

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/377546172>


PAMR: Persian Abstract Meaning Representation Corpus

Article in *ACM Transactions on Asian and Low-Resource Language Information Processing* · January 2024
DOI: 10.1145/3638288

CITATION
1

READS
104

5 authors, including:




Nasim Tohidi

University of Michigan

20 PUBLICATIONS 87 CITATIONS

SEE PROFILE




Chitra Dadkhah

K. N. Toosi University of Technology

41 PUBLICATIONS 538 CITATIONS

SEE PROFILE




Reza NourAlizadeh Ganji

Khaje Nasir Toosi University of Technology

2 PUBLICATIONS 5 CITATIONS

SEE PROFILE



Ehsan Ghaffari Sadr

Khaje Nasir Toosi University of Technology

1 PUBLICATION 1 CITATION

SEE PROFILE



PAMR: Persian Abstract Meaning Representation Corpus

NASIM TOHIDI

Faculty of Computer Engineering, Artificial Intelligence Department, K. N. Toosi University of Technology, Tehran, Iran,
n.tohidi@email.kntu.ac.ir, ORCID: 0000-0003-4499-9947

University of Michigan, Ann Arbor, Michigan, US, tohidin@umich.edu

CHITRA DADKHAH

Faculty of Computer Engineering, Artificial Intelligence Department, K. N. Toosi University of Technology, Tehran, Iran,
dadkhah@kntu.ac.ir, ORCID: 0000-0002-9836-9388

REZA NOURALIZADEH GANJI

Faculty of Computer Engineering, Artificial Intelligence Department, K. N. Toosi University of Technology, Tehran, Iran,
reza.nooralizadehganji@email.kntu.ac.ir, ORCID: 0000-0002-8892-7796

EHSAN GHAFARI SADR

Faculty of Computer Engineering, Artificial Intelligence Department, K. N. Toosi University of Technology, Tehran, Iran,
ehsan.ghaffarisadr@email.kntu.ac.ir, ORCID: 0000-0003-0872-2393

HODA ELMI

Faculty of Computer Engineering, Artificial Intelligence Department, K. N. Toosi University of Technology, Tehran, Iran,
h.elmi@email.kntu.ac.ir, ORCID: 0000-0002-3812-360X

One of the most used and well-known semantic representation models is Abstract Meaning Representation (AMR). This representation has had numerous applications in natural language processing tasks in recent years. Currently, for English and Chinese languages, large annotated corpora are available. Besides, in some low-resource languages, related corpora have been generated with less size. Although, till now to the best of our knowledge, there is not any AMR corpus for the Persian/Farsi language. Therefore, the aim of this paper is to create a Persian AMR (PAMR) corpus via translating English sentences and adjusting AMR guidelines and to solve the various challenges that are faced in this regard. The result of this research is a corpus, containing 1020 Persian sentences and their related AMR which can be used in various natural language processing tasks. In this paper, to investigate the feasibility of using the corpus, we have applied it to two natural language processing tasks: Sentiment Analysis and Text Summarization.

Keywords: Abstract Meaning Representation, Persian, Text, Corpus, Low-resource Language, Natural Language Processing

1 INTRODUCTION

From a mere decade ago, there has been a surge in the Natural Language Processing (NLP) society's tendency to concentrate on language understanding. Additionally, the meaning representation issue has been considered one of the hot topics in the NLP field [1]. In this regard, various meaning representation models have been proposed. Thus, the sever demand for broad-coverage and more comprehensive semantic banks leads to launch of various projects, some of which are the Groningen Meaning Bank (GMB) [2], Universal Conceptual Cognitive Annotation (UCCA) [3], the Semantic Treebank (ST) [4], the Prague Dependency Treebank [5], Universal Networking Language (UNL) [6, 7] and Abstract Meaning Representation (AMR) Sembank [8].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2375-4699/2024/01-ART

<https://doi.org/10.1145/3638288>

Some of the main features of these projects are as follows:

- **Concept:** The overwhelming majority of models consider words of each language as its concepts. However, AMR and UNL models use Propositional Bank (PropBank) [9] frames and English WordNet¹ synonym sets (synsets) in this regard, respectively.
- **Relation:** Each model has its own approach for defining relations in each natural language. GMB applies VerbNet [10] roles, AMR applies frame-specific relations of PropBank [9] and UNL has assigned a set of more than 30 most used relations.
- **Entity:** Some of the mentioned models uses a certain number of pre-defined entity types. For instance, GMB and AMR use 7 and 80 entity types, respectively.
- **Format:** Semantics in GMB are formalized with Discourse Representation Theory (DRT) [11], using full First-Order Logic (FOL). Meanwhile, both ST and GMB exploit the universal quantifier concept. While most other models do not follow logical models necessarily.
- **Granularity:** GMB and UCCA projects concentrate on annotating short texts, in order that similar entities can take part in events defined in distinct sentences; the remaining projects work on annotating individual ones.
- **Derivation:** Both UNL and AMR ignore the relation between strings and their meanings. Although, GMB and ST annotate phrases and words instantly.
- **Automatic vs Manual:** The three AMR, UCCA, and UNL annotations are completely generated manually. In contrast, ST and GMB work automatically, however, the generated representations could be reviewed and refined by humans [12].
- **Bottom-up vs Top-down:** AMR annotators build meanings from the top-down approach quickly, by commencing with the primary gist of the sentence, however, the AMR Editor also authorizes bottom-up structures. UCCA and GMB annotators both follow the bottom-up approach.
- **Supplementary:** All the mentioned projects have annotation guidelines like [13], graphical Sembanking tools like [2], and Sembanks from various genres.

Overall, one of the most successful models in this field has been the Abstract Meaning Representation (AMR) [8] compared with other related models. To commence with, Banarescu et al. [8] introduced this representation model, which embeds several important semantic concepts like *co-references*, named entity relations, and semantic relations in one sentence. In other words, AMR was proposed in order to detect the semantic relation among words in the form of arguments and, in this regard, it uses the framesets of the English Propositional Bank [9]. This successful representation is robust and stores the mentioned information for each sentence in a rooted, labeled, directed, and acyclic graph. Moreover, the word *abstract* in AMR means that this representation does not represent any syntactic information about a sentence. Therefore, syntactically different sentences with the same meaning (synonym sentences) would be represented by exactly the same graph [8]. As a consequence, in the AMR annotation framework, words that do not contribute to the sentence's meaning, such as infinitive *to* in the English language, are not included in the AMR annotation.

This characteristic will result in remarkable improvements in several NLP tasks, mostly the ones that are based on semantic similarity between sentences [14, 15]. Further, two salient characteristics of AMR are as follows [8]:

1. The AMR graph could be interpreted easily by humans or automatically.
2. The base form of AMR is vastly reliant on the English language.

Currently, the available AMR corpus for English is large enough for almost any NLP task, including about 39,000 sentences each paired with their AMR graph. In addition, several papers have been published in recent years that aim in defining AMR for other languages, some of which have taken advantage of the annotation and the existing corpus for

¹ <https://wordnet.princeton.edu/>

English [16, 17]. On the other hand, some other research attempted to adjust the primary guidelines of AMR to non-English languages, using its cross-lingual features [18, 19].

Table 1 represents the distribution of different applications that AMR has been used in them. This table has been prepared by reviewing about 120 related papers published in different languages in recent years.

Table 1: The distribution of the different applications of AMR

Applications	Percent
Machine Comprehension	31 %
Text Summarization	18 %
Question Answering	18 %
Entity Linking and Linked Data	13 %
Machine Translation	9 %
Information Retrieval	9 %
Other	2 %
Total	100 %

As it can be seen in Table 1, AMR has been used in different applications of natural language processing [20-24]. Besides, there are different languages in which AMR has been applied like English [8], Chinese [25], German [16], Spanish [18], Portuguese [26], Japanese [27], and Turkish [28]. Even some studies tried to work with AMR in a multilingual approach, such as [29].

As the study done in [30], for the application of AMR in six main NLP tasks, proves and by reviewing the scientific works based on AMR and its positive effects on the quality of NLP tasks, the significance of creating similar corpora for other languages is undeniable. In other words, an annotated corpus not only gives reusable and comparative data in order to enhance the quality of the existing techniques or even develop new ones, but also it provides the opportunity for considering it as a benchmark for evaluating novel methods.

The wide range of AMR applications in various languages has encouraged us to produce a Persian AMR corpus. It is notable that considering Persian as a low-resource language, the main aim here is to show how an AMR dataset for Persian can be produced, what are main challenges will be faced in this regard, what are their solutions, and how it can be used in NLP tasks. However, the produced dataset, with its current size, like some existing similar datasets for other natural languages, is also useful and can be applied to various NLP tasks.

The rest of this paper is organized as follows. In Section 2, the related studies for other natural languages are presented briefly. Then in section 3, a brief introduction to AMR is presented. In section 4, the AMP corpus, its production phases, and the most significant challenges in the production process are explained. In section 5, the experiments are done and the application of PAMR in two downstream NLP tasks has been briefly explained. Section 6 gives a brief overview of the annotation guidelines and steps. Finally, section 6 gives the conclusion and possible future works.

2 RELATED WORKS

In recent decades, scientists who have studied semantic analysis of the whole sentence usually have used small and restricted-domain banks such as GeoQuery [31]. These corpora are not informative enough for pragmatic research. As mentioned before, AMR is one of the most successful models for the English language and it has been applied in several NLP tasks. Hence, projects that concentrate on generating AMR annotated corpus for both English and non-English languages will be elaborated in this section.

One of the primary works that attempted to generate an AMR annotated corpus for a language other than English was presented by Xue et al. [17]. The principal aim of this study was to assess the potential of AMR to be applied as an interlingua. In this regard, they selected 100 English sentences from the Penn Treebank and annotated them by AMR after which translated them into Chinese and Czech languages. These translated sentences were also annotated with AMR. Eventually, they compared the extent of compatibility of AMR between English and these two languages, and they recognized that Chinese was more compatible with English than Czech.

Consequently, some studies worked on Chinese AMR annotation and two years later in 2016, the Chinese version of *The Little Prince* Corpus along with the related annotation guideline for Chinese AMR (CAMR) was published [25]. The purpose was to produce a major CAMR Corpus, which, by applying the index of every word in a sentence, has the manual annotation of concept/relation-to-word alignments. CAMR is primarily based on the English annotation technique with some adjustments for managing special phenomena in the Chinese language. Due to the complexity and the extent of application of this language, in the following years, more research was done to make AMR even more compatible with this language [32]. The size of the CAMR corpus is 10,149 sentences [33, 34]. Undoubtedly, it results in a considerable parsing capability. In addition, further research is still ongoing such as Song et al. [35], which tried to improve the predicate lexicon in order to enhance the CAMR.

In 2015, an AMR parser for English was developed by Vanderwende et al. [16]. The goal was to convert representations in logic form into AMR. Besides, in this work, they created an AMR-annotated corpus for Japanese, French, Spanish, and German languages.

In 2018, an AMR parser for English was proposed by Damonte and Cohen. It exploited parallel corpora in order to train AMR parsers for Chinese, Italian, German, and Spanish languages. The final results confirmed that the proposed parsers could successfully handle structural variations among these languages. Additionally, they introduced a parser evaluation technique that does not require gold standard data for the target languages.

Later in 2018, some studies concentrated on producing AMR-annotated corpora for some specific languages, like Spanish. In this regard, an AMR annotation of *The Little Prince* book was generated manually by Migueles-Abraira et al. [18], by using the English AMR project guidelines. The fundamental aim was to investigate the English specifications and to provide some adjustments to handle the related phenomena in the Spanish language. By way of another example, the first corpus with AMR annotations in Portuguese was built by Anchiêta and Pardo [36]. Like the Chinese corpus, it has alignments from *The Little Prince* book for Portuguese and English. Hence, the approach included receiving the related AMR annotation for every sentence from the annotations of the English corpus and modifying it to become compatible with Portuguese. In this project, the Verbo-Brasil [37] was used in order to annotate certain concepts as the main lexical resource, which is considered an efficient alternative to the English PropBank for Portuguese.

In 2019, studies that try to produce AMR-based resources have been carried out in Vietnamese [38] and Turkish [39]. In the former study, the authors introduced a meaning representation label set by adjusting the English pattern and considering the particular features of the Vietnamese language. In the latter one, the authors defined the first Turkish AMR corpus through manually annotating a hundred sentences from the Turkish translation of *The Little Prince* book and comparing the outputs with the same sentences in the English similar corpus.

Later in 2019 and 2020, similar studies were done for the Korean language [40]. Introductory studies have been conducted to present some grammatical characteristics of the Korean language, like Case-stacking, Copula construction, and its negation, and to illustrate the way they can be addressed by the grammatical schema of AMR [41]. Following studies have developed Korean guidelines, as a result, it lays the foundation for building the Korean corpus including 1,253 sentences of *The Little Prince* novel.

In 2021, multilingual AMR systems were developed by Sheth et al. [42] by projecting English AMR annotations to languages with barely enough resources. They bootstrapped transformer-based multilingual word embedding, especially ones from cross-lingual RoBERTa. Besides, they proposed a new method for Non-English-to-English AMR alignment, applying the contextual word alignment between tokens of English and Non-English languages. As a result, they could reach vastly competitive results that surpassed the most successful reports for Chinese, German, Spanish, and Italian.

Recently, Vu et al. [43] represented legal texts in the form of AMR. They studied AMR parsing and generation methods in the legal domain. Plus, they introduced JCivilCode which is a human-annotated legal AMR dataset. They created and verified this dataset by a group of legal and linguistic experts. Moreover, they introduced their domain adaptation method and applied it in the training phase and decoding phase of a neural AMR-to-text generation model. According to their report, the method could improve the quality of text generated from AMR graphs compared to related works.

Considering all these efforts made in various languages, the importance of having an AMR corpus in each and every language for NLP to improve the quality of applications is undeniable. In light of this fact, in this research, a similar corpus has been produced for the Persian language.

3 ABSTRACT MEANING REPRESENTATION

There are two types of meaning representations: formal representations and distributed representations. AMR [8] is an example of the first one. It represents the semantic of an English utterance as a set of relations between predicates and entities, that are packaged in a graph-based format. In the graph, nodes represent entities, and predicates of the utterance. AMR uses a neo-Davidsonian format for predicate meanings and treats predicate as an atomic value. AMRs are rooted, labeled, directed graphs, that allow co-references to be modeled by reentrancy. To represent the human reading and writing structure, the AMR format uses the PENMAN notation [44]. Moreover, AMRs abstract away from morphological and syntactic levels. So, different sentences could have the same AMR, if they exactly have the same semantics, even with different syntactic structures [30]. As another result of it, no particular alignment between graph components and the related sentence has been provided. In addition, AMRs' predicates are annotated based on framesets that have been specified in Propbank [45]. In addition, the predicate-argument structure could be used widely in this representation. For instance, the word *teacher* is represented as “(person :ARG0-of (t / teach-01))”, which is equivalent to *person(teacher)* in FOPL.

An example of a representation generated using AMR for the sentence *The Japanese Government stated on April 8, 2002, its policy of holding no nuclear warheads* is shown in Figure 1.

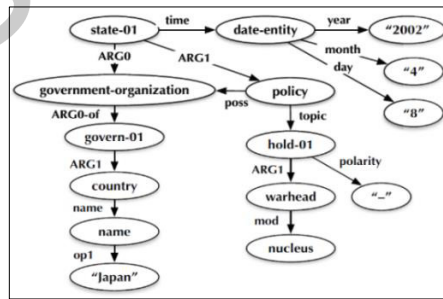


Figure 1: An example of AMR for an English sentence [46]

As it can be seen in Figure 1, in AMR graphs each node represents a concept (such as *policy* and *Japan*), and each edge represents the type of relationship between these concepts (such as *ARG1* or *mod*). It should be noted that in Figure 1 just the graphical view is shown, however, AMR has its corresponding frame representation (PENMAN notation), which makes it easier to apply the corpus in different applications. For instance, the related frame representation of the example in Figure 1 is illustrated in Figure 2.

```
(s / state-01
:ARG0 (g / government-organization
:ARG0-of (g2 / govern-01
:ARG1 (c / country
:name (n2 / name :op1 "Japan"))))
:ARG1 (p / policy
:poss g
:topic (h / hold-01 :polarity -
:ARG1 (w / warhead
:mod (n / nucleus))))
:time (d / date-entity :year 2002 :month 4 :day 8))
```

Figure 2: An example of the frame representation of AMR for an English sentence [46]

In general, nodes are recognized with their variable. For instance, the letter *s* is the instance of the entity *state* which is labeled as *state-01* in Figures 1 and 2. The variable *01* here points to the sense number of this concept in PropBank. Further, the labeled edges linking nodes are relations, like *ARG0*. Moreover, nodes that do not have variables are constants, such as *Japan*. They are commonly used to represent names, negations, or numbers. Usually, AMR concepts can be related to a single word in the sentence, that constitutes a one-to-one mapping (lexical concepts). However, sometimes there are concepts, which cannot easily be associated with any specific word in the sentence (abstract concepts). These concepts usually indicate inferred knowledge which is invoked by certain phrases or implicit relationships between disparate clauses. For instance, in Figures 1 and 2 the concept *country* is an inferred named entity type for *Japan*.

Albeit it has been claimed that AMR cannot be considered as an interlingua [8], the characteristic which it abstracts away from surface morpho-syntactic differences, makes it very attractive to implement cross-lingual AMR banks based on resembling principles.

4 THE PERSIAN AMR (PAMR) CORPUS

The Persian language has some characteristics that make it distinctive from some other natural languages [47, 48]:

- Persian is a pro-drop language with canonical Subject-Object-Verb (SOV) word order and various often exceptions that have made Persian an almost free word-order language. For example, in the sentence *the woman called the man*, the word order is SVO. In some languages, *the man* is in the objective case. However, English does not detect the objective either by marking or by inflection. Therefore, on condition that the word order is changed to OVS like *the man called the woman*, the meaning will completely change. This phenomenon does not exist in Persian, since «مرد» which in English means *the man* is marked in the objective case and as a result, words can be arranged in all other possible word orders without causing any change in the initial meaning like:

• زن مرد را صدا کرد . (SOV) *The woman called the man*
 • مرد را زن صدا کرد . (OSV) *The man was called by the woman*
 • صدا کرد زن مرد را . (VSO) *Called the man, the woman*

In these sentences, the word «زن», means *the woman*, is the subject, the word «مرد», means *the man*, is an object and the verb is «صدا کرد», meaning *called*.

- Verbs are marked for aspect, tense, agree with the subject in person, and number with some exceptions. For instance, in Persian, the perfect aspect describes a past action with relevance to non-past time (present or future) and it brings attention to the consequences of a past action. For example, *I have had lunch* points *I am not hungry now*. In addition, the grammatical person marks the participant(s) of an action. To be more precise, Persian has three degrees of grammatical persons: 1. first person (the person talking) 2. second person (the person being talked to) 3. third person (the person being talked about), and has two numbers: 1. singular 2. plural. Thus, it has six grammatical persons, and all these points are described by a verb. For instance, consider the verb «خورده ام» meaning *I have eaten*, this verb has perfect aspect and present tense. Besides, the suffix «ام» in this verb shows the first person and singular number.
- Persian is written right-to-left, and its Alphabetic letters have 1 to 4 writing forms, depending on the position of them within the word which may be initial, medial, or final (isolated). Besides, some sounds are denoted with more than one letter. For example, the same letter which denote /t/, can be written as «ت» in «مدت», as «تـ» in «توان» and as «تـ» in «کتان»; and the letters «ث», «س» and «ص» denote /s/.
- There are a lot of scripts for writing Persian texts, differing in using or omission of spaces between or within words, the style of writing words, using various forms of characters, etc. For instance, the verb meaning *I go* can be written as «می روم», «می روم» and «میروم».
- Commonly in Persian none of the short vowels are written in a sentence. Therefore, facing homographs and homonyms are examples of ambiguities in this language.
- There is no definite article in a Persian sentence, unlike English where most of the nouns appear with one in it.
- There is not any distinction between female or male pronouns in Persian and there is no rule for using uncountable nouns in singular form and words that are uncountable may even appear in plural form.
- Persian is a generative and derivational language that in it many new words may be created by concatenating words and affixes. For example, the word «همدردی» meaning *Sympathy* is made from three parts: «ی»+«درد»+«هم».
- In Persian words and phrases may be eliminated in a sentence regarding a semantic or syntactic symmetry. Besides, eliminating the subject is very usual in Persian sentences, and in these occasions the agreement between person and number which is embedded in the verb can play the subject role. For instance, the subject «من» (in English *I*) in the sentence «من مطالعه می کنم» (which means *I study*) can be omitted without changing the meaning or making any ambiguity.
- In Persian in several cases adjectives can be placed instead of nouns without any lexical change and this may lead to semantic or structural ambiguities in noun phrases. For example, the word «دارا» meaning *rich* or *wealthy* can be used as a noun like «داراها زندگی راحتتری دارند» (which means: *The wealthy have a more comfortable life*).

It is undeniable that working on Persian language processing is a developing field [49, 50]. However, several language resources, applications, and tools have been proposed and produced for Persian during recent decades, still, this language is far from languages like English in terms of current technologies and can be considered as a low-resource language in various tasks [47, 51].

This paper mainly aims to distinguish whether it is conceivable to annotate meaning representation for Persian sentences by considering the existing AMR guidelines. In the same vein, the crucial point is to identify aspects of meaning in Persian sentences which cannot be represented using the existing AMR guidelines. After specifying the related linguistic aspects, some guidelines should be modified or new ones should be introduced, thus, the AMR-based annotation of these sentences would be feasible. Here, to perform this analysis, like almost all related research in other natural languages, *The Little Prince Corpus* has been selected¹. The English version of it includes 1652 sentences of this

¹ The corpus is available in this link.

novel along with the related AMR for each of them. It is abundantly clear that working on this freely accessible corpus will give researchers the opportunity to compare various representations and results of applications in different natural languages of the same text.

4.1 Phases

In general, in this research, the production process of the PAMR corpus was divided into three phases, as can be seen in Figure 3. The first phase was related to translating English sentences to Persian sentences, which have been explained before. This phase was done by 10 experts in linguistics and translation. In the second phase, human agents were trained to annotate sentences which have fairly simple annotations and they started annotating the translated sentences of *The Little Prince* Corpus. Human agents in this phase were 3 Ph.D. and 37 Master students in NLP and linguistics. In this phase, firstly, some training sessions were held and as the agents have the linguistics background, in a couple of sessions they get familiar with AMR. Afterward, some training materials and example Persian sentences with their related representation were prepared and sent to agents to make sure that they are ready for doing the same process for new sentences. Then, each human agent was assigned some sentences and has been asked to upload the generated representation in a shared folder. Moreover, they could add comments on parts and cases that they face any challenge, or they were doubtful about the generated representation. The output of this phase was a set of initial representations that still needed further revision of 1020 sentences from this corpus.

Then the third phase was done by a limited number of human agents (2 Ph.D. students and 2 final-year Master's students) who were completely experts in this field, and they were familiar with AMR structure and specifications. These agents studied several papers and training materials related to annotating AMR corpora. Further, their thesis subject was related to meaning representation, so they had a strong background in this area. Each of them was assigned part of the generated corpus and they were asked to discuss each mistake that they face in a shared file to make the process of final review consistent in the whole corpus. They reviewed all the representations again and corrected them carefully. This effort, which took approximately 11 months (from March 2021 to January 2022), led to the production of the final corpus.

It should be noted that in all phases, the whole process was done under the profound and direct supervision of the NLP Lab head, Dr. Chitra Dadkhah, artificial intelligence department at K.N. Toosi University of Technology.

For example, the representation of the sentence *Why should any one be frightened by a hat?* was generated in phase 2 shown in Figure 4. As can be seen in the red parts, the representation is not accurate, as its concentration had been on the English sentence. In the English sentence the nature of the predicate “frighten” is passive, however, the corresponding predicate, «ترساندن» in the Persian translation is not. Further, the first argument of the predicate, «کلاه» is annotated wrongly. Therefore, in the 3rd phase, corrections are applied and the representation in Figure 5 was generated.

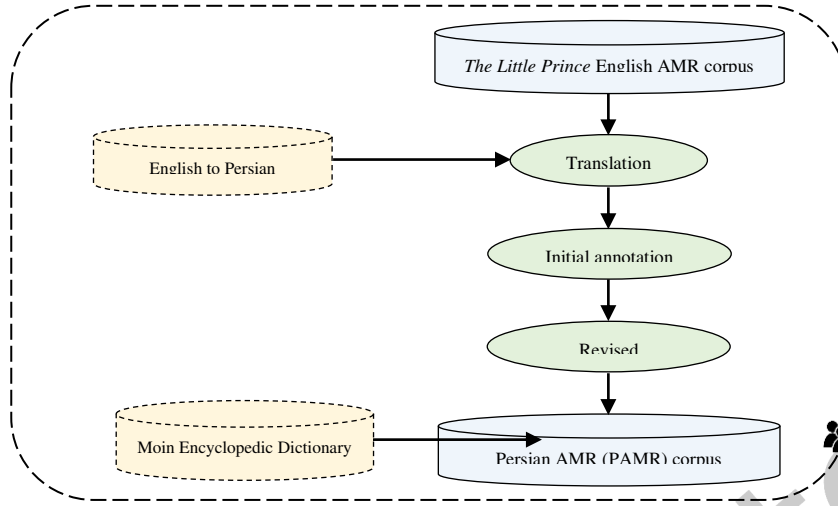


Figure 3: Phases of PAMR corpus production

12. Why should any one be frightened by a hat ? "

" چرا کسی باید از یک کلاه بترسد ؟

ت / ترساندن- (۰۱)

:ARG0 (ک / کلاه)

:ARG1 (ی / یک)

:mod (ه / هر)

:ARG1-of (د / دلیل)

:ARG0 (ب / بازنمایی-ناشناخته)

Figure 4: An example representation produced in phase 2.

12. Why should any one be frightened by a hat ? "

" چرا کسی باید از یک کلاه بترسد ؟

ت / ترسیدن- (۰۱)

:ARG0 (ش / شخص)

:ARG1 (ک / کلاه)

:mod (ی / یک)

:ARG1-of (د / دلیل)

:ARG0 (ب / بازنمایی-ناشناخته)

Figure 5: An example representation produced in phase 3.

Table 2 illustrates some of the most important features of the final produced corpus (PAMR). Meanwhile, considering the average length of the sentences, it can be concluded that on average the sentences represented in this corpus are not long.

Table 2: Features of PAMR corpus

Feature	PAMR
Number of sentences	1020
Average sentence length	12.84
The shortest sentence length	1
The longest sentence length	57
Number of distinct concepts	1638
Number of distinct relations	69

Table 3 presents the top 10 most frequent relations annotated in the PAMR corpus. It is worth noting that the order of the 10 most frequent relations in PAMR is very similar to *The Little Prince* English AMR corpus, which shows the similarity of semantic concept structures in both languages.

Table 3: 10 most frequent relations in PAMR corpus

Relation	Frequency	Percent
ARG1	1572	21.16
ARG0	1277	17.19
mod	676	9.10
ARG2	578	7.78
ARG1-of	378	5.09
op1	320	4.30
time	249	3.35
polarity	228	3.07
op2	224	3.01
degree	194	2.61

4.2 Challenges

During the construction process of this corpus, several challenges were raised, some of the most important of which are mentioned in the following.

Firstly, all the sentences of the corpus were translated into Persian. This process was not free of challenge, because there are several correct translations for each sentence available, and selecting one of them was not easy. An example is as follows:

English sentence: *Here you may see the best portrait that, later, I was able to make of him.*

Some possible translations in Persian are as follows:

- اینجا، شما می‌توانید بهترین تمثالی که من بعدتر توانستم از او بکشم را ببینید.
- شما ممکن است اینجا بهترین شکلی را ببینید، که بعدها توانستم از او بکشم.
- شما ممکن است اینجا بهترین تصویری را ببینید که من بعدتر توانستم از او بکشم.

Some of these differences in translation are related to synonymous words, some are related to different syntactic structures (which are not considered in AMR) and some are because of different ways of expressing a concept in Persian. For instance, in these three translations the word «اینجا», which means *here*, each time appears in different positions. Additionally, the word *portrait* has been translated to three different words («تصویر», «شکل», «تمثال») in each of them. Besides, the main verb «ببینید» (which means *see*) in the first translation is used at the end of the sentence, although in translations number 2 and 3 is used in the middle of the sentence.

Therefore, we decided to choose the closest translation, in terms of corresponding lexicons, to the English sentence in the corpus as the final selected sentence. In other words, the main attempt here was to translate each sentence word by word. Our purpose in applying this approach was to facilitate the process of generating representation for Persian sentences from the existing English representations, as when you have the same vocabularies in most cases the representations will be similar to a large extent. However, it was not possible to follow this approach in all English sentences. For example, about specific phrases, proverbs, and cases where the translation of a phrase into the target language did not have the same number of words (such as compound verbs that are very frequent in Persian, like the verb «زندگی کردن» that in English is equal to *live*), this approach could not be applied. Hence, the related representations of the Persian language were produced independently by experts. One related example from the corpus is shown in Figure 6.

<p>من مدت زیادی میان بزرگ‌ترها زندگی کرده‌ام .</p> <p>زندگی کردن-۰۱ / ز</p> <p>:ARG۰ (م / من)</p> <p>:mod (مدت / ۲م)</p> <p>:mod (زیاد / ۲ز))</p> <p>:location (میان / ۳م)</p> <p>:op۱ (ب / بزرگ ۰۲)</p> <p>:degree (تر / ۲ت)))</p>	<p># ::snt I have lived a great deal among grown - ups .</p> <p>(1 / live-01</p> <p>:ARG0 (i / i)</p> <p>:mod (d / deal</p> <p>:mod (g2 / great))</p> <p>:location (a / among</p> <p>:op1 (g / grown-up)))</p>
---	--

Figure 6: An example of differences that translation from English to Persian made in PAMR

In Figure 6, the phrase *grown-up* is translated to «بزرگ‌تر», thus, the related representation is different in these two languages. The main reason is that in Persian, this concept is expressed sequentially (similar to comparative and superlative adjectives in English), meaning that one person is more grown mentally than others, but in English the structure of the word is different. It should be noted that as this structure (the word «بزرگ» plus the suffix «تر» which in English is equal to the comparative suffix *er*) affects the meaning directly, it cannot be omitted from the representation. Meanwhile, as it is clear in Figure 4, in PAMR the infinitive form of verbs, such as «زندگی کردن» (in English *live*), is used. Moreover, prefixes and suffixes, like «ام», «ها», «تر», and «ی», have been removed from all words. In cases where prefixes and suffixes affected the meaning of the sentence and needed to be represented, the relevant structural standards were used in this regard, which an example is shown in Figure 7.

<p>۶۶. But I had never drawn a sheep . (lpp_۱۹۴۳.۶۶)</p> <p>اما من هرگز یک بره طراحی نکرده بودم .</p> <p>ت / تضاد</p> <p>:ARG۲ (ط / طراحی کردن-۰۱)</p> <p>:ARG۰ (م / من)</p> <p>:ARG۱ (ب / بره)</p> <p>time (ه / هرگز)</p> <p>polarity (-)</p>
--

Figure 7: An example of representing meaning of a prefix in PAMR

In Figure 7, it can be seen that the prefix «نـ» which makes the main verb «طراحی کرده بودم», in English *had drawn*, negative is removed in PAMR, however, its meaning is represented by the concept *polarity* and its value «-».

Moreover, here the word *but* is not mentioned as a concept directly, because in various sentences this meaning can be included with other words and phrases and in AMR the aim is to abstract the representation away from these structures.

Therefore, the meaning of the word «اما» (in English *but*) is represented with «تضاد» exactly like the English version which used the *contrast* word in this regard. This approach leads to have the same representation for the English sentence *However, I had never drawn a sheep.* or in Persian *ولی من هرگز یک بره طراحی نکرده بودم.* as they have the same meaning.

Secondly, we should handle the sense number challenge in generating AMR for Persian sentences. In this respect, at first glance, it seems that like the English corpus, we can use the Persian Propositional Bank (PropBank) [52]. This bank seemed the most suitable option because in various languages, like Chinese, Korean, Vietnamese, and Brazilian-Portuguese, AMR is created based on resources that are related to PropBank. Hence, this annotation schema could be beneficial in terms of adaptability between various language resources in different research. However, applying Persian PropBank was not possible for our purpose, because the structure of the Persian PropBank is completely different from the English one and there is not any sense list for each Persian concept in it. In order to tackle this problem, one of the most well-known Persian encyclopedic dictionaries, called Moin, is exploited. Thus, the produced PAMR annotation is fully based on the sense lists mentioned in this resource. In Figure 4, it can be seen that the 2nd sense of the word «بزرگ» (which means *big*) is used in the representation, as in Persian the first sense is related to size and the second one is about age.

Thirdly, another issue was that in AMR specifications for English, some pre-defined concepts are defined to express a particular concept or a series of particular entities. In this regard, during the annotation of the PAMR corpus, the (non-lexical) concepts and relations in the English AMR specifications are generally adopted. It should be noted that there are two groups of concepts in English AMR: lexical concepts and abstract concepts. The former ones are based on word tokens in each sentence and the latter ones are not connected to a particular lexical item. In general, lexical concepts are stem (or lemma) forms of word tokens with or without word sense numbers. Abstract concepts are derived from the context and are not explicitly related to a specific lexical item of a sentence. For instance, the word *country* could be an abstract concept when the word *Spain* is mentioned in a sentence. Although, all kinds of these concepts are not named entities and there are also some abstract and pre-defined concepts for time expressions, numbers, dates, besides concepts that represent discourse relations. Obviously, it is not possible to use the lexical concepts of English AMR for PAMR annotation, however, the abstract concepts have been indirectly applied in PAMR. To be more precise, in order to use these concepts in representations, equivalent pre-defined Persian concept names were used, which the most important ones can be seen in Table 4. The supplementary list of abstract concepts will be available in the PAMR annotation specification.

Eventually, other challenges were related to the writing structure of the Persian language. In Persian, some vowels are not written, and as a result, there are words in exactly the same written form and different pronunciations. For example, the word «گل» with one pronunciation can be equivalent to the word *flower* and with another pronunciation can be equivalent to the word *mud* in English. For handling this challenge, these words were treated exactly as words with different meanings, and they were annotated with sense numbers in the corpus. Other related challenges to the writing structure of the Persian language are the writing direction, which is from right to left, and the use of half-space in it, as well as the combination of this structure with the standard format of AMR in different languages. In other words, in a single representation of a Persian sentence, a combination of Persian words (concepts) and standard relationships defined with English phrases had to be used (like the representation illustrated in Figure 5). In this regard, a standard format was defined initially, and human agents were trained to create and review representations with this pre-defined structure. It is

worth noting that this format is fully similar to the English corpus, however, for coding and storing Persian sentences and representations in the text files the UTF-8 formatting¹ is used.

Table 4: Equivalent Persian concept names for the most used abstract concept names of English AMR.

English	Persian
have-rel-role	دارای-نقش-ارتباطی
Have-purpose	دارای-هدف
Temporal-quantity	کمیت-زمانی
Distance-quantity	کمیت-فاصله
Be-located-at	واقع-شدن-در
Amr-unknown	بازنمایی-ناشناخته
By-oneself	توسط-خود
Have-condition	دارای-شرط
Have-org-role	دارای-نقش-سازمانی
Have-quant	دارای-مقدار
Have-rel-role	دارای-نقش-ارتباطی
Have-polarity	دارای-قطبیت
Have-degree	دارای-درجه
Have-concession	دارای-اذعان
Have-frequency	دارای-تکرار
Instead-of	به-جای
Request-confirmation	درخواست-تأیید
Monetary-quantity	کمیت-پولی
Interrogative	پرسشی
Imperative	امری
Expressive	رسا
Multi-sentence	چند-جمله‌ای
Country-region	کشور-منطقه
Date-entity	موجودیت-تاریخ
Rate-entity	موجودیت-نرخ
Relative-position	موقعیت-نسبی

5 EXPERIMENTS

In order to show the usefulness and to evaluate the quality of the annotated dataset, we have applied the PAMR dataset in two downstream NLP tasks.

5.1 PAMR for Data Augmentation

Data Augmentation (DA) is one of the primary uses of AMR in NLP. The term data augmentation refers to a group of methods that, when applied to an existing dataset, can enable the generation of additional, synthetic data. This technique has recently garnered a lot of interest. Researchers that have trained NLP models using methods like Deep Neural Networks have paid a lot of attention to DA in the hopes of improving the model performance. Because it is believed that DA can aid in avoiding overfitting, addressing imbalances in datasets, and providing more data.

¹ <https://www.rfc-editor.org/rfc3629.html>

There is a possibility for DA algorithms to make use of the AMR tree structure nodes to create extra-textual data for each phrase in annotated PAMR corpora, mainly because AMR provides a high-quality semantic representation by way of a structured, directed graph. We reach our goal by associating each Persian sentence with its corresponding AMR tree in PAMR. The associated concepts with a sentence then can be extracted from the rooted directed graph and looked up for their synonyms in a thesaurus. It would be possible to repeat these steps as often as needed. Finally, to create fresh synthetic textual data, every term presented within a sentence would be swapped to its corresponding synonym.

The Hazm Python package¹ normalizes and tokenizes each Persian sentence in PAMR, while the Motaradef-Motazad thesaurus² in Persian is used to look for synonyms for each concept. Following the processing of all 1020 sentences of Persian in PAMR, a total of 888 additional textual records are produced with the assistance of AMR structures. Table 5 shows several samples of original Persian sentences and their related generated sentences.

Table 5: Original Persian sample sentences and their related generated sentences.

Source in English	Source in Persian	Generated
My drawing was not a picture of a hat .	نقاشی من تصویر یک کلاه نبود.	صورتگری من پرتره یک کلاه نبود .
So, then I chose another profession, and learned to pilot airplanes.	بنابراین، من حرفه دیگری را انتخاب کردم و خلبانی هواپیما را یاد گرفتم .	از این رو ، من پیشه دیگری را انتخاب کردم و خلبانی بالن را یاد گرفتم .
That is funny! "	این بامزه است ! "	این خوشطعم است ! "
I say plainly, " watch out for the baobabs! "	من به وضوح گفتم ، «مواظب کاج ها باشید»	من به وضوح گفتم ، « مترصد کاج ها باشید »
" Your cigarette has gone out. "	«سیگار ت خاموش شده است.»	« سیگار ت بی فروغ شده است . »
She was a coquettish creature!	او موجودی عشوهر گر بود .	او موجودی افسونگر بود .
And her mysterious adornment lasted for days and days.	و آرایش اسرار آمیز او روزها و روزها به طول انجامید .	و بزک پررازورمز او روزها و روزها به درازا انجامید .
But the conceited man did not hear him.	اما مرد مغرور به او گوش نکرد .	اما مرد پرادعا به او گوش نکرد .
And now I think I will go on my way. "	و حالا فکر می کنم که به راه خودم بروم .	و اکنون اندیشه می کنم که به جاده خودم بروم .
" Your planet is very beautiful, " he said.	او گفت : «سیاره شما بسیار زیبا است.»	او گفت : « اختر شما انبوه پریچهر است . »

The concept that additional generated text should be semantically like source textual data was respected by the similarity-based approach, which was utilized by us in the process of evaluating the DA method that had been applied previously. Then, to generate an embedding representation of each source and augmented sentence, we utilized a pretrained Persian language model known as ParsBERT [53]. After that, both the source sentences and the augmented sentences were input into ParsBERT, and the representation of those sentences was obtained from the attention outputs of classification tokens ([CLS]) in the final layer of ParsBERT. Finally, we determined the cosine similarity between the representations of the source texts and the augmented texts. According to our research, the average degree of similarity between all augmented sentences and the sentences that served as their inspiration is 90.87 percent, which demonstrates that the method of augmentation is effective.

¹ <https://pypi.org/project/hazm/>

² <http://www.fars-encyclopedia.com/modules.php?name=kitab&op=showbook&bid=57>

An additional downstream task, such as sentiment analysis, might benefit from the application of the DA technique that makes use of the suggested PAMR corpora. To depict the aforementioned usefulness, a Persian sentiment lexicon known as PerSent [54] is initially utilized to automatically annotate both the original and augmented data. After that, preprocessed texts that have already been labeled with a sentiment class are introduced into a deep learning model that is comprised of a ParsBERT [54] transformer module and multi-layer perceptron network. The purpose of this model is to identify the sentiment associated with the texts. And finally, a softmax layer is used to determine the classification probability of the subclass that correlates with it. Figure 8 shows the pseudocode for the sentiment analysis method.

Algorithm 1 Sentiment Analysis on PAMR corpora.

Input: source data X_s , sentiment lexicon L_s , opinion strength of the i th word in the sentiment lexicon $O_s^i \in [-1, +1]$, the initial model parameters θ^{init} , the maximum number of training epochs K .
Output: sentiment label y .

```

1: for each text in  $X_s$  do
2:   for each word in text do
3:     if word in  $L_s$  do
4:       assign opinion strengths  $O_s$  to each extracted word;
5:     else do
6:       assign zero polarity;
7:     end
8:   end
9:   annotate the sentiment class of a text-based on overall opinion strengths  $O_s$  of all words;
10: end
11: Train classifier model on annotated  $X_s$  with  $\theta^{init}$  and for  $K$  epochs;
12: Identify sentiment label ( $y$ ) of a text with trained classifier model;

```

Figure 8: The pseudocode of sentiment analysis task using PAMR corpora

The sentiment analyzer model underwent two training sessions: one using the original data and another using the augmented data. The resulting outcomes of the trained model on each dataset were obtained and subsequently presented in Table 6. The experimental results indicate that the utilization of the augmented data resulted in a notable enhancement of the model performance, with an increase of 12 percent observed in both the macro F1 and accuracy metrics. Under the assumption that the model's architecture, hyperparameters, and experimental conditions remained essentially the same, it can be argued that the improvements that have been observed can be solely attributed to the effectiveness of the employed data augmentation method, which was specifically designed based on the characteristics of the proposed PAMR corpora.

Table 6: the evaluation results of the sentiment analyzer on the original and the augmented data using PAMR

Dataset	Precision		Recall		F1-score		Accuracy
	Macro	Weighted	Macro	Weighted	Macro	Weighted	
Original	0.82	0.75	0.68	0.74	0.73	0.74	0.74
Augmented	0.83	0.86	0.88	0.86	0.85	0.86	0.86

5.2 PAMR for Automatic Text Summarization

Automatic Text Summarization (ATS) is divided into three categories: Extractive, Abstractive, and Hybrid. One of the methods in abstractive summarization is the use of Encoder-Decoder architecture [55]. However, the existing challenge of this method is that it does not provide explicit semantic modeling of the original and summarized document. The proposed mechanism to solve this problem is to use an abstract meaning representation of each sentence in document as Natural Language Generation (NLG) step of summarization.

The steps of the proposed summarization method based on PAMR are as follows:

1. Apply Girvan-Newman community detection algorithm on sentence-sentence network and determine centers of communities [56].
2. Each center of the communities is converted into an abstract meaning representation graph.
3. Combine the PAMR graphs and generate a single graph as a Unified-Graph.
4. Identify the effective nodes and edges of Unified-Graph.

After community detection algorithm applied, we select centers of each community to shorten the volume of input sentences to find important sentences. Each selected sentence should be parsed into its abstract meaning representation (graph representation). Two graphs can be connected to each other if they have at least one common node. For instance, *Nicole went home* and *Nicole wrote the letter* can be merged by means of the *Nicole* node. Thus, the graph of merging these two sentences contains 5 nodes: 1) *Nicole*, 2) *go*, 3) *write*, 4) *school*, 5) *letter*. For identifying the important sentences from centers of the detected community, we keep track of frequent bigrams in each sentence of the Unified-Graph and label each edge in Unified-Graph as the number of bigrams. For example, consider two sentences *Elizabeth went home* and *Elizabeth went to school*, in the Unified-Graph, the edge *Elizabeth-go* has a frequency of 2 as counter label. The important sentences are those that have more frequent edges in the Unified-Graph.

For the experiment, we chose Chapter 1 of *The little prince* book as dataset. Some examples from the first chapter of *The Little Prince* book are presented in Table 7. The sentences in this chapter are parsed using PAMR corpus and used as input for the summarization procedure. After parsing all sentences, a graph will be created. This graph will include nodes (words in the parse tree) and edges (relation of each word pair). After the construction of the Unified-Graph with a counter label on each edge (to count the repetition of each pair of words) a threshold will be applied to remove non-frequent edges. The remaining edges will be used to construct the summary.

Table 7: Examples of sentences from the Little Prince book.

Source in English	Source in Persian
It was a picture of a boa constrictor in the act of swallowing an animal.	آن تصویر ، عکس یک مار بوآ در هنگام بلعیدن یک حیوان بود.
In the book it said : " Boa constrictors swallow their prey whole , without chewing it.	در این کتاب گفته شده مارهای بوآ بدون اینکه طعمه های خود را بجوند ، به طور کامل آنها را می بلعند.
It was a picture of a boa constrictor digesting an elephant.	آن تصویری از یک مار بوآ بود که یک فیل را هضم می کرد.
But since the grown - ups were not able to understand it , I made another drawing : I drew the inside of the boa constrictor , so that the grown - ups could see it clearly.	اما چون بزرگترها قادر به درک آن نبودند ، من تصویر دیگری کشیدم : من درون مار بوآ را کشیدم ، تا بزرگترها آن را به وضوح ببینند.
My Drawing Number Two looked like this : The grown - ups ' response , this time , was to advise me to lay aside my drawings of boa constrictors , whether from the inside or the outside , and devote myself instead to geography , history , arithmetic and grammar.	نقاشی شماره دوی من اینگونه به نظر می رسید: این بار ، پاسخ بزرگترها این بود که ، به من توصیه کنند نقاشی های مار بوآ را ، چه از درون و چه از بیرون ، کنار بگذارم و وقت خود را صرف جغرافیا ، تاریخ ، حساب و دستور زبان کنم.
Then I would never talk to that person about boa constrictors , or primeval forests , or stars.	در نتیجه من هرگز راجع به مار بوآ یا درباره جنگل های بکر یا ستارگان با آنها حرف نمی زدم.

For evaluation, we consider ROUGE-1, ROUGE-2, and ROUGE-L as criteria. The precision of these metrics is 26.50, 25.80, and 26.50, respectively. In comparison with similar methods, it has a decrease in precision by about 15 percent, but it can have improvement by combining it with Encoder-Decoder architecture as mentioned before. It is worth noting that the difference of datasets in different languages is not ineffective in these results. The efficiency of using PAMR in automated text summarization can be seen in Table 8.

Table 8: The evaluation results with and without using PAMR in automated text summarization

Metric	With using PAMR	Without using PAMR
ROUGE-1	26.50	22.3
ROUGE-2	25.80	21.0
ROUGE-L	26.50	22.3

ROUGE-1, ROUGE-2, and ROUGE-L refer to the overlap of unigrams (each word), bigrams, and L-longest common subsequence between the system and reference summaries, respectively. The results in Table 8 clearly show the superiority of the PAMR-based technique for text summarization.

6 A BRIEF ANNOTATION GUIDELINE

As mentioned in the previous sections in detail, generally, for annotating more sentences and corpora similar to PAMR, the following steps should be taken:

- Selecting and training some human agents to be familiar with the AMR formal and conceptual structure.
- Training the human agents who are familiar with the general concept, with several examples, particularly the mentioned challenges.
- Selecting the desired text to be represented by AMR (preferably texts which are annotated by AMR in other languages to make evaluation simpler).
- Preparing a suitable translation for the text (if it is not in Persian).
- Providing the human agents with the translation and the needed resources for annotating.
- Assigning a desired number of sentences of the chosen text to the trained agents with a deadline. The assigned sentences to each agent should be consecutive and from the same topic to make the provided representation more accurate.
- Creating a shared folder among the agents and experts, so that all team members work on annotating in an integrated and consistent manner.
- Reviewing the provided representation by experts in the field and making the needed revisions.
- Applying the corpora on some NLP tasks and evaluating the corpus quality according to the experimental results.

7 CONCLUSION AND FUTURE WORKS

AMR is a successful annotation framework that was proposed in order to represent the meaning of a sentence in a natural language (English) with a directed, single-rooted and acyclic graph. The most important feature of it is that it abstracts away from the syntactic structure of a sentence.

This research aims to pave the way for the use of the AMR technique in the Persian language and for NLP tasks and applications in this natural language. The main goal was to investigate how to create PAMR corpus according to the current English corpus named *The Little Prince*. In the same vein, for PAMR annotation, the vocabulary and general approach used for annotating English AMRs are generally adopted. Though, the used predicate senses and argument labels have been taken from the Persian PropBank, most of which are vastly resembling in convention to that of the English PropBank.

The corpus produced in this research includes 1020 Persian sentences from the *The Little Prince* corpus, which has been prepared in three phases and after overcoming various challenges in a period of 11 months using two Moin and Aryanpour Persian dictionaries. These challenges are related to translation, sense numbers, English tags and expressions, and Persian writing structure. In this paper, we applied the corpus in two main NLP tasks: Sentiment Analysis and Text

Summarization, and studied the results. Hopefully, the outcome of this research and the produced corpus could have positive effects in improving the quality of NLP applications in the Persian language in the future.

For future work, the first step is to continue the annotation process and produce large corpora in various domains, which leads to easily use them in various NLP tasks, like, machine translation, automatic text summarization, and question answering systems. In this paper, the related information for generating other related corpora is discussed and by this experience, the team is well trained to produce future corpora faster and more accurately. As well, annotating the remaining sentences of the *The Little Prince* corpus and other existing corpora, such as the Bio AMR corpus¹, with PAMR will be vital and could help in making accurate comparisons of the final AMRs to the English and Persian versions. Besides, like AMR corpora in other languages, reviewing and updating the current representations is also crucial and should not be ignored. Finally, implementing an editor for PAMR annotation would extremely boost annotators.

REFERENCES

- [1] Tohidi, Nasim; Dadkhah, Chitra, "Integrated Semantic Representation (ISR-Model): Syntax-Independent Model for Natural Language," *Journal of Soft Computing and Information Technology*, vol. 12, no. 2, pp. 74-88, 2023.
- [2] Basile, Valerio; Bos, Johan; Evang, Kilian; Venhuizen, Noortje, "A platform for collaborative semantic annotation," in *In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012.
- [3] Abend, O.; Rappoport A., "UCCA: A Semantics-based Grammatical Annotation Scheme," in *In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, 2013.
- [4] Butler, Alistair; Yoshimoto, Kei., "Banking Meaning Representations from Treebanks," *Linguistic Issues in Language Technology*, vol. 7, no. 6, pp. 1-22, 2012.
- [5] Böhmová, Alena; Hajič, Jan; Hajičová, Eva; Hladká, Barbora, *The Prague dependency treebank*, vol. 20, Springer, 2003, p. 103–127.
- [6] Uchida, H.; Zhu, M.; Senta, T. D., "an electronic language for communication, understanding and collaboration," UNL: Universal Networking Language, IAS/UNU Tokyo, 1996.
- [7] R. Martins, "Le Petit Prince in UNL," in *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.
- [8] Banarescu, Laura; Bonial, Claire; Cai, Shu; Georgescu, Madalina; Griffitt, Kira; Hermjakob, Ulf; Knight, Kevin; Koehn, Philipp; Palmer, Martha; Schneider, Nathan, "Abstract Meaning Representation for Sembanking," in *In proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, 2013.
- [9] Kingsbury, Paul; Palmer, Martha, "From TreeBank to PropBank," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, 2002.
- [10] Palmer, Martha; Bonial, Claire; Hwang, Jena D., "VerbNet: Capturing English verb behavior, meaning and usage," in *The Oxford Handbook of Cognitive Science*, Oxford University Press, 2017.
- [11] Kamp, Hans; Van Genabith, Josef; Reyle, Uwe, "Discourse Representation Theory," in *Handbook of Philosophical Logic*, Springer, Dordrecht, 2011, pp. 125-394.
- [12] Venhuizen, Noortje J.; Brouwer, Harm, "Implementing Projective Discourse Representation Theory," in *The 18th Workshop on the Semantics and Pragmatics of Dialogue, SemDial 2014 - DialWatt*, 2014.
- [13] Knight, Kevin; Badarau, Bianca; Baranescu, Laura; Bonial, Claire; Bardocz, Madalina; Griffitt, Kira; Hermjakob, Ulf; Marcu, Daniel; Palmer, Martha; O'Gorman, Tim; Schneider, Nathan, "Abstract Meaning Representation (AMR) Annotation Release 3.0," Linguistic Data Consortium, Philadelphia, 2020.
- [14] Konstas, Ioannis; Iyer, Srinivasan; Yatskar, Mark; Choi, Yejin; Zettlemoyer, Luke, "Neural AMR: Sequence-to-Sequence Models for Parsing and Generation," in *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017.
- [15] J. Flanagan, "Parsing and Generation for the Abstract Meaning Representation," Carnegie Mellon University, Pittsburgh, 2018.
- [16] Vanderwende, Lucy; Menezes, Arul; Quirk, Chris., "An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus," in *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Denver, Colorado, 2015.
- [17] Xue, Nianwen; Bojar, Ondřej; Hajič, Jan; Palmer, Martha; Uřešová, Zdenka; Zhang, Xiuhong, "Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech," in *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.
- [18] Migueles-Abaira, Noelia; Agerri, Rodrigo; de Ilaraza, Arantza Diaz, "Annotating Abstract Meaning Representations for Spanish," in *Proceedings*

¹ <https://amr.isi.edu/download/2018-01-25/amr-release-bio-v3.0.txt>

of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2018.

- [19] Damonte, Marco; Cohen, Shay B., "Cross-Lingual Abstract Meaning Representation Parsing," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, 2018.
- [20] Pust, Michael; Hermjakob, Ulf; Knight, Kevin; Marcu, Daniel; May, Jonathan, "Parsing English into Abstract Meaning Representation Using Syntax-Based Machine Translation," in *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015.
- [21] Artzi, Yoav; Lee, Kenton; Zettlemoyer, Luke, "Broad-coverage CCG semantic parsing with AMR," in *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015.
- [22] Pham, Viet; Nguyen, Long H.B.; Dinh, Dien, "Semantic Convolutional Neural Machine Translation Using AMR for English-Vietnamese," in *CCCIS 2020: Proceedings of the 2020 International Conference on Computer Communication and Information Systems*, 2020.
- [23] Zhang, Zixuan; Ji, Heng, "Abstract Meaning Representation Guided Graph Encoding and Decoding for Joint Information Extraction," in *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, 2021.
- [24] Tohidi, Nasim; Dadkhah, Chitra, "A Short Review of Abstract Meaning Representation Applications," *Journal of Modeling & Simulation in Electrical & Electronics Engineering*, vol. 2, no. 3, pp. 1-9, 2022.
- [25] Li, Bin; Wen, Yuan; Bu, Lijun; Qu, Weiguang; Xue, Nianwen, "Annotating the Little Prince with Chinese AMRs," in *Proceedings of LAW X – The 10th Linguistic Annotation Workshop*, Berlin, Germany, 2016.
- [26] Cabezudo, Marco Antonio Sobrevilla; Pardo, Thiago, "Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese," in *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Italy, 2019.
- [27] W. Winiwarter, "JAMRED: a Japanese abstract meaning representation editor," in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, Brussels Belgium, 2015.
- [28] Heinecke, Johannes; Shimorina, Anastasia, "Multilingual Abstract Meaning Representation for Celtic Languages," in *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, Marseille, France, 2022.
- [29] Oral, Elif; Acar, Ali; Eryigit, Gülşen, "Abstract meaning representation of Turkish," *Natural Language Engineering*, vol. First View, pp. 1-30, 2022.
- [30] Tohidi, Nasim; Dadkhah, Chitra, "A Study on Abstract Meaning Representation Applications," in *The first Conference on Artificial Intelligence and Smart Computing*, Online, 2022.
- [31] Wong, Yuk Wah; Mooney, Raymond J., "Learning for semantic parsing with statistical machine translation," in *In Proceedings of the main conference on Human Language Technology, Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, New York, USA, 2006.
- [32] Bin, Li; Yuan, Wen; Li, Song; Li-jun, Bu; Weiguang, Qu; Nianwen, Xue, "Construction of Chinese Abstract Meaning Representation Corpus with Concept-to-word Alignment," *Journal of Chinese Information Processing*, vol. 31, no. 6, pp. 93-102, 2017.
- [33] Li, Bin; Wen, Yuan; Song, Li; Qu, Weiguang; Xue, Nianwen, "Building a Chinese AMR Bank with Concept and Relation Alignments," *Linguistic Issues in Language Technology*, vol. 18, no. 1, 2019.
- [34] C. Wang, "Abstract Meaning Representation Parsing," PhD thesis, Brandeis University, 2018.
- [35] Song, Li; Dai, Yuling; Liu, Yihuan; Li, Bin; Qu, Weiguang, "Construct a Sense-Frame Aligned Predicate Lexicon for Chinese AMR Corpus," in *The 12th Language Resources and Evaluation Conference*, Marseille, France, 2020.
- [36] Anchieta, Rafael; Pardo, Thiago, "Towards AMR-BR: A SemBank for Brazilian Portuguese Language," in *The Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.
- [37] Duran, Magali Sanches; Aluisio, Sandra M., "Automatic Generation of a Lexical Resource to support Semantic Role Labeling in Portuguese," in *The Fourth Joint Conference on Lexical and Computational Semantics*, Denver, Colorado, 2015.
- [38] Linh, Ha; Nguyen, Huyen, "A Case Study on Meaning Representation for Vietnamese," in *The First International Workshop on Designing Meaning Representations*, Florence, Italy, 2019.
- [39] Azin, Zahra; Eryigit, Gülşen, "Towards Turkish Abstract Meaning Representation," in *The 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Florence, Italy, 2019.
- [40] Choe, Hyonsu; Han, Jiyoung; Park, Hyejin; Oh, Taehwan; Park, Seokwon; Kim, Hansaem, "Establishment of Korean abstract semantic representation guidelines and corpus for graph structure representation of sentence meaning," *Journal of the Information Science Society*, vol. 47, no. 12, pp. 1134-1141, 2020.
- [41] Choe, Hyonsu; Han, Jiyoung; Park, Hyejin; Kim, Hansaem, "Copula and Case-Stacking Annotations for Korean AMR," in *The First International Workshop on Designing Meaning Representations*, Florence, Italy, 2019.
- [42] Sheth, Janaki; Lee, Young-Suk; Astudillo, Ramón Fernandez; Naseem, Tahira; Florian, Radu; Roukos, Salim; Ward, Todd, "Bootstrapping Multilingual AMR with Contextual Word Alignments," in *The 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, 2021.
- [43] Vu, Sinh Trong; Nguyen, Minh Le; Satoh, Ken, "Abstract meaning representation for legal documents: an empirical research on a human-annotated dataset," *Artificial Intelligence and Law*, vol. 30, no. 2, pp. 221-243, 2022.
- [44] W. C. Mann, "An overview of the Penman text generation system," in *In Proceedings of the Third AAAI Conference on Artificial Intelligence*, 1983.
- [45] Palmer, Martha; Gildea, Daniel; Kingsbury, Paul, "The Proposition Bank: An Annotated Corpus of Semantic Roles," *Computational Linguistics*, vol. 31, no. 1, pp. 71-106, 2005.

- [46] Liao, Kexin ; Lebanoff, Logan; Liu, Fei, "Abstract Meaning Representation for Multi-Document Summarization," in *The 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, 2018.
- [47] M. Shamsfard, "Challenges and Opportunities in Processing Low Resource Languages: A Study on Persian," *Computer Science*, 2019.
- [48] Basiri, Ehsan; Kabiri, Arman, "Words Are Important: Improving Sentiment Analysis in the Persian Language by Lexicon Refining," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 4, pp. 1-18, 2018.
- [49] Tohidi, Nasim; Dadkhah, Chitra; Rustamov, Rustam B., "Optimizing the Performance of Persian Multi-objective question answering system," in *The 16th International Conference on Technical and Physical Problems of Engineering*, Istanbul, Turkey, 2020.
- [50] Tohidi, Nasim; Dadkhah, Chitra; Rustamov, Rustam B., "Optimizing Persian multi-objective question answering system," *International Journal on Technical and Physical Problems of Engineering (IJTPE)*, vol. 13, no. 46, 2021.
- [51] Tohidi, Nasim; Hasheminejad, Seyed Mohammad Hossein, "A Practice of Humman-Machine Collaboration for Persian Text Summarization," in *The 27th International Computer Conference, the Computer Society of Iran*, Tehran (Virtually), 2022.
- [52] Mirzaei, Azadeh; Moloodi, Amiraeid, "Persian Proposition Bank," in *10th edition of the Language Resources and Evaluation Conference*, 2016.
- [53] Farahani, Mehrdad; Gharachorloo, Mohammad; Farahani, Marzieh; Manthouri, Mohammad, "ParsBERT: Transformer-based Model for Persian Language Understanding," *Neural Processing Letters*, vol. 53, p. 3831–3847, 2021.
- [54] Dashtipour, Kia; Raza, Ali; Gelbukh, Alexander; Zhang, Rui; Cambria, Erik; Hussain, Amir, "PerSent 2.0: Persian sentiment lexicon enriched with domain-specific words," in *Advances in Brain Inspired Cognitive Systems - 10th International Conference, BICS 2019, Proceedings*, Guangzhou, China, 2020.
- [55] Abolghasemi, Majid; Dadkhah, Chitra; Tohidi, Nasim, "HTS-DL: Hybrid Text Summarization System using Deep Learning," in *The 27th International Computer Conference, the Computer Society of Iran*, Tehran, Online, 2022.
- [56] Girvan, M.; Newman, M. E. J., "Community structure in social and biological networks," *Processing of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821-7826, 2002.