E - ISSN: 2953-8203
P - ISSN: 2953-819X

YEREVAN STATE
UNIVERSITY

JOURNAL OF

IRANIAN LINGUISTICS

Volume 2 - Issue 2 - 2025

JOURNAL OF IRANIAN LINGUISTICS

Volume 2 | issue 2

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

E - ISSN: 2953-8203
P - ISSN: 2953-819X

JOURNAL OF IRANIAN LINGUISTICS
VOLUME 2 | ISSUE 2

CONTENT

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

JIL 2/2 (2025): 87-107

Journal of Iranian Linguistics

The Nonverbal Element in Persian Verbal Multiword Expressions: A Corpus Annotation Approach

Vahide Tajalli*
Shahid Beheshti University
Mehrnoush Shamsfard
Shahid Beheshti University
Yalda Yarandi
Shahid Beheshti University
Mahtab Sarlak
Shahid Beheshti University
Arezoo Haghbin
Shahid Beheshti University

doi.org/10.46991/jil/2025.02.04

Abstract: This article presents a linguistic framework for the identification and annotation of Persian (Farsi) Verbal Multiword Expressions (VMWEs), developed in alignment with the standards and methodologies set by the PARSEME Corpus—an international research network focused on the systematic analysis of multiword expressions across languages. The study aims to bridge the gap between universal annotation guidelines and language-specific grammatical features by tailoring the PARSEME framework to the structural and semantic properties of Persian. By extracting the characteristics of Persian VMWEs, particularly their nonverbal elements (preverbs) and their diverse syntactic and morphological patterns, this work contributes to a more refined understanding of Persian verbal idiomaticity and the advancement of natural language processing tasks. The article details the development of annotation guidelines that reflect both cross-linguistic categories and Persian-specific grammatical phenomena and the process of annotating a corpus of 5,617 sentences encompassing a wide range of Persian VMWEs including light verb constructions, verbal idioms, and prefix verbs. The practical applications of these guidelines in natural language processing are discussed, highlighting their potential to enhance machine understanding of complex verbal constructions, improve syntactic parsing accuracy, and support downstream tasks such as machine translation, information extraction, and semantic role labeling.

Keywords: Compound Verb; Nonverbal Element; Persian; Preverb; Text Corpus; Verbal Multiword Expression

Conflict of Interest
The authors declare no conflicts of interest.

Vahide Tajalli
E-mail: vtajalli@ut.ac.ir
ORCID: https://orcid.org/0000-0003-3118-2903
Mehrnoush Shamsfard
E-mail: m-shams@sbu.ac.ir
ORCID: https://orcid.org/0000-0002-7027-7529
Yalda Yarandi
E-mail: yalda.yarandi@gmail.com
ORCID: https://orcid.org/0009-0008-3088-8166
Mahtab Sarlak
E-mail: sarlak3@gmail.com
ORCID: https://orcid.org/0009-0002-0166-096X
Arezoo Haghbin
E-mail: haghbin33@gmail.com
ORCID: https://orcid.org/0009-0000-5285-1702

Received: 07.11.2025
Revised: 09.12.2025
Accepted: 25.12.2025

1. Introduction

Verbal Multiword Expressions (VMWEs), also known as complex predicates, complex verbs, or compound verbs, play a crucial role in natural language understanding, as they often convey meanings that go beyond the sum of their parts.

Persian is an Indo-European language with relatively few simple verbs and a large number of VMWEs. New verbs are typically formed by combining nouns or adjectives with a light verb (Karimi-Doostan, 2005). These constructions are referred to as "compound verbs" in Persian grammatical tradition. This language contains around 200 simple verbs (Mohammad and Karimi, 1992; Samvelian and Faghiri, 2013), most of which can function simultaneously as light and heavy verbs. Compound verbs have gradually replaced simple verbs since the thirteenth century. In some cases, the use of the simple verb is limited to written or elevated registers (Folli et al, 2005).
As mentioned, a Persian compound verb is formed by a light verb (LV) and a nonverbal element (NV). The NV precedes the LV (1-a) which has partly or entirely lost its original meaning. Different inflections including negation and future are normally prefixed or suffixed to the LV (1-b).

(1) a. *mæn* *dust-æm* *ra*[1] ***dæ'væt*** ***kærd-æm***.[2]
I friend-my ACC invite do.PST-1SG
'I invited my friend.'

b. *mæn* *dust-æm* *ra* ***dæ'væt*** ***næ-kærd-æm***.
I friend-my ACC invite NEG-do.PST-1SG
'I did not invite my friend.'

Identifying VMWEs in Persian poses several challenges, including a lack of recourses, morphological complexity, lexical variation, idiomatic expressions, and semantic ambiguity. Addressing these challenges requires linguistic resources along with innovative approaches that take into account the language's unique characteristics.

---

[1] - *ra* is Persian object marker.

[2] Persian examples are transliterated using a modified Latin-based system in which long vowels are represented as a, i, u; short vowels as æ, e, o; and the glottal stop is marked with an apostrophe (ʾ). In addittion, in the glossed examples, grammatical abbreviations are employed as follows: ACC (accusative), PRES (present tense), PST (past tense), SG (singular), PL (plural), NEG (negation), and PROG (progressive aspect).

In this article, we aim to present a framework for identifying Persian VMWEs in accordance with the guidelines of PARSEME Corpus. PARSEME (Parsing and Multiword Expressions) is a research network dedicated to studying multiword expressions. It aims to develop methods for their identification, analysis, and representation to enhance natural language processing tasks. PARSEME annotation largely relies on Universal Dependencies for the annotation of morpho-syntax, due to the shared objectives of universality. The annotation flow follows a decision diagram driven by linguistic tests. An expression is annotated as a VMWE if it can pass the tests on binary decision trees. By understanding VMWEs, which include idioms, collocations and phrasal verbs, PARSEME contributes to improving machine comprehension of language constructs. This corpus, in its version 1.3, contained the VMWEs of 26 different languages including Persian (Savary et al., 2023). The Persian part included 3617 sentences with VMWE tags which required revision based on both PARSEME guideline and Persian grammar, therefore, a framework was needed to enhance the accuracy of the previous annotations and add new sentences, thereby increasing the dataset size. In order to formulate the framework and annotate the Persian data, we studied the linguistic details of the Persian VMWEs, particularly the characteristics of the non-verbal elements, and identified the various forms they can take.

A Persian VMVE corpus, designed for a low-resource language, has significant implications for computational linguistics, natural language processing, and linguistic research. It provides valuable insights into Persian syntax, semantics, and pragmatics. As a foundational dataset, it assists in training algorithms for machine translation, enhancing their ability to handle idiomatic verbs and expressions to produce more accurate and natural translations. Moreover, it supports the development of tools for sentiment analysis, information extraction, and text summarization.

The present article is organized as follows: Section 2 provides an overview of previous research on processing Persian VWMEs. Section 3 outlines the steps involved in corpus development and explains the annotation process. section 4 introduces the annotation framework. Section 5 presents the findings and section 6 concludes the article.

2. Related Work

Studies on VMWEs comprise a significant part of Persian linguistic literature, among which some research focuses on the processing and detection of these expressions in Persian texts.

Iranpour Mobarakeh and Minaei (2009) focused on identifying verbs in Persian textual corpora. They mention that compound verbs consist of a light verb, which is morphologically similar to other verbs but differs from them

in terms of meaning, and a nonverbal element, which has no morphological suffixes—unlike the light verb. Hence, identifying the verbal part of compound verbs is relatively easy; however, the NV that precedes the verbal part complicates the identification of compound verbs. In the first step, the model developed in their research uses the structural information of Persian verbs to identify them. According to the derivational structure of Persian, its words can be stemmed systematically. The presented stem finder applies rules to extract verb stems and uses a vocabulary to improve the accuracy of the results. In the next step, using the n-gram approach, the disambiguation of co-occurring characters is discussed to enhance efficiency.

Rasooli et al. (2011) employ machine learning algorithms to identify Persian compound verbs. They highlight the diverse lexical semantics as the main challenging feature of the light verb constructions. They believe one of the difficulties of research in this field is the lack of reliable data sets for supervised and unsupervised learning. In their study, two unsupervised learning methods are applied to automatically identify Persian compound verbs. One is based on a modified version of the PMI concept, a method that calculates the probability of two words occurring together in the same corpus, known as PPMI**;** the other is based on the k-means clustering algorithm. They show that criteria such as PMI are insufficient for identifying Persian compound verbs, due to sparse data and the flexibility of word spacing in the Persian verbal construction. Using the k-means algorithm, they show that the average number of nouns between the NV and the LV significantly affects the identification efficiency of the compound verbs.

Mansoory et al. (2011) discuss the development of the Persian WordNet for verbs within FarsNet (Shamsfard, 2007), a semi-automated framework for building the Persian WordNet. They describe different types of Persian compound verbs, as well as the syntactic and semantic properties of each type. They then address the specific characteristics and behaviors of these verb types in order to develop a semantic lexicon. Finally, they present a method of using such linguistic properties in the automatic extraction of compound verbs and their relations from large text corpora and dictionaries with the aim of enriching the Persian WordNet of verbs. They treat compound verbs as lexical phenomena rather than syntactic ones; therefore, they can belong to synonym sets that include both compound and simple verbs.

Samvelian and Faghiri (2013) introduce PersPred as a manually annotated syntactic and semantic database for Persian compound verbs. They propose a framework for storing and describing Persian compound verbs. PersPred 1 contains more than 700 combinations of the verb *zædæn* (hit) with a noun, organized in a spreadsheet. Adopting a Construction-based approach, they study the way the productivity of Persian compound verbs can be explained despite their idiomaticity and the link generally established between compositionality and productivity.

In their study, Sarlak et al. (2023) propose both non-contextual and contextual methods to identify VMWEs. For the non-contextual strategy, they utilize a VMWE dataset based on Persian WordNet, created by collecting all compound verbs from FarsNet. They extract 21,462 VMWEs from FarsNet. To identify VMWEs in a sentence, they extract n-grams (n=2,3,4) and search for the presence of all components of a multi-word verb within the n-gram. However, not all cases found are VMWEs, particularly when intermediate words are present. For the contextual approach, their method addresses a sequence labeling problem aimed at recognizing the components thet comprise VMWEs. Through comparative analysis, the study evaluates two neural architectures, BiLSTM and ParsBERT (Farahani et al., 2021), for VMWE identification. Results indicate that a fine-tuned BERT model outperforms the BiLSTM model.

Eshaghi and Karimi-Doostan (2021) have developed a synchronic, monolingual corpus of 6000 Persian LVCs. They employ VBA macro codes to extract the LVCs consisting of 21 Persian LVs: *dashtæn*: have, *kærdæn*: do, *shodæn*: become, *gæshtæn*: turn, *gozashtæn*: put, *keshidæn*: pull, *didæn*: see, *dadæn*: give, *bakhshidæn*: give, grant, *gereftæn*: get, *yaftæn*: obtain, *amædæn*: come, *aværdæn*: bring, *residæn*: arrive, *ræftæn*: go, *oftadæn*: fall, *ændakhtæn*: throw, *bordæn*: take, *khordæn*: collide, *zædæn*: hit, and *bæstæn*: tie.

In the present paper, following a universal framework, we aim to investigate Persian VMWEs in a wider area comparing to the previous studies and focus on the nature of the non-verbal element in these expressions.

3. Corpus Development

The primary objective of this work was the *quantitative and qualitative expansion* of the PARSEME corpus. For the quantitative expansion, we *selected 2,000 sentences from PerUDT* (Safari et al., 2022) to add to the PARSEME Persian dataset, *based on their notable similarity in Universal Dependencies Treebank (UDT) structure and Part-of-Speech (POS) labels to the PARSEME corpus*. This deliberate expansion not only increased the corpus's volume but also enhanced its overall quality, enabling more comprehensive analysis and research.

The approach adopted in this study was grounded in Sarlak et al.'s (2023) model, specifically designed for the identification of Persian VMWEs within sentences. Utilizing their model, we accurately annotated PerUDT sentences to identify VMWEs. To ensure corpus balance, we included 1,500 sentences containing VMWEs and 500 sentences without any. Following the initial annotation process, we conducted a thorough review of each tag, making necessary corrections and categorizing each VMWE according to a defined set of guidelines. This included refining the initial PARSEME tags to

accurately designate the type of each VMWE, thereby ensuring the corpus's integrity for further linguistic research and applications.

The guidelines provided in this paper were instrumental, offering detailed descriptions and examples of VMWE types, which fostered a shared understanding among annotators and ensured consistency throughout the annotation process. To annotate the data, we employed FLAT2, a web-based annotation tool based on FoLiA format, which supports a wide range of linguistic annotations. This tool enabled us to efficiently revise and enrich the corpus with precise labels, ensuring the quality and accuracy of our linguistic data.

4. Guidelines

In this section, we provide a brief overview of the PARSEME guidelines, which form the basis for our categorization of Persian VMWEs.

PARSEME Categories of VMWEs:

1. **Universal categories**, valid across all languages in the corpus:
   a. Light verb construction (**LVC**)
      - The LV is semantically bleached (**LVC.full**) [to make a decision]
      - The LV adds a causative meaning (**LVC.cause**) [to cause a problem]
   b. Verbal Idioms (**VID**)
      - Verbal phrases with more than two components [to make ends meet]
      - Inflexible Verbal phrases with an adjective or an adverb as the NV [to come clean]
      - Verbal phrases with a non-eventive concrete noun as the NV [to give a hand]
      - Verbal phrases with a cranberry word as the NV [go astray]
      - Proverbs and conventionalized phrases [I beg your pardon]
2. **Quasi-universal categories**, valid for some languages in the corpus:
   a. inherently reflexive verbs (**IRV**):
      - Reflexives verbs in languages like French [s'évanouir]
      - Verbs including a reflexive pronoun [to help oneself]
   b. verb-particle constructions (**VPC**) with two subcategories:
      - the particle totally alters the verb's meaning (**VPC.full**) [to do in]
      - the particle adds a partly predictable meaning (**VPC.semi**) [to end up]

c. multi-verb constructions (**MVC**) [to let go]

Persian was one of the languages included in the PARSEME corpus, and based on its guidelines and Persian grammar, we aimed to categorize Persian VMWEs. As mentioned in Section 1, our primary dataset consisted of *3,617 sentences from PARSEME 1.2* and *2,000 sentences from PerUDT,* annotated according to separate instructions. Table 1 *presents the distribution of different VMWE types* found in this primary dataset.

Table 1: The number of different types of VMWEs in the primary dataset

	PARSEME	**PerUDT**	**Total**
Sentences	3617	2000	5617
VMWE	3450	1933	5383
LVC.full	3432	1933	5365
VID	17		17
IRV	1		1

As can be observed, the PerUDT dataset contained only LVCs, whereas PARSEME 1.2 included LVCs along with a small number of other VMWE categories. The following section outlines the challenges we encountered and the decisions we made in aligning the structural and semantic characteristics of Persian VMWEs with the categorization framework provided by the PARSEME guidelines.

4.1. Idiomaticity

Unpredictability or non-compositionality of the meaning is the most common criterion for defining idioms (Karimi, 1997). This concept, when applied to verbal expressions, exists on a continuum ranging from fully idiomatic to partially compositional verbs. In Persian, the lexical meaning of VMWEs is often not entirely predictable from the meaning of their individual component (Samvelian and Faghiri, 2013). Both the LV and the NV contribute distinct semantic roles**.** The pairing of a specific verb with a specific noun is frequently idiosyncratic, as there is often no clear semantic rationale for the selection of a particular verb (Samvelian and Faghiri, 2014).

(2) a. ***gush*** ***kærdæn*** (to listen)
ear doing
b. ***cheshm*** ***kærdæn*** (to hurt someone with an evil eye)
eye doing

As outlined in the PARSEME framework, the selection of NV plays a more decisive role in determining whether a construction qualifies as a VMWE.

Since semantic non-compositionality is difficult to assess directly, it is typically approximated through lexical and morpho-syntactic inflexibility. Accordingly, in example 3-a, the construction is classified as an LVC because the NV is an abstract noun. In contrast, example 3-b is categorized as a verbal idiom (VID) due to the presence of an adjective as the NV**.**

(3) a. ***hæds*** ***zædæn*** (to guess)
guess hitting

b. ***kutah*** ***amædæn*** (to give up)
short coming

Inflexibility refers to the inability to substitute the NV with similar lexical items while preserving the VMWE's meaning. For example, in 3-b, *bolænd amædæn* (lit. 'long coming')—as the opposite of *kutah amædæn* (lit. 'short coming')—does not yield a meaningful expression in Persian. This illustrates that the NV in *kutah amædæn* is not freely interchangeable, highlighting the idiomatic and fixed nature of the construction.

4.2. The Category of the NV

In Persian VMWEs, the NV is most often an abstract noun, as noted in the PARSEME guidelines for LVCs.

(4) ***pærtab*** ***kærdæn*** (to throw)
throw doing

Sometimes a preposition establishes the connection between the noun and the verb:

(5) ***be*** ***færamushi*** ***sepordæn*** (to forget)
to forgetfulness delivering

These verbal constructions are categorized as LVCs. However, other types of NVs are also used in Persian VMWEs. Constructions featuring these alternative NVs are typically labeled as verbal idioms in the PARSEME framework. In the following subsections, we examine the use of adjectives, adverbs, and concrete nouns as NVs in Persian VMWEs to reveal their structural and semantic properties.

4.2.1. Adjectives as the NV

Persian adjectives can be classified into two main types: simple adjectives and predicative adjectives.

4.2.1.1. Simple Adjectives

According to Persian grammar, adjectives can serve as the non-verbal element in Light Verb Constructions. For example:

(6) a. ***tæmiz*** ***kærdæn*** (to clean)
clean making
b. ***khamush*** ***kærdæn*** (to turn off)
off making

These types of verbal constructions are not classified as VMWEs in the PARSEME framework. Due to their flexible structure—where the NV can be substituted with similar adjectives while preserving the overall meaning—they do not meet the criteria for verbal idioms. Such constructions resemble expressions in English like *make somebody happy* **or** *make something clean***,** which denote a change of state rather than idiomatic meaning. Accordingly, we did not treat them as compound verbs in the Persian corpus either.

However, certain two-part verbal constructions consisting of an adjective and a verb exhibit inherently idiomatic meanings. In line with the PARSEME guidelines, these constructions are considered inflexible and are therefore categorized as VIDs. For example, in (7-a), *kæm aværdæn* (lit. 'little bringing') conveys an idiomatic meaning in Persian. Its hypothetical opposite, *ziyad aværdæn* (lit. 'much bringing'), does not form a meaningful expression, underscoring the inflexibility of the original construction. This lack of substitutability confirms its classification as a VID.

(7) a. ***kæm*** ***aværdæn*** (to give in)
little bringing

b. ***deraz*** ***keshidæn*** (to lie down)
long stretching

4.2.1.2. Predicative Adjectives

A subset of Persian adjectives, known as predicative adjectives, differs from standard adjectives in several key ways. These adjectives appear exclusively within the structure of VMWEs and are not used independently. They cannot be employed in comparative or superlative forms, do not function as noun modifiers, and cannot be intensified by degree modifiers (Karimi-Doostan, 2011). Below are two illustrative examples:

(8) a. ***færamush*** ***kærdæn*** (to forget)
forgotten making

b. ***mæhsub*** ***kærdæn*** (to consider)
considered making

The constructions **khaterat-e færamush* (lit. 'forgotten memories**'**) or **khaterat-e færamush-tær* (lit. 'more forgotten memories**'**) are ungrammatical in Persian.

Although these adjectives are not cranberry words in the strictest sense, given their morphological independence and clear semantic content, they are exclusively used within compound verb constructions. Therefore, we have classified them as cranberry words for the purposes of this study, and the corresponding verbal multiword expressions were counted as VIDs.

4.2.2. Concrete Nouns as the NV

Based on the guidelines, a non-eventive concrete noun cannot function as the NV in an LVC. However, if the concrete noun is eventive, i.e. if it denotes an event, it may be incorporated into an LVC. In 9-b the concrete noun ***jaru*** (broom) serves as the core of the construction by introducing the event as a whole.

(9) a. *maman* *otagh* *ra* ***jaru*** ***kærd-Ø.***
mom room ACC broom do.PST-3SG
'Mom swept the room.'

b. ***Jaru****-ye* *otaq* *do* *sa'æt* *tul* *keshid-Ø.*
broom-of room two hours length draw.PST-3SG
'It took two hours to sweep the room.'

In Persian, a light verb (LV) is frequently combined with the names of tools and objects to express the action performed using that tool.

(10) a. ***mesvak*** ***zædæn*** (to brush the teeth)
toothbrush hitting

b. ***telefon*** ***kærdæn*** (to make a phone call)
telephone doing

This is one of the productive mechanisms in Persian for forming multiword verbal expressions through the use of eventive concrete nouns. Consequently, such constructions are classified as light verb constructions.

Concrete nouns can be also used in the structure of VIDs. Our data indicated that 62% of concrete noun NVs in these constructions were body parts including eye, head, hand, foot, etc.

(11) a. ***æz*** ***pa*** ***dæramædæn*** (to be exhausted, to be killed)
of foot getting out

b. ***dæst*** ***shostæan*** (to be disappointed)
hand washing

In the same way as the adjectives, flexible cases with the sense of "making a change" including the ones in example 12 were excluded from the VMWE group, even though they are considered compound verbs in the traditional Persian grammar.

(12) a. ***sæng*** ***kærdæn*** (to change to a stone)
stone making

b. ***ard*** ***kærdæn*** (to change to flour)
flour making

4.2.3. Adverbs as the NV

In a small portion of the data, adverbs appeared as the non-verbal component of verbal multiword expressions. These constructions were also annotated as VIDs.

(13) a. ***pish*** ***amædæn*** (to happen)
forward coming

b. ***pæs*** ***oftadæn*** (to faint)
back falling

According to the PARSEME guidelines, verbal structures containing adverbs with directional meaning were not annotated as verbal multiword expressions.

(14) a. ***æghæb*** ***ræftæn*** (to go back)
back going

b. ***birun*** ***keshidæn*** (to pull out)
out pulling

4.3. Overlapping structures

In overlapping structures, a single LV or NV may occur in more than one verbal multiword expression. This may happen in the following cases.

4.3.1. Coordinate structures

There are instances in which an LV is implied rather than explicitly stated. In such cases, a single LV simultaneously serves multiple non-verbal elements, as illustrated in example 15.

(15) a. *Hæme-ye* *khane* *ra*
all-of house ACC
gærdgiri ***væ*** ***jaru*** ***kærd-æm.***
dusting and broom do.PST-1SG
'I dusted and swept the whole house.'

b. *dær* *anja* ***ærj*** ***væ*** ***mænzelæt*** ***dasht-Ø.***
in there esteem and dignity have.PST-3SG
'S/he had esteem and dignity there.'

4.3.2. An LV as a VMWE

There are Persian verbal multiword expressions in which the LV itself is a VMWE (Moloodi & Kouhestani, 2017). In such cases, the entire construction consists of three components, forming a layered multiword expression.

(16) $VMWE_1$ = NV + $VMWE_2$
kahesh ***[peyda*** ***kærdæn]*** (to decrease)
decrease found making

The tendency to form compound verbs in Persian has led to the coexistence of two sets of verbs, simple and complex, for a range of verbal concepts. This phenomenon is comparable to English, where both "decide" and "make a decision" convey the same meaning. For instance, the simple verb ***yaftæn*** (lit. 'to find**'**) functions as an LV in many verbal multiword expressions. Over time, however, it has been replaced by the compound verb ***peyda kærdæn*** (lit. 'found making**'**), which itself constitutes a VMWE.

yaftæn = **peyda kærdæn** (to find)

$[$kahesh yaftæn$]_{LVC}$ = [kahesh $[$peyda kærdæn$]_{VID}$ $]_{LVC}$ (to decrease)
decrease finding decrease found making

Peyda (found) functions as a predicative adjective. Therefore, in the construction ***kahesh peyda kærdæn*** (to experience a decrease), the expression contains a VID embedded within an LVC.

4.4. Agreement on the NV

As noted in Section 1, inflectional markers in Persian LVCs are typically prefixed or suffixed to the LV. However, there exists a unique type of verbal multiword expression in which an enclitic pronoun is attached to the non-verbal element. In these constructions, the subject agrees with the clitic in both person and number, while the verb consistently appears in the third person singular form (Rasekh-Mahand, 2014). The clitics involved are object clitics, distinct from the subject clitics found in standard simple or compound verbs. In this respect, the structure resembles reflexive verbs: it features an experiencer subject, lacks imperative forms, and cannot occur without the clitic. Moreover, these expressions rarely possess a usable infinitive form, or the infinitive fails to retain the same meaning. More precisely, they lack a prototypical form that would allow for straightforward classification as VMWEs.

Exceptional compound verb = NV + Object clitic + LV (3SG)

(17) a. *bæche* ***khab****-æsh* ***bord****-Ø*.
kid sleep-him take.PST-3SG
'The kid fell asleep.'

b. *mæn æz bagh* ***khosh****-æm* ***amæd****-Ø*.
I of garden good-me come.PST-3SG
'I liked the garden.'

c. *mæn* ***særd****-æm* ***æst***.
I cold-me be.PRES.3SG
'I am cold.'

d. *mærd* ***bavær****-æsh* ***shod****-Ø*.
man belief-him become.PST-3SG
'The man believed it.'

We decided to treat these constructions as standard VMWEs within the corpus. By omitting the clitic, we have an unusable combination of NV+LV as illustrated in the following examples, and we categorized them accordingly.

(18) a. ***khosh*** ***amædæn*** (to like) [VID]
good coming

b. ***bavær*** ***shodæn*** (to believe) [LVC]
belief becoming

4.5. Prefix verbs

In Persian, there are verbs that include prefixes. Their structure resembles that of particle verbs, except that the prefix is directly attached to the verb.

(19) a. ***bær-dashtæn*** (to pick up)
on-having

b. ***dær-gozæshtæn*** (to pass away)
in-passing

On the other hand, their behavior is more like LVCs in that other affixes, auxiliaries and clitics can appear between the prefix and the verb.

(20) a. ***bær****-mi-****dasht****-Ø*.
on-PROG-have.PST-3SG
'S/He was picking'

b. ***bær****-næ-****dasht****-Ø*.
on-NEG-have.PST-3SG
'S/He did not pick up'

These verbs are often regarded as a third category, distinct from both simple and compound verbs. In this corpus, they are classified as a subgroup of VMWEs, specifically termed verb-particle constructions (VPCs). This subgroup is further divided into two types:

1. VPC.full: the prefix totally changes the meaning of the verb.

(21) ***bær-dashtæn*** (to pick up)
on – having

2. VPC.semi: the prefix either does not add any meaning or adds a partly predictable meaning to the verb.

(22) ***bær-æfrukhtæn*** (to ignite)
on - igniting

4.6. Serial verb constructions

There are a few serial verb constructions in informal Persian, including:

(23) a. ***bezæn*** ***ber-im****!*
hit go-1PL
'Let's go!'

b. ***begir-æm*** ***bekhab-æm****.*

take-1SG sleep-1SG
'I'm gonna get some sleep.'

In these constructions, one verb functions as the main verb and carries the core semantic content of the combination (Anosheh, 2019). We chose to classify them as Multi-Verb Constructions (MVCs), However, no instances of serial verb constructions were found in our data, as this type of VMWE typically appears in informal contexts, whereas our corpus consisted of formal Persian sentences.

4.7. Reflexive verbs

Persian does not have reflexive clitics like those found in French. However, the pronoun 'oneself' can appear within verbal phrase structures. These instances were classified as Reflexive-Inflected Verbs (RIVs) under the VMWE framework.

(24) a. ***be*** ***khod*** ***amædæn*** (to come to one's senses)
to oneself coming

b. ***khod*** ***ra*** ***gereftæn*** (to be full of oneself)
oneself ACC taking

4.8. Passive voice

There are two mechanisms for forming passive constructions out of Persian transitive active VNWEs.

1. If the LV is ***kærdæn*** (to do, to make) or one of a few similar verbs, it is replaced by ***shodæn*** (to become) in the passive form and no auxiliary verb is used.

(25) active: ***dæ'væt*** ***kærdæn*** (to invite)
invite doing

passive: ***dæ'væt*** ***shodæn*** (to be invited)
invite becoming

In this group, the passive form was labeled as a VNWE if its corresponding active form met the criteria for a verbal multi-word expression, as illustrated in example (25).

2. In other cases, the verb ***shodæn*** is added to the verbal construction as a passive auxiliary and the LV changes to a past participle.

(26) active: ***ejare*** ***dadæn*** (to rent out)
rent giving

passive: ***ejare*** ***dad-e*** ***shodæn*** (to be rented out)
rent given becoming

In this group, only the NV and the past participle were tagged as a VMWE. Auxiliary verbs, like in other instances, were left unannotated.

Finally, we arrived at a classification scheme for Persian. In total, 5,617 VMWEs were tagged in our dataset. Of these, 83% were identified as LVCs, while the remaining 17% fell into other categories. The classification of verbal expressions, along with the percentage distribution of each type, is presented below.

1. **Light Verb Constructions** (LVC)
    - Verbal phrases with abstract nouns as the NV (98.4%)
    - Verbal phrases with eventive concrete nouns as the NV (1.6%)

2. **Verbal Idioms** (VID)
    - Verbal phrases with inflexible constructions and simple adjectives as the NV (13.18%)
    - Verbal phrases with inflexible constructions and adverbs as the NV (3.6%)
    - Verbal phrases with inflexible constructions and non-eventive concrete nouns as the NV (34.83%)
    - Verbal phrases with predicative adjectives as the NV (36.10%)
    - Verbal phrases with more than one NV (10.11%)
    - Proverbs and conventionalized phrases (2.17%)

3. **Verb-particle constructions** (VPC)
    - Prefix verbs in which the prefix totally changes the meaning of the verb (71.84%)
    - Prefix verbs in which the prefix adds a partly predictable meaning to the verb (28.16%)

4. **Reflexive verbs** (IRV)
    - Verbal phrases containing reflexive pronouns

5. **Multi-Verb Constructions** (MVC)

- Serial verb constructions

The following table shows the number of each type of VMWE and their subcategories in the final data.

Table 2: The number of different types of VMWEs in the final data

Category	**Subtype / Construction**	**Count (#)**	**Total**
LVC	Abstract N as NV	**4366**	**4434**
	Eventive concrete N as NV	**68**	
VID	Non-eventive concrete N as NV	**193**	**554**
	Simple Adj as NV	**73**	
	Adv as NV	**20**	
	Predicative Adj as NV	**200**	
	More than one NV	**56**	
	Proverbs (prov)	**12**	
VPC	Full	**227**	**316**
	Semi	**89**	
IRV	—	**10**	**10**
MVC	—	**0**	**0**

5. Analysis and Findings

This article outlined the development of a set of annotation guidelines to apply to a corpus of 5,617 sentences, encompassing various types of Persian

verbal multiword expression. Below, we present a summary of the key observations encountered during the annotation process.

- Persian contains a large number of verbal multiword expressions, the majority of which are light verb constructions.
- The verb ***kærdæn*** (to do, to make) is the most frequently used and the most productive Persian light verb.
- Most non-verbal elements used in forming Persian verbal multiword expressions are abstract nouns.
- After light verb constructions, verbal idioms are the second most frequent type of compound verbs in Persian, and their non-verbal components are typically adjectives or concrete nouns.
- The annotated data showed that 62% of concrete nouns in idiomatic verbal constructions were body parts like eye, head, hand and foot.
- Persian verbal multiword expressions exhibit different types of overlapping structures. In addition to coordinate structures (see Example 16), the data included three types of embedded structures:

  a. The light verb is an LVC

  (27) ***tæht-e*** ***tæ'sir*** ***[qærar*** ***dadæn]***$_{\textbf{LVC}}$ (to impress)
  under influence setup giving

  b. The light verb is a VPC

  (28) ***færyad*** ***[bær-aværdæn]***$_{\textbf{VPC}}$ (to yell)
  yell on – bringing

  c. The light verb is a VID

  (29) ***ronæq*** ***[peyda*** ***kærdæn]***$_{\textbf{VID}}$ (to prosper)
  prosperity found making

- Some flexible constructions, although considered VMWEs based on Persian grammatical criteria, were not annotated as such under the PARSEME framework:

  a. Adjective + verb

  (30) a. ***momken*** ***budæn*** (to be possible)
  possible being

b. ***khamush*** ***kærdæn*** (to turn off)
off making

b. Concrete noun + verb

(31) a. ***be*** ***gush*** ***residæn*** (to be heard)
to ear reaching

b. ***eynæk*** ***zædæn*** (to wear glasses)
glass hitting

c. Direction + verb

(32) a. ***æghæb*** ***ræftæn*** (to go back)
back going

b. ***birun*** ***keshidæn*** (to pull out)
out pulling

- Prefix verbs, which are typically treated as simple verbs in most corpora including the primary data of this study, were annotated as verbal multiword expressions due to their behavioral similarity to particle verbs.

- Persian has two mechanisms for forming the passive voice from compound verbs. If the light verb is *kærdæn* (to do, to make) or similar, it is replaced by *shodæn* (to become). If the light verb is something else, *shodæn* is added as an auxiliary verb.

6. Conclusion

In this study, we investigated Persian verbal multiword expressions (VMWEs) across a broad range. To this end, we followed the PARSEME corpus guidelines and adapted them to align with the grammatical features of Persian. Drawing on both language-specific properties and universal categories present in Persian, we developed annotation instructions and applied them to a corpus of 5,617 sentences. Additionally, we examined the characteristics of nonverbal elements across various types of Persian VMWEs.

7. Limitations & Future Work

We have obtained this instruction by studying articles on Persian VMWEs, PARSEME guidelines and examining the data of two corpora. There may still be cases not covered and be presented in future studies.

Acknowledgment

This work received advisory support from the CA21167 COST action UniDive, funded by European Cooperation in Science and Technology (COST).

BIBLIOGRAPHY

Anousheh, M. (2019), "Serial Verb Construction in Persian: A Minimalist Approach", *Journal of Researches in Linguistics*, vol. 11, no. 1, pp. 73–91. In Persian.

Eshaghi, M., and Karimi-Doostan, G. (2021), "The Productivity of Persian Light Verbs", *Journal of Language Researches*, vol. 12, no. 2, pp. 1–28. In Persian.

Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2021), "ParsBERT: Transformer-Based Model for Persian Language Understanding", *Neural Processing Letters*, vol. 53, pp. 3831–3847.

Folli, R., Harley, H., and Karimi, S. (2005), "Determinants of Event Type in Persian Complex Predicates", *Lingua*, vol. 115, no. 10, pp. 1365–1401.

Iranpour Mobarakeh, M., and Minaei-Bidgoli, B. (2009), "Verb Detection in Persian Corpus", *International Journal of Digital Content Technology and its Applications*, vol. 3, no. 1, pp. 58–65.

Karimi, S. 1997. "Persian Complex Verbs: Idiomatic or Compositional", *Lexicology*, vol. 3, pp. 273–318.

Karimi-Doostan, G. (2005), "Light Verbs and Structural Case", *Lingua*, vol. 115, no. 12, pp. 1737–1756.

Karimi-Doostan, G. (2011), "Separability of Light Verb Constructions in Persian", *Studia Linguistica*, vol. 65, no. 1, pp. 70–95.

Mansoory, N., Shamsfard, M., and Rouhizadeh, M. (2012), "Compound Verbs in Persian WordNet", *International Journal of Lexicography*, vol. 25, no. 1, pp. 50–67.

Mohammad, J., and Karimi, S. (1992), "Light Verbs Are Taking Over: Complex Verbs in Persian", in: *Proceedings of the Western Conference on Linguistics (WECOL)*, vol 5, ed. J. Nevis, and V. Samiian, Fresno: California State University, pp. 195–212.

Moloodi, A., and Kouhestani, M. (2017), "The Role of Metaphor and Metonymy in the Semantics of Persian Adjectival Preverbs: A Cognitive Linguistics Approach", *Language Art*, vol 2, no. 2, pp. 91–105.

Rasekh, M. (2014), "Persian Clitics: Doubling and Agreement." *Journal of Modern Languages*, vol. 24, no. 1, pp. 16–33.

Rasooli, M. S., Faili, H., and Minaei-Bidgoli, B. (2011), "Unsupervised Identification of Persian Compound Verbs", in: *Advances in Artificial Intelligence: 10th Mexican International Conference on Artificial Intelligence,*

MICAI 2011, Puebla, Mexico, November 26 - December 4, 2011, Proceedings, Part 1, ed. I. Batyrshin, and G. Sidorov, Heidelberg: Springer, pp. 394–406.

Safari, P., Rasooli, M. S., Moloodi, A., and Nourian A. (2022), "The Persian Dependency Treebank Made Universal", in: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, ed. N. Calzolari et al., Marseille: European Language Resources Association (ELRA), pp. 7078–7087.

Samvelian, P., and Faghiri, P. (2013), "Introducing PersPred, a Syntactic and Semantic Database for Persian Complex Predicates", in: *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013),* ed. V. Kordoni, C. Ramisch, and A. Villavicencio, Stroudsburg: Association for Computational Linguistics, pp. 11–20.

Samvelian, P., and Faghiri, P. (2014), "Persian Complex Predicates: How Compositional Are They?", *Semantics-Syntax Interface*, vol 1, no. 1, pp. 43–74.

Sarlak, M., Yarandi, Y., and Shamsfard, M. (2023), "Predicting Compositionality of Verbal Multiword Expressions in Persian", in: *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, ed. A. Bhatia et al., Stroudsburg: Association for Computational Linguistics, pp. 14–23.

Savary, A., Khelil, C. B., Ramisch, C., Giouli, V., Mititelu, V. B., et al. (2023), "PARSEME Corpus Release 1.3", in: *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, ed. A. Bhatia et al., Stroudsburg: Association for Computational Linguistics, pp. 24–35.

Shamsfard, M. (2007), "Developing FarsNet: A Lexical Ontology for Persian", *Proceedings of the Fourth Global WordNet Conference / GWC (Szeged, Hungary, January 22-25, 2008)*, ed. A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, and P. Vossen, Szeged: University of Szeged, Department of Informatics, pp. 413-418