

Morphological Operations in the Persian Verbal System

Marzieh Rabiei Max Silberztein
Marie and Louis Pasteur University

Abstract

Persian is a morphologically rich language whose verbal system presents several challenges for computational modeling. In particular, the interaction between lexical preverbs, grammatical prefixes, and orthographic conventions often obscures the internal morphological boundaries of compound verbs. Additionally, prefixation in verbs with stems beginning with *l* or *l̤* (alef-initial stems) may trigger systematic morphophonological alternations that are not directly visible in the written form.

This paper proposes a formal computational description of these processes using electronic dictionaries and morphological grammars implemented in the NooJ linguistic development environment. We formalize two morphological operations designed to capture key structural properties of Persian verbal morphology and implement the corresponding operators: operator *<F>* identifies internal lexical boundaries and dynamically positions grammatical prefixes within compound verbs involving a restricted set of lexical preverbs, and operator *<G>* models conditional stem alternations in prefixed contexts through a controlled glide insertion mechanism.

The resulting description has been implemented and applied to a corpus of Persian verbal forms, producing an accurate representation of morpho-orthographic interactions in Persian verbs.

1 Introduction

This study falls within the field of Computational Linguistics, whose objective is to provide explicit and formal descriptions of natural languages through electronic dictionaries and structured morphological and syntactic grammars. In rule-based approaches, linguistic knowledge is encoded in a transparent and formally controlled way, allowing automatic parsing, annotation, and generation [1]. Formalization is therefore both a computational and theoretical commitment to making linguistic structure explicit.

In this context, the present work focuses on the Persian verbal system, which constitutes the morphologically richest component of the language (Windfuhr, 2009; Lazard, 1992). Persian verbs encode tense, aspect, mood, polarity, and person agreement, and combine verbal stems with both grammatical prefixes (*mi-*, *ne-*, *be-*, etc.) and lexical preverbs (*bar-*, *dar-*, *farâ-*, *zir-*, etc.). This interaction generates complex morphosyntactic configurations in which lexical structure, prefixation, and orthographic constraints must be respected simultaneously. Lexical preverbs form semantically integrated units, while grammatical prefixes must be inserted within these structures without violating the lexical integrity.

From a computational perspective, these properties raise a central question: How can the morpho-orthographic structure of Persian verbs be formally modeled in a way that preserves lexical structure, correctly positions grammatical prefixes, and accounts for prefix-triggered stem alternations? To address this question, we propose a formal modeling approach based on electronic dictionaries and morphological grammars. The study introduces mechanisms for dynamic boundary detection in prefixed verbs and conditional stem mutation in alef-initial verbs. The model is evaluated on a corpus of literary texts in order to assess both descriptive adequacy and computational robustness.

2 Persian Verbal Morphology and Orthography

2.1 Verbal Prefixes and Lexical Preverbs

Persian verbal morphology is characterized by the interaction between two types of preverbal elements: grammatical prefixes and lexical preverbs. Although both precede the verbal stem, they differ in their morphological status, semantic contribution, and computational treatment. Grammatical prefixes express inflectional categories such as aspect, mood, or polarity. The most frequent are *mi-* (imperfective), *ne-* / *na-* (negation), and *be-* (sub-

junctive or imperative). These prefixes are productive and combine systematically with verbal stems across paradigms without modifying lexical meaning (Windfuhr, 2009). In contrast, lexical preverbs form part of the lexical identity of the verb. Elements such as *بر-*, *در-*, *فر-*, *باز-*, *پس-*, and *پیش-* combine with a verbal base to create semantically distinct units that cannot be reduced to simple compositional combinations (Samvelian, 2012; Karimi-Doostan, 1997).

This distinction is illustrated in Table 1.

The addition of a lexical preverb creates a new lexical unit with its own semantic identity. These forms must therefore be treated as lexical entries, whereas grammatical prefixes operate at the inflectional level. Failing to distinguish these two layers leads to incorrect segmentation and prefix placement.

2.2 Orthographic Constraints

Persian orthography introduces constraints that directly affect morphological analysis. These concern (i) the absence of short vowels, (ii) obligatory attachment of lexical preverbs, and (iii) the use of the Zero Width Non-Joiner (ZWNJ).¹

Standard Persian writing does not represent short vowels, leaving vowel interpretation to the reader (Windfuhr, 2009; Lazard, 1992). As a result, a single orthographic form may correspond to multiple morphological analyzes, increasing the ambiguity in stem identification and inflection.

Many lexical preverbs are obligatorily attached to the verbal base. Forms such as *برداشتن* [bar dâštan] and *درآوردن* [dar âvardan] are written as single units, concealing internal morphological boundaries that remain relevant for inflectional processes.

The ZWNJ (*nim-fâsele*) prevents incorrect ligatures while preserving morphological cohesion. For example, *پیش رفتن* requires a ZWNJ between the preverb and the base, whereas *پیشرفت* is orthographically incorrect and *پیش رفتن* incorrectly separates the lexical unit (Academy of Persian Language and Literature, 2016).

These constraints show that orthography is not merely superficial. The absence of vowels, the obligatory attachment, and the usage of ZWNJ directly impact segmentation and the application of

rules. Therefore, any formal model of Persian verbal morphology must integrate orthographic structure as part of the morphological system.

3 Computational Approaches to Persian Morphology

Computational resources for Persian have been developed within several frameworks. Corpus-based resources, including the Bijankhan corpus (Bijankhan et al., 2004), the Hamshahri corpus (AleAhmad et al., 2003), and the Persian Dependency Treebank (Seraji et al., 2016), provide annotated data for tasks such as part-of-speech tagging and syntactic parsing. Morphological resources have also been developed specifically for Persian. For example, PerLex 2 provides a linguistically motivated morphological lexicon, used together with the MELtfa morphosyntactic tagger (Sagot et al., 2011). These resources constitute important lexical and annotation infrastructures, although their primary objective differs from the explicit generation and analysis of the morpho-orthographic processes considered here.

Beyond these corpus and lexical resources, Persian morphology has also been investigated from both descriptive and computational perspectives. Mahmoodi-Bakhtiari (2018) provides a comprehensive description of Persian morphology, including verbal inflection, derivation, and compounding. From a computational perspective, Shamsfard et al. (2010) integrate morphological analysis into the STeP-1 Persian text-processing toolkit, while also addressing orthographic normalization and spacing variation. More recently, Haghdoost et al. (2019) developed a morphological network for Persian on the basis of a morpheme-segmented lexicon. These studies illustrate the diversity of computational and linguistic approaches that have been proposed to model Persian morphology.

A particularly relevant point of comparison is the finite-state morphological analyzer proposed by Megerdooomian (2004). Using Xerox finite-state technology, this system provides a two-level analysis of Persian morphology and addresses complex verbal conjugation, morphological dependencies, orthographic variation, tokenization, and phonological alternations at morpheme boundaries. It therefore demonstrates that many of the linguistic phenomena addressed in the present study are well suited to explicit rule-based modeling.

The present work does not claim that these

¹The Zero Width Non-Joiner (ZWNJ) is a non-printing character used in Persian orthography to prevent the cursive connection between characters while preserving morphological boundaries.

Simple Verb	Meaning	Compound Verb	Meaning	Structural Effect
داشتن [dâštan]	to have	برداشتن [bar dâštan]	to pick up	lexical reanalysis
آوردن [âvardan]	to bring	درآوردن [dar âvardan]	to extract / bring out	semantic specialization
دادن [dâdan]	to give	پس دادن [pas dâdan]	to return	directional reinterpretation
رفتن [raftan]	to go	پیش رفتن [pish raftan]	to progress	metaphorical extension

Table 1: Lexical preverbs and semantic differentiation in Persian compound verbs

underlying linguistic phenomena are themselves novel. Rather, its contribution lies in their formalization within the NooJ framework through two structure-sensitive operators. Operator <F> dynamically locates internal lexical boundaries in verbs containing lexical preverbs in order to position grammatical prefixes, whereas operator <G> implements context-dependent glide insertion in prefixed alef-initial stems. Whereas finite-state implementations encode morphotactic and phonological constraints through finite-state lexicons, continuation paths, boundary markers, and composed rules, the present implementation expresses the targeted operations directly within NooJ morphological grammars through operator-based transformations.

The aim is therefore not to establish superiority over existing Persian morphological analyzers, but to provide an explicit NooJ implementation of these interactions and to evaluate its behavior on Persian textual data. A direct empirical comparison with existing finite-state systems on a shared gold-standard dataset remains an important direction for future work.

4 Formalizing Persian Morphology With NooJ

To formalize the morpho-orthographic structure described in the previous sections, we have used the NooJ linguistic development environment (Silberztein, 2016), a rule-based platform dedicated to the construction of electronic dictionaries and formal morphological, syntactic, and semantic grammars. This framework has allowed us to explicitly formalize morphosyntactic constraints using controlled rules and guarantees that our morphological description will be integrated in a future syntactic formalization of Persian syntax.

The following sections focus on the operator-based mechanisms implemented within this framework to model Persian-specific morphological phenomena. In NooJ, inflectional and derivational grammars are formalized as enhanced context-free grammars, in which rules associate sets of suffixes

to be concatenated to a lexical entry with their corresponding linguistic properties.

For example, to produce the inflected forms of the lexical entry *drink*, we need to delete its final letters before adding the suffix *unk*. In NooJ, this is done using the operator (backspace):

```
DRINK = <E>/STD | s/PR+3+s
| ing/G | <B3>unk/PP |
<B3>ank/PT ;
```

By adding the suffix <B3>unk to the lexical entry *drink* (resp. *stink*), one produces the word form *drunk* (resp. *stunk*), and then associates it with the property code PP (Past Participle).

NooJ offers several other generic operators, such as <L> (move cursor to the left), <P> (move to the previous wordform), and <C> (change case). Besides these generic operators available for any language, NooJ also provides operators that are specific to certain languages. We will discuss in Sections 5 and 6 the addition of two specific operators for Persian.

5 Prefix Handling and Boundary Detection

5.1 Linguistic Motivation

Persian includes a class of lexical preverbs such as *بر-*, *در-*, *فرا-*, *باز-*, *پیش*, and *پس*, which attach to a verbal base and form semantically integrated lexical units. Constructions such as *برداشتن* [bar dâštan] and *درآوردن* [dar âvardan] are therefore not syntactic sequences but lexicalized verbs (Lazard, 1992; Samvelian, 2012).

Grammatical prefixes (*mi-*, *na-*, *be-*) must be inserted between the preverb and the base without breaking the lexical unit.

For example:

برداشتن [bar dâštan, ‘to pick up’] →
برنداشته [bar-na-dâšte, ‘not picked up’]

where negation applies to the base *داشتن* [dâštan] rather than to the entire form. The incorrect form:

نبرداشته [*na-bar-dâšte]

illustrates the consequence of applying the pre-fixation at the word level instead of at the level of the verbal base.

This reveals a structural asymmetry: although the verb appears as a single orthographic unit, it contains an internal boundary that must be accessed for inflection.

5.2 Implementation

We introduce the operator <F> as a structure-sensitive positioning operator for internal boundary detection. When a lexical preverb is detected, <F> relocates the rule cursor to the boundary between the preverb and the verbal base, while preserving the lexical preverb:

- if a lexical preverb is detected, <F> shifts the cursor after the preverb;
- otherwise, <F> applies at the beginning of the word.

This operator relies on structure-sensitive positioning rather than linear insertion. For example, <F> correctly produces the negative past participle form برنداشته [bar-na-dâšte, ‘not picked up’] from the lexical entry برداشتن [bar dâštan, ‘to pick up’].

This mechanism provides a formal solution to prefix positioning in compound verbs and constitutes a necessary component of any computational model of Persian verbal morphology.

6 Glide Insertion in Prefixal Contexts

6.1 Linguistic Motivation

A second difficulty in modeling Persian verbal morphology concerns verbs beginning with $\bar{a}$ or $\bar{a}$ (alef-initial stems) in prefixal contexts.

Consider the verb آوردن [âvardan]. When the prefix ب is added, simple concatenation yields:

آوردن [âvardan] + ب [be-] →
بآورد [*be-âvarad]

However, the correct form is as follows:

بیاورد [bi-âvarad]

In this configuration, the initial $\bar{a}$ is replaced by a glide sequence $\bar{a}$. This corresponds to glide insertion in prefixal contexts, whereby Persian avoids the adjacency of two vowel nuclei.

This behavior is not uniform across all alef-initial verbs. For example:

ایستادن [istâdan, ‘to stand’] →

بایست [be-ist, ‘stand’] /

بایستم [be-istam, ‘I stand’]

Here, the initial alef is preserved, and no glide insertion occurs. The phenomenon is therefore lexically conditioned and cannot be reduced to a general rule.

6.2 Phonological Characterization

The prefixal context may create a hiatus between the vowel of the prefix and the initial vowel of the verbal stem. Persian resolves this configuration through glide insertion. However, this process is not uniform and must be distinguished into two distinct patterns depending on the phonological nature of the stem-initial element.

Case 1: Alef-madda-initial stems ($\bar{a}$) In this class of verbs, the stem begins with $\bar{a}$, which represents a long vowel. In prefixal contexts, this vowel is not deleted; instead, it is preserved and realized through a glide structure:

$\bar{a} \rightarrow \bar{a}$

For example:

آوردن [âvardan] → بیاورد [bi-âvarad]

In this case, the vocalic quality is maintained and surfaces as the sequence $\bar{a}$.

Case 2: Alef-initial stems ($\bar{a}$) In this second case, the stem begins with $\bar{a}$, which functions as a vowel carrier rather than a fully specified vowel. The actual vowel quality is determined by its phonological context (e.g., /a/, /o/).

In prefixal contexts, this initial $\bar{a}$ is deleted and replaced by the glide $\bar{a}$, while the associated vocalic quality is preserved and transferred onto the resulting structure.

For example:

افتادن [oftâdan] → بیفتد [bi-oftad]

Here, the initial carrier $\bar{a}$ disappears, the glide $\bar{a}$ is inserted, and the original vowel quality is maintained within the new structure.

These two patterns demonstrate that glide insertion in Persian is not a uniform process but is conditioned by the phonological nature of the stem-initial element. Consequently, any computational model must explicitly distinguish between $\bar{a}$ -initial and $\bar{a}$ -initial stems in order to correctly model prefix-triggered stem alternations.

6.3 Implementation

We introduce the operator <G> as a context-sensitive transformation operator for glide insertion in prefixed alef-initial verbs. When the relevant prefixed context is detected, <G> triggers the insertion of the glide *ی* and the corresponding stem-initial alternation. Its application is lexically conditioned by the phonological class of the stem and therefore does not apply uniformly to all alef-initial verbs.

7 Interaction

Operators <F> and <G> address complementary aspects of Persian verbal morphology: boundary detection and stem alternation. In many cases, both mechanisms must apply simultaneously. For example, in the verb درآوردن [dar âvardan], the negative form درنیاردم [dar-na-yâvardam] requires:

- <F> to detect the lexical preverb - در [dar] and correctly position the prefix;
- <G> to trigger the insertion of the glide that affects the stem آورد [âvard] in the prefixal context, producing forms such as نیاردم [na-yâvardam].

Without <F>, prefixation would incorrectly target the entire word; without <G>, the resulting form would be phonologically ill-formed. The correct output therefore depends on the interaction of both operators.

This interaction shows that Persian prefixation requires access to internal structure and conditional transformation, highlighting the necessity of structure-sensitive morphological modeling.

8 Evaluation

We used the NooJ software to apply our linguistic resources to a corpus of Persian texts. In natural data, most words are written without short vowels, and the space character is often used instead of the correct ZWNJ (Zero Width Non-Joiner) character.

In order to construct a linguistically controlled evaluation setting, we explicitly encoded vowels in our dictionaries and grammars and systematically used the correct ZWNJ character. However, when applied to real texts, NooJ processes both vowel absence and orthographic variation, including inconsistent use of spacing and ZWNJ.

The evaluation shows that the proposed operators correctly handle prefix positioning and stem

alternation across a wide range of verbal forms. In particular, <F> successfully detects internal boundaries in compound verbs, while <G> accurately models glide insertion in prefixal contexts.

These results demonstrate that a rule-based, structure-sensitive approach can effectively capture morpho-orthographic interactions in Persian verbal morphology, even in the presence of orthographic variability.

The results show that the combination of <F> and <G> allows correct generation and analysis of prefixed forms while preserving lexical integrity and handling orthographic constraints such as attachment and ZWNJ.

The evaluation was conducted on a contemporary Persian literary corpus containing 172,548 word forms (230,085 tokens in total).

The system identified 16,253 verbal forms using the implemented morphological resources.

Among these, 361 forms required the application of the <F> operator for correct prefix positioning, while 183 forms involved the <G> operator, corresponding to the alef-initial stems undergoing glide insertion. These counts provide a targeted evaluation of the two phenomena addressed by the proposed operators, namely internal prefix positioning in lexical-preverb constructions and glide insertion in prefixed alef-initial stems.

The system did not recognize a total of 192 verbal forms, corresponding to a recall of 98%. These unrecognized forms mainly reflect gaps in the current lexical coverage and indicate the need for further extension of the verbal lexicon. For example, forms such as تجمع کنند [tajammo' konand] and باز نیافته بود [bâz na-yâfte bud] correspond to compound verb constructions composed of a nominal element and a simple verb. These compound verbs are not yet fully included in the current lexical resources. Although approximately 2000 compound verbs have already been encoded in the dictionary, some constructions remain uncovered. In such cases, NooJ correctly identifies the simple verbal component, but does not fully analyze the entire compound structure.

In addition, 134 verbal forms were incorrectly analyzed, corresponding to a precision of 99%. These errors are mainly due to orthographic variation, especially the use of space instead of the Zero Width Non-Joiner (ZWNJ), a problem that can be addressed within the NooJ framework.

Table 2 summarizes the main quantitative re-

extension would require adaptation to its specific lexical and morphophonological properties.

Overall, this work underscores the importance of structure-sensitive modeling for morphologically rich and low-resource languages. It provides a robust foundation for future developments, including the integration of richer lexical resources and the extension of the model toward syntactic and semantic processing.

References

- Academy of Persian Language and Literature. 2016. Persian orthography guidelines. Tehran.
- Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2003. Hamshahri: A standard persian text collection. In *Proceedings of the ACL Workshop on Computational Approaches to Arabic Script-based Languages*.
- Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2004. Lessons from building a persian written corpus: Peykare. *International Journal of Corpus Linguistics*, 9(2):195–213.
- Hamid Haghdoost, Ebrahim Ansari, Zdeněk Žabokrtský, and Mahshid Nikraves. 2019. Building a morphological network for Persian on top of a morpheme-segmented lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 91–100, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.
- Gholamhossein Karimi-Doostan. 1997. Light verb constructions in persian. *Studies in the Linguistic Sciences*, 27(2):149–173.
- Gilbert Lazard. 1992. *A Grammar of Contemporary Persian*. Mazda Publishers.
- Behrooz Mahmoodi-Bakhtiari. 2018. **Morphology**. In Anousha Sedighi and Pounesh Shabani-Jadidi, editors, *The Oxford Handbook of Persian Linguistics*, pages 273–299. Oxford University Press.
- Karine Megerdooian. 2004. Finite-state morphological analysis of Persian. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 35–41, Geneva, Switzerland. COLING.
- Benoît Sagot, Géraldine Walther, Pegah Faghiri, and Pollet Samvelian. 2011. Développement de ressources pour le persan : PerLex 2, nouveau lexique morphologique et MELtfa, étiqueteur morphosyntaxique. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 73–78, Montpellier, France. ATALA.
- Pollet Samvelian. 2012. Auxiliation in persian. *Lingua*, 122(5):514–535.
- Mojgan Seraji, Beata Megyesi, and Joakim Nivre. 2016. Building a persian dependency treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.
- Mehrnoush Shamsfard, Hoda Sadat Jafari, and Mahdi Ilbeygi. 2010. STeP-1: A set of fundamental tools for Persian text processing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Max Silberztein. 2016. *Formalizing Natural Languages: The NooJ Approach*. Wiley.
- Gernot Windfuhr. 2009. *The Iranian Languages*. Routledge.