# Predicting Compositionality of Verbal Multiword Expressions in Persian

**Mahtab Sarlak*** and **Yaldasadat Yarandi*** and **Mehrnoush Shamsfard**
NLP Research Laboratory, Shahid Beheshti University
{ma.sarlak, y.yarandi}@mail.sbu.ac.ir
{m-shams}@sbu.ac.ir

## Abstract

The identification of Verbal Multiword Expressions (VMWEs) presents a greater challenge compared to non-verbal MWEs due to their higher surface variability. VMWEs are linguistic units that exhibit varying levels of semantic opaqueness and pose difficulties for computational models in terms of both their identification and the degree of compositionality. In this study, a new approach to predicting the compositional nature of VMWEs in Persian is presented. The method begins with an automatic identification of VMWEs in Persian sentences, which is approached as a sequence labeling problem for recognizing the components of VMWEs. The method then creates word embeddings that better capture the semantic properties of VMWEs and uses them to determine the degree of compositionality through multiple criteria. The study compares two neural architectures for identification, BiLSTM and ParsBERT, and shows that a fine-tuned BERT model surpasses the BiLSTM model in evaluation metrics with an F1 score of 89%. Next, a word2vec embedding model is trained to capture the semantics of identified VMWEs and is used to estimate their compositionality, resulting in an accuracy of 70.9% as demonstrated by experiments on a collected dataset of expert-annotated compositional and non-compositional VMWEs.

## 1 Introduction

In today's world, multiword expression detection and embedding are trending topics, particularly among the research conducted on natural language processing. Multiword expressions (MWEs) are word combinations that display some form of idiomaticity, in which the semantics of some of the MWEs cannot be predicted from the semantics of their component. These expressions comprised of at least two words, inclusive of a headword and syntactically related words that display some degree of lexical, morphological, syntactic, and/or semantic idiosyncrasy (Sag et al., 2002). In this paper, we focus on verbal MWE (VMWE) which is a multiword expression such that its syntactic head is a verb and its other components are directly dependent on the verb (Sag et al., 2002). Identifying a VMWE in a Persian sentence poses many challenges, like in other languages(Constant et al., 2017). One of the primary ones is the violation of the compositionality principle, leading to the inability to deduce the semantic meaning of the VMWE from the meanings of its individual components as shown in (1).

(1) دست روی دست گذاشتن
lit. put hand on hand
**doing nothing**

Discontiguous VMWEs pose an extra challenge, as shown in the example (2).

(2) او اقدام به خودکشی کرد
lit. he attempt to suicide did
he **attempted** suicide

In (2), identifying the compound verb "اقدام کرد" (attempt did => attempted suicide) becomes challenging through traditional approaches. Finally, the assignment of grammatical roles to certain word sequences can be entirely dependent on the sense of the words and the context in which they are used.

* These two authors contributed equally to this work



*Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 14–23

(3) او دستش را بلند کرد
lit. he his hand tall did
He **raised** his hand
(4) او آثار هنری را بلند کرد
lit. He artworks tall did
He **stole** the artworks

For instance, in (3) and (4), although the sense of the word "بلند" (tall) is the same in both examples, the expressions "بلند کرد" (I did tall) have different meaning depending on the context (raised and stole, respectively). Furthermore, representing VMWEs as unified units in embeddings is challenging due to the limitation of traditional static embeddings generating one embedding per token, while VMWEs consist of multiple tokens. Alternative representation methods need exploration. Additionally, as previously mentioned, VMWEs can possess both idiomatic and literal meanings, leading to syntactic ambiguity. This creates a problem for the generation of embedding vectors that accurately capture the semantic meaning of such expressions. **Contribution:** The contributions of this paper are two-fold. First, we propose non-contextual and contextual methods to identify VMWEs. For the non-contextual strategy, we use a VMWE dataset based on Persian WordNet, while LSTM and BERT models are used as the contextual methods. Though the BERT model uses contextual embedding for each word, our LSTM model has a non-contextual embedding layer in its network. In our second contribution, we aim to measure the degree of compositionality of a VMWE by analyzing the semantic similarity between its components and the expression as a whole. To do this, we utilize two word-level and character-level embedding methods: word2vec and fasttext, which capture the semantic meaning of the VMWEs by concatenating detected VMWEs in the training corpus. We then determine the compositionality of a VMWE by using six different metrics. Finally, we have gathered a dataset that includes around 55 VMWEs, which have been tagged as either compositional or non-compositional, to evaluate the accuracy of our predictions.

In Section 2, a review of existing methods is presented. The proposed algorithm for identification and prediction of compositionality is detailed in Section 3 and 4, respectively. The effectiveness of the introduced approaches is assessed through experiments, the results of which are presented in Section 5. Finally, in Section 6, the results are discussed and concluding remarks are drawn.

## 2 Related Work

**VMWEs identification:** There are generally two types of methods to identify VMWEs in a sentence: language-dependent and language-independent methods. In terms of language-dependent methods, (Chaghari and Shamsfard, 2013) introduced an unsupervised method to identify Persian VMWEs by defining a set of linguistic rules. (Saljoughi Badlou, 2016) also introduced a language-dependent method to identify Persian MWEs by creating regular expressions by Persian linguistic rules and searching extracted MWEs from Wikipedia article titles and FarsNet (Shamsfard, 2007). Moreover, (Salehi et al., 2012) introduced a method that utilized a bilingual parallel corpus and evaluated the efficacy of seven linguistically-informed features in automatically detecting Persian LVCs with the aid of two classifiers.

In recent years, deep learning has demonstrated remarkable success in sequence tagging tasks, including MWE identification (Ramisch et al., 2018; Taslimipoor and Rohanian, 2018). RNNs and ConvNets have shown significant progress in this area. (Gharbieh et al., 2017) achieved their best results on the DiMSUM (Schneider et al., 2016) dataset using a ConvNet architecture to identify MWEs. (Taslimipoor and Rohanian, 2018) proposed a language-independent LSTM architecture to identify VMWEs, which includes both convolutional and recurrent layers, and an optional high-level CRF layer. Additionally, (Rohanian et al., 2020) focused on using MWEs to identify verbal metaphors and proposed a deep learning model based on attention-guided GCNs, which incorporate both syntactic dependencies and information about VMWEs.

Supervised techniques like deep learning require vast amounts of labeled data. The fine-tuning step of the BERT model has the capability to tackle this issue, making it a powerful tool. ParsBERT, developed by (Farahani et al., 2021), is a monolingual Persian language model based on Google's BERT architecture that utilizes the same BERT-Base settings. It was trained on over 2 million diverse documents, allowing it to perform various tasks, including sentiment analysis, text classification, and named entity recognition.

**VMWEs compositionality prediction:** Compositionality prediction of MWEs has garnered considerable attention in recent years. One popular method for measuring the compositionality of MWEs is through the use of word embeddings. (Salehi et al., 2015) were among the first to explore this approach by comparing the performance of two embedding models, word2vec and MSSG, in predicting the degree of compositionality of MWEs in English and German datasets. Their hypothesis was that the similarity between MWEs and their component words' embedding vectors would be indicative of the MWEs' compositionality. They then found that combining string similarity with the word embedding approach was comparable to existing state-of-the-art methods (Salehi and Cook, 2013). A study by (Nandakumar et al., 2018) provides a similar examination, using word-level, character-level, and document-level embeddings to calculate the compositionality of MWEs in English. Their results suggest that the word2vec (Mikolov et al., 2013) model, followed by fasttext (Bojanowski et al., 2017) and infersent(Conneau et al., 2017), outperformed other embedding models. (Cordeiro et al., 2019) improved that method and proposed that multi-word expressions (MWEs) should be preprocessed into a single unit prior to model training. This has a drawback that a comprehensive list of MWEs be available beforehand to accurately identify and consolidate them into a single token. Additionally, any alterations to the set of MWEs would mandate retraining of the model. Consequently, this study aims to determine the degree of compositionality of each VMWE by first identifying them and training an embedding model to capture their semantic information. The resulting embedding vectors are then utilized to predict the compositionality of each VMWE.

Despite numerous studies on predicting MWEs compositionality, much of the research has been concentrated on English and European language corpora. To the best of our knowledge, there has been no investigation on compositionality prediction of VMWEs in Persian, which is a low-resource language. Thus, in this work, we aim to address these two issues by leveraging the methods established in previous MWE studies.

## 3 VMWE Identification

In this section, we first present the datasets utilized in the proposed approach for VMWE identification, followed by a detailed description of the methods and models employed for this task. To detect VMWEs, a combination of a non-contextual method and two deep learning models are employed. These deep learning models treat the VMWE detection task as a sequence labelling problem, where the goal is to assign a relevant tag to each token in the sequence. To accomplish this, an IOB-like labelling format was used to tag the VMWEs in sentences, where the beginning component of the expression is tagged as 'B', its other components are tagged as 'I', and the words in the sentence that do not belong to any VMWE receive an 'O' tag. Additionally, sentences containing two VMWEs with mixed components were removed for simplicity (e.g. 5). The two deep learning models used are an LSTM-based architecture and a BERT-based model.

(5) در تمام طول زندگی اش نقش بازی کرد
lit. in all length his life role play did
He impersonated during all his life
VMWE1 : بازی کرد (play did => play)
VMWE2 : نقش بازی کرد (role play did => impersonate)

### 3.1 Dataset for the identification of VMWE

In terms of datasets, the Parseme Corpus (Savary et al., 2017) serves as the annotated corpus of tagged VMWEs, comprising 3226 sentences. The VMWEs in this corpus were manually annotated by a single annotator per file. Every verb-particle construction (VPC) that is fully non-compositional, where the particle modifies the meaning of the verb, is tagged, and a number bonds the components of the VMWE. Additionally, Persian Dependency Treebank (PerDT) contains 30 thousand tagged sentences (Rasooli et al., 2013). PerDT was tagged using both rule-based and manual strategies. The first strategy utilized the dependency tree to identify the components of VMWEs by extracting words with LVP[1], NVE[2], and VPRT [3] tags and their connected verbs, resulting in the detection of 32056 VMWEs in the training set of the corpus. A manual annotation of VMWEs was also performed on 1000 sentences of

[1] Light Verb Particle
[2] Non-Verbal Element
[3] Verb-Particle Construction

the corpus. Although this method resulted in fewer tagged sentences, it was more accurate and reliable compared to the previous strategy. We evaluated our non-contextual method on the Parseme Corpus and trained neural networks on both corpora.

### 3.2 Non-contextual method

The first strategy for identifying VMWEs

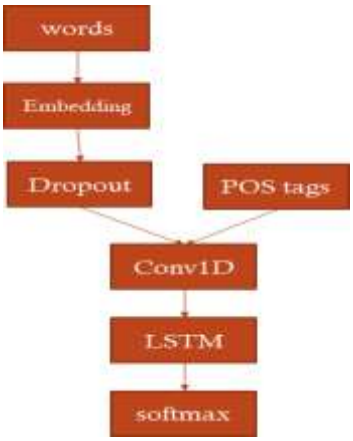


Figure 1: The architecture of ConvNet + LSTM

involves a straightforward approach that seeks to identify such expressions within a sentence. To achieve this, a dataset of VMWEs was created by collecting all compound verbs in FarsNet, which is the Persian wordnet With 100,000 words developed by the natural language processing laboratory at Shahid Beheshti University. We extracted 21462 VMWEs from FarsNet. To identify VMWEs in a sentence, the n-grams (for n=2,3,4) were extracted and searched for the presence of all components of a multi-word verb within the n-gram. Not all cases that are found are VMWEs, and not all VMWEs can be found in this way, especially if there are intermediate words. However, this approach can help identify potential VMWEs. The effectiveness of this approach will be evaluated in the evaluation section.

### 3.3 Long Short-Term Memory (LSTM)

A neural network architecture comprised of a convolution network and an LSTM network was utilized. The network was designed with an embedding layer as the initial component, which is demonstrated to produce better results than utilizing a standalone embedding model. To enhance the accuracy of predictions, the inputs to the network were augmented with POS tags. The architecture of the layers is illustrated in Figure 1. The first layer encompasses a combination of token vectors derived from the embedding layer, concatenated with 50-dimension features and a dropout rate of 0.2. The output of this layer and the POS tags were then concatenated as a numerical code at the end of the embedding vector of each word and then, fed into a ConvNet layer containing 200 neurons and a filter size of 1. No dropout was applied to the ConvNet layer and the activation function used was Rectified Linear Unit (ReLU). The output of the convolutional layer was then fed into a bi-directional LSTM network with 100 neurons and a recurrent dropout rate of 0.5.

### 3.4 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained neural model based on self-attention blocks. It has achieved state-of-the-art results on various natural language processing tasks, such as question answering (Devlin et al., 2018) and Multi-Genre Natural Language Inference (Nangia et al., 2017), due to its ability to embed each token in a sentence contextually, it can capture the meaning of each token within its context. The advantage of BERT is that it is a general architecture that can be applied to multiple problems, and its pre-training on raw, unlabeled texts minimizes the need for labeled data. Additionally, BERT has been pre-trained in 104 languages, including Persian. In this study, we utilize the ParsBERT model, pre-trained on Persian text, to identify VMWEs in Persian sentences. The ParsBERT model is fine-tuned on datasets specifically for the task of tagging tokens that are part of a VMWE.

## 4 Predicting the Compositionality of VMWEs

The primary objective of this paper is to predict the compositionality of VMWEs. Our assumption is that the degree of compositionality of a multiword expression can be determined by evaluating the semantic similarity between its constituent components and the expression itself. This evaluation is conducted by comparing the similarity of the embedding vectors of the corresponding word tokens. To accomplish this, we follow the studies of (Salehi et al., 2015) and (Nandakumar et al., 2018) and investigate six metrics to determine the compositionality of

VMWEs. In this section, the criteria for the task and a description of the datasets are presented.

### 4.1 Methodology

One of the defining challenges of VMWEs is their compositional nature, where the semantic meaning of a VMWE can be dissimilar from the meanings of its individual components. Therefore, the objective of this research is to determine the degree of compositional property by analyzing the embedding vectors of both the VMWEs and their components.

We begin with the preparation of four different corpora for training embedding models. The detected VMWEs are pre-processed by removing all spaces and semi-spaces[4], and replacing them with an underscore symbol to consider the VMWE as a single word. Two word-level and character-level embedding models, namely word2vec and fasttext, are then trained on the processed corpora.

To assess the compositionality of the VMWEs, six different criteria are leveraged to predict the compositionality of the VMWEs based on the generated VMWE-specific embedding vectors. It is assumed that the compositionality of an MWE can be captured by computing the relative similarity between the MWE's component embedding vectors and the embedding vector of the MWE. Consequently, the majority of the proposed metrics focus on calculating this similarity, followed by the determination of a threshold that indicates whether a VMWE is compositional or not based on the computed metric value. We compare the performance of different criteria in distinguishing compositional and non-compositional VMWEs. All similarity calculations between two vectors are performed using cosine similarity. Additionally, the embedding models are trained on the original corpora to obtain the embedding vectors of all VMWE components. In this study, the overall compositionality of VMWEs is computed using six metrics. In order to evaluate the used embedding vectors, we introduced a new metric called Syn_sim. This is in addition to two previously introduced metrics, Direct_pre and Direct_post, by Salehi et al. (2015) and Nandakumar et al. (2018). Furthermore, (Rossyaykin and Loukachevitch, 2019) and (Loukachevitch and Parkhomenko, 2018) proposed DFsing and DFsum, while Loukachevitch and Parkhomenko (2018) suggested DFcomp. These criteria are explained in more detail.

**Syn_sim:** Intuitively, we can demonstrate that an embedding effectively captures the semantic meaning of a VMWE if it's similar to the embedding vector of that VMWE's synonymous simple verb, which is extracted through Farsnet. We directly compare two different similarity metrics: (1) the similarity between the VMWE's embedding vector and that of the synonymous simple verb; and (2) the similarity between the synonymous verb and 'combined' vector, which is computing an element-wise sum over VMWE's components embedding vector. We calculate these two similarities of the embeddings of the VMWE and its synonymous simple verb using the following three formulas:

$$combined_{vector} = \sum_{i=1}^{N} w_i \quad (1)$$

$$sim\_syn\_vmwe = cos(vmwe,\ syn\_verb_1) \quad (2)$$

$$sim\_syn\_combined = \cos(combined_{vector},\ syn\_verb_1) \quad (3)$$

Where: vmwe, $w_i$, and $syn\_verb_1$ are the embeddings for the VMWE, i-th components of VMWE, and synonymous simple verb, respectively. In all cases, if the sim_syn_vmwe is greater than the sim_syn_combined, it means that the constructed VMWE's vector provides a better representation than the combined vector; Thus, the use of the introduced embedding model leads to a better result as it produces better semantic-aware representation for VMWEs.

**Direct_pre:** Assuming that compositional VMWEs tend to have a similar context with their components, we compare the vector embedding of the VMWE with the 'combined' vector of its components by calculating the cosine similarity between them. Formally:

$$direct\_pre = cos(vmwe, combined_{vector}) \quad (4)$$

**Direct_post:** The similarity between the vector embedding of a VMWE and each of its components is first measured. Then the overall compositionality of the VMWE is computed by combining the similarity scores below.

$$direct_{post} = \alpha\ cos(vmwe, w1) + (1 - \alpha) * \cos(vmwe, w2) \quad (5)$$

[4] In Persian typography, a semi-space is a zero-width-space character that separates two sides without leaving any space between them.

Where w1 and w2 denote the embedding for the first and second component of the VMWE, respectively. Here, we assume that the VMWE consists of two components as most of Persian VMWEs are light verb constructions (LVCs), but the formula can be easily generalized to concider more than two components.

**DFsum:** The similarity between the vector embedding of a VMWE and the element-wise sum of normalized vectors of its components is computed. Formally:

$$combined\_vector\_norm = \sum_{i=1}^{N} \frac{w_i}{|w_i|} \quad (6)$$

$$DFsum = cos(vmwe, combined\_vector\_norm) \quad (7)$$

**DFcomp:** The similarity between the VMWE's components' word vectors is computed. Formally:

$$DFcomp = cos(w_1, w_2) \quad (8)$$

**DFsing:** The similarity between the vector embedding of a VMWE and the vector of the most similar single word (sim_word) is calculated as below :

$$DFsing = cos(vmwe, sim\_word) \quad (9)$$

### 4.2 Dataset for compositionality prediction

For our experiment, we use four current Persian corpora, namely Bijankhan, HmBlogs, PARSEME, and PerDT to statistically study the occurrences of VMWES in Persian texts.

**Bijankhan:** The dataset of Bijankhan is a tagged corpus that is gathered from daily news and common texts (Bijankhan, 2004). This corpus contains about 2.6 million tagged words with 550 Persian part-of-speech tags.

**HmBlogs:** A tokenized corpus of 500 million sentences and 6.5 billion tokens is gathered by (Khansari and Shamsfard, 2021) We use the first 1 million sentences of it.

**Compositional and non-compositional VMWE dataset:** A self-gathered dataset of compositional and non-compositional verbs was identified by linguists, which annotated for compositionality on a binary scale. According to (Karimi, 1997) and (Sharif, 2017), 33 compositional and 22 non-compositional verbs were extracted in an infinitive form.

## 5 Results and Discussion

This section showcases the evaluation outcomes achieved during the testing phase for identifying VMWEs and predicting their compositionality. The evaluation was performed on the Parseme corpus test-set for all identification techniques.

### 5.1 VMWE Identification Evaluation

We trained our identification networks using the Parseme and PerDT corpora, identifying 2451 VMWEs and 1669 unique ones in Parseme, and using IOB format for tagging. We also tagged and used VMWEs from PerDT for the train set. Table 1 and Table 2 specify the results. The first row of Table 1 shows the results of the non-contextual

| | **Token_based** | | | **VMWE_based** | | | **Sentence_based** |
|---|---|---|---|---|---|---|---|
| | p | r | **f1** | **p** | **r** | **f1** | **accuracy** |
| **Non-Contextual** | - | - | - | 34.19% | 43.71 % | 38.36% | - |
| **LSTM** | 61.50%<br>69.95%<br>63.39% | 49.23 %<br>50.40 %<br>51.03 % | 54.71 %<br>58.59 %<br>56.54 % | 72.00%<br>85.52%<br>72.34% | 60.07 %<br>63.67 %<br>61.07 % | 65.50%<br>72.99%<br>66.23% | 51.11%<br>58.05%<br>53.61% |
| **BERT** | 94.04%<br>90.34%<br>94.88% | 84.25 %<br>74.54 %<br>77.86 % | 88.87 %<br>81.68 %<br>85.53 % | 92.37%<br>91.43%<br>93.25% | 85.99 %<br>77.90 %<br>79.09 % | 89.07%<br>84.13%<br>85.59% | 71.38%<br>63.88%<br>68.61% |

Table 1: VMWE Identification Results

method on the Parseme dataset. For the other rows, the first row of each method was trained on Parseme corpus, while the other rows used both corpora to train the models. However, the second and third rows consider the rule-based and manually tagged PerDT, respectively. It is not surprising that contextual methods utilizing neural networks exhibit a substantial improvement over non-contextual methods. The LSTM model performs relatively better with a train-set size increase, achieving about 73% F1-score. The BERT model has the highest F1-score of 89.07% on the PARSEME train-set. The BERT model performs better on PARSEME due to inaccuracies in manual and rule-based tagging methods, caused by the absence of expert annotators and limited expert evaluation. Additionally, BERT's sensitivity to incorrect data is higher than the LSTM model as it is pre-trained on Persian, resulting in lower performance for the second and third rows.

| | Seen proportion | CDSV | CDUV |
|---|---|---|---|
| **1** | 33.33% | 89.00% | 62.56% |
| **2** | 73.12% | 80.42% | 46.75% |

Table 2: Proportion of seen VMWEs in Parseme and the percentage of correct detection of seen(CDSV) and unseen verbs(CDUV)

We also analyzed the results based on seen and unseen verbs. Table 2 shows the evaluation results of the best model (BERT fine-tuned on Parseme) on seen and unseen verbs by two approaches.

- We considered seen verbs as verbs whose exact forms (like their persons, tenses etc.) exist in the train set.
- For finding seen verbs, we turn the core (the main verb) of all verbal expressions in the test and train set to their infinitive form and then check whether the expression exists in the train set.

### 5.2 Compositionality Prediction of VMWEs

| Criterion | threshold | accuracy |
|---|---|---|
| Direct pre | 0.23 | **0.709** |
| Direct post | 0.27 | 0.655 |
| DFcomp | 0.23 | 0.618 |
| DFsum | 0.23 | **0.709** |

Table 3: Evaluation results of the criteria

The experiments began with analysing the top most similar words or expressions to some of the frequent VMWEs to find the best embedding model capable of capturing VMWE's semantics. By increasing the corpus size, we observe that the top most similar expressions of a VMWE are closer to the meaning of that VMWE. Take for example, the meaning of similar top expressions using word embedding models trained on relatively more minor corpora such as Parseme and PerDT is far different from the semantic meaning of the verb. Besides, most of the VMWEs in Persian are considered Light verb constructions (LVCs), which consist of a semantically reduced verb and a NVE. Also, a limited set of light verbs, around 20 Persian full verbs (Family, 2006), can be combined with an NVE to form a VMWE. Most of the top most similar expressions obtained using fasttext generated embedding vectors have a similar verbal

| VMWE | syn_verb | sim_syn_vmwe | sim_syn_combined |
|---|---|---|---|
| در_نظر_گرفتند (in consider got => considered) | شمردن (considering) | 0.81 | 0.62 |
| خشمگین_شده ( angry become => get angry) | برافروختن (getting angry) | 0.88 | 0.63 |
| بیان_می_کرد expression was doing => was ) (expressing | فرمودن (saying) | 0.83 | 0.50 |

Table 4:The degree of similarity with a synonymous simple verb

element with different NVE due to the character-level attitude of fasttext embedding models. Therefore, the semantics of the VMWE is not well-captured by fasttext. This being the case, for analyzing the compositionality of VMWE, only the word2vec model trained on Hmblog, which is the largest corpus, is considered. To assess the compositional nature of a verb in the dataset, the median value of each proposed criterion is calculated for the five most frequently occurring inflections of the verb. This median value is then used to determine the degree of compositionality of the infinitive verb, as measured by the given metric. Table 3 presents our experiment results for Direct_pre, Direct_post, DFsum, and DFcomp using the optimal threshold. The most accurate threshold was determined for each criterion within the calculated range of values. Direct_pre and DFsum achieved the highest accuracy of 70.9% among the proposed metrics, distinguishing between compositional and non-compositional verbs. A Direct_pre criterion value or DFsum above 0.23 indicates a compositional verb, while a value below indicates a non-compositional verb. Although Direct_post is also accurate, DFcomp had the lowest accuracy and did not effectively separate the two categories.

### 5.3 Analysis of Proposed Criteria

Further analysis Syn_sim reveals that out of 75152 non repetitive VMWE in the corpus, synonymous simple verbs for 4384 VMWE have been extracted; among them, for 3558 VMWE, the similarity of the synonymous simple verb to the VMWE is greater than the similarity of the synonymous simple verb to the combined vector (Table 4). Therefore, in 81% of VMWEs, the VMWE embedding vector constructed by the proposed method provides a better representation than the combined vector. Table 5 shows Direct_pre results for various VMWEs, where the values are highly similar to those of the DFsum metric. Non-compositional verbs in column one typically have a lower calculated criterion than compositional verbs in column five. However, some non-compositional verbs such as "چشم_زدن" (eye hitting => jinxing) have unexpectedly high calculated values due to their low occurrence frequency. This shows that higher occurrence frequency is likely to result in a more accurate calculated value, and should be taken into consideration when predicting compositionality. Moreover, DFcomp overestimates non-compositional verbs compared to compositional ones, and DFsing is unsuitable as the most similar expressions are often compound verbs.

| non-compositional | Direct_pre | DFcomp | freq | compositional | DFcomp | Direct_pre | freq |
|---|---|---|---|---|---|---|---|
| چشم_زدن (eye hitting => jinxing) | **0.23** | 0.22 | 7 | نگاه_کنید (look do => look) | 0.30 | 0.37 | 296 |
| فریب_خورده (deception ate => deceived) | **0.25** | 0.40 | 28 | تغییر_کند (change do => change) | 0.33 | 0.43 | 130 |
| دوست_دارم (friend have => to like) | **0.10** | 0.56 | 1032 | خاک_کرد (soil did => buried) | 0.16 | 0.23 | 3 |
| شکست_خورده (failure ate => failed) | 0.17 | 0.51 | 132 | فکر_کنید (think do => think) | 0.24 | 0.40 | 258 |
| زمین_خوردن (land eating => falling down) | 0.13 | 0.29 | 50 | قرار_دادن (put have => putting up) | 0.32 | 0.38 | 1806 |
| چانه_زدن (chin hiting => to bargaining) | 0.14 | 0.4 | 62 | آمده_به_دنیا (to world came => born) | 0.25 | 0.51 | 105 |

Table 5: Samples of Direct_pre and DFcomp results

## 6 Conclusion

To conclude, this paper presented an approach to predicting the compositional nature of VMWEs in Persian. The proposed method utilized automatic identification of VMWEs, followed by the creation of word embeddings that better capture the semantic properties of these expressions, and multiple criteria to determine their degree of compositionality. The study compared two neural architectures, BiLSTM and ParsBERT, and found that a fine-tuned BERT model outperformed the BiLSTM model with an F1 score of 89%. Moreover, the paper demonstrated the effectiveness of a word2vec embedding model in capturing the semantics of identified VMWEs and used criteria, resulting in an accuracy of 70.9% on a collected dataset of expert-annotated compositional and non-compositional VMWEs. These findings have important implications for further research in predicting the compositional nature of multiword expressions.

## Limitations

The limitations of our approach are mainly attributed to the limited annotated dataset of compositional and non-compositional VMWEs used in our experiments, which may not be representative of the full population of VMWEs in the Persian language. Moreover, the high prevalence of VMWEs in Persian and the varying perspectives among linguists on their compositional status add to the limitations of our results. Furthermore, the reliance on word embeddings for our approach may lead to potential inaccuracies in capturing the semantic information of words, especially for Persian which is a low-resource language. The limited data available for training word embeddings may not accurately reflect the language usage, resulting in a higher risk of inaccuracies for common words in the language that may not appear frequently in the training corpus. Moreover, as a further research we should evaluate the rule-based method against neural network-based models thoroughly, which requires more expert- annotated dataset. In addition, for future research endeavors, it is imperative to conduct a comprehensive evaluation of rule-based approaches in comparison to neural network-based models. However, such an evaluation would necessitate a more substantial dataset annotated by domain experts. Given these limitations, the results should be interpreted with caution, and further research is needed to fully understand the complexities of VMWEs in the Persian language.

## References


Mahmood Bijankhan. 2004. The role of the corpus in writing a grammar: An introduction to a software. *Iranian Journal of Linguistics*, 19(2):48–67.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

A Chaghari and Mehrnoush Shamsfard. 2013. Identification of verbs in Persian language sentences. *Journal of Computer Science and Engineering*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

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

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Neiloufar Family. 2006. *Explorations of semantic space: The case of light verb constructions in Persian*. PhD Thesis, Paris, EHESS.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53:3831–3847.

Waseem Gharbieh, Virendrakumar Bhavsar, and Paul Cook. 2017. Deep learning models for multiword expression identification. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 54–64.

Simin Karimi. 1997. Persian complex verbs: Idiomatic or compositional. *LEXICOLOGY-BERLIN-*, 3:273–318.

Hamzeh Motahari Khansari and Mehrnoush Shamsfard. 2021. HmBlogs: A big general Persian corpus. *arXiv preprint arXiv:2111.02362*.

Natalia Loukachevitch and Ekaterina Parkhomenko. 2018. Recognition of multiword expressions using word embeddings. In *Russian Conference on Artificial Intelligence*, pages 112–124. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Navnita Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. A comparative study of embedding models in predicting the compositionality of multiword expressions. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 71–76.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R. Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. *arXiv preprint arXiv:1707.08172*.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, and Voula Giouli. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics.

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.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le Ha. 2020. Verbal multiword expressions for identification of metaphor. In ACL.

P. O. Rossyaykin and N. V. Loukachevitch. 2019. Measure clustering approach to MWE extraction. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 562–575.

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

Bahar Salehi, Narjes Askarian, and Afsaneh Fazly. 2012. Automatic identification of Persian light verb constructions. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 201–210. Springer.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 266–275.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983.

Pourya Saljoughi Badlou. 2016. *Recognizing MultiWord Expressions in Persian*. Ph.D. thesis, Shahid Beheshti University.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, and Ivelina Stoyanova. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559.

Mehrnoush Shamsfard. 2007. Developing FarsNet: A lexical ontology for Persian. *GWC 2008*:413.

Babak Sharif. 2017. Persian Compound Verb Formation from a Cognitive Grammar Viewpoint. *Language Related Research*, 8(2):149–170.

Shiva Taslimipoor and Omid Rohanian. 2018. Shoma at parseme shared task on automatic identification of vmwes: Neural multiword expression tagging with high generalisation. *arXiv preprint arXiv:1809.03056*.