset.