Building a Large Persian Verb Collection: A Generative Approach

Sayed Nasir Khalifehsoltani, Ali Cholmaghani, Ali Vahdani, Reza Moallemi
Department of Computer Engineering
SheikhBahee University
Isfahan, Iran
n_khalifesoltani@yahoo.com, ali.cholmaghani@gmail.com, alivahdani@gmail.com, moallemi.reza@gmail.com

Abstract— **Persian is an Iranian language within the Indo-Iranian branch of the Indo-European languages which is one of the major languages in the Heart Land (Middle East). Nevertheless, there are proportionately few studies on retrieval and analyses of Persian documents. In this case, suitable resources for Persian are becoming a vital necessity for the progress of language engineering researches. In this paper we tried to make a Persian Verb Collection -as a linguistic resource- which is needed in some NLP researches like verb and sentence detection, POS tagging, Lexicography and etc. We used a generative approach for building this collection from Persian Stems by Verbal Structures.**

Keywords-Persian Language, Verb Conjugation, Natural Language Processing, Linguistic Resource

I. INTRODUCTION

Persian is spoken today primarily in Iran, Afghanistan and Tajikistan, but was historically a more widely understood language in an area ranging from the Heart Land (Middle East) to India. Significant populations of speakers in other Persian Gulf countries as well as large communities around the World. Nevertheless, there are proportionately few available linguistic resources for thirsty researchers.

There are many articles about verb morphology in some languages like [1, 2, 3, 4, 5, 6, 7]. Karine Megerdoomian described Persian morphology in [8].

The richly inflected morphological system of Old Iranian has been drastically reduced in Persian. The language has no grammatical gender or articles, but person and number distinctions are maintained. Nouns are marked for specificity: there is one marker in the singular and two in the plural. Objects of transitive verbs are marked by a suffix. The morphological features of Arabic words are preserved in loans. Verbs are formed using one of two basic stems, present and past; aspect is as important as tense: all verbs are marked as perfective and imperfective. The latter is marked by means of prefixation. Both perfective and imperfective verb forms appear in three tenses: present, past and inferential past. The language has an aorist (a type of past tense), and has three moods: indicative, subjunctive, counterfactual. Passive is formed with the verb “to become”, and is not allowed with specified agents. Verbs agree with the subject in person and number. Persian verbs are normally compounds consisting of a noun and a verb.

Although some articles [9, 10, 11, 12] are dedicated to Persian verbs, but there is no any equivalent verb specific available resource in Persian. In this paper we tried to build a large Persian Verb Collection as a linguistic resource with a generative approach. Section II presents some background concepts about Persian verbs and their inflection system. Section III is dedicated to our solution for extracting stems and generating verbs from them with described formulation. Appendix I shows a brief view of our verb generation structure.

II. BACKGROUND CONCEPTS

A. Persian Verbs

Persian is an ancient language of Indo-European family. You can find many grammatical similarities between Persian and the other languages of this family. However, Persian is similar more to its coeval languages like Latin than to relatively newer languages. Persian has a SOV word order: normal sentences are structured Subject-preposition-Object-Verb. If the object is specific, then the order is "(S) (O + "rā") (PP) V". Persian is a pro-drop language, thus the subject is optional. The object marker râ (را) is used to indicate specific direct objects in simple sentences [13]. However, Persian can have relatively free word order, often called "scrambling". This scrambling characteristic has allowed Persian a high degree of flexibility for versification and rhyming. Although this is not so strict in the colloquial usage of the language, the verb is usually the last element in a Persian sentence.

Indo-European languages usually inflect verbs for several grammatical categories in complex paradigms, although some, like English, have simplified verb conjugation to a large extent. Verbs are probably the most complex topic in the Persian grammar, although (not surprisingly) they work a lot like verbs in many other Indo-European languages, and not like e.g. Arabic verbs at all. Persian verbs are conjugated for each person, so they have three forms for the singular (1st, 2nd, 3rd) and three forms for the plural (1st, 2nd, 3rd) in all tenses (excepting in the Imperative mood, where only the 2nd person singular and 2nd person plural are present). Because of that, personal pronouns are not usually necessary, except for clarification or special emphasis. Persian conjugation does not reflect gender, though.

Persian verbs are conjugated for both aspect and tense. Aspect consists of perfective and imperfective forms; tenses consist of present, past and reported past. The verbal system comprises four moods: indicative, subjunctive, conditional and imperative. The full conjugation system is based on two different stems, the present stem and the past stem. Like English, Persian has affixitive morphology. In other words, suffixes and prefixes are concatenated to Persian words to modify the meaning. Persian verbs are modified more extensively than English verbs. Persian verbs vary form according to tense, person, negation, and mood. Therefore, a given verb may have scores of variations. [12]

B. *Persian Verbal Inflection*

The inflectional system for the Persian verbs consists of simple forms and compound forms; the latter are forms that require an auxiliary verb. The simple forms are divided into two groups according to the stem they use in their formation: the tenses that use the Present Stem and those formed on the Past (or Aorist) Stem. The Present Stem needs to be specified in the lexicon since it cannot be derived, while the Past Stem is easily derivable from the infinitival form of the verb. The citation form for the verb is the infinitive.

In addition to the verb stems, the following elements also participate in the formation of the verbal inflectional system in Persian [14]:

- **Prefixes**: the imperfective prefix *mee* (می) and the morpheme *b* (ب) or *bee*, which characterizes the subjunctive and the imperative. Negation is marked by the *n* (ن) or m (م) or *nee* prefix.
- **Personal Inflections**: present, past and imperative personal inflections are used in conjugating the Persian verb. All verb forms are marked for person and number.
- **Suffixes**: the suffix *and e* marks the present participle ending and *e* (written *h*) (ه) is used to form the past participle.
- **Causation morpheme**: causatives are obtained by adding the affix *ân* (ان) or *âni* (انی) to the end of the Present Stem of the verb. Personal inflections and suffixes can then be attached to the Causative Present Stem to derive all verbal forms for the causative construction.
- **Auxiliaries**: Persian conjugation uses a number of auxiliaries in the compound forms. The enclitic form of the auxiliary *budan* (be) (بودن) is the one used in the formation of the perfect forms of all verbs. The verb *khâstan* (want) (خواستن) is used as an auxiliary in forming the future tenses. The auxiliary *shodan* (become) (شدن) forms the passive constructions.

The complete inflectional system can be obtained by the various combinations of these elements as shown in Appendix I.

III. Our Approach

As we mentioned, in Persian, each verb has two stems: Past and Present. By knowing these two stems of a given verb, we can conjugate it. Among the simple verbs, the tenses that are formed using the Present Stem are the present, the simple subjunctive, the imperative and the present participle. On the Past Stem are formed the preterit, the imperfect, the past participle. Among the compound forms, the future is formed on the Past Stem. All the other compound forms are based on the past participle.

The Past Stem always obtains regularly by removing -an from the infinitive e.g. (رفتن) raftan (to go) = raft (رفت). There isn't such a rule for obtaining the present stem of verbs but they can be classified into subgroups whose Present Stem is obtained according to a regular pattern with no or few exceptions. However, a verb whether regular or irregular has one and only one Present Stem for all persons. Therefore, as opposed to languages like French, Italian and Spanish, Persian does not have irregular verb conjugations. The past participle forms by replacing the infinitive suffix (-an) with -e. In other words, by adding -e to the Past Stem e.g. (رفتن) raftan = rafte (رفته).

Obtaining the Present Stem is tricky, though. There are some guidelines, such as dropping the endings (ـیدن) -idan, (ـادن) -âdan, (ـستن) -(e)stan etc., but there are so many exceptions and irregularities that it is usually easier just to learn the present indicative by heart and derive the present stem directly from it (by dropping both the prefix and the suffix).

Making of the Persian Verb Collection includes two main steps: colleting stems and conjugating them through the structure. Finally, we discussed about disambiguation of generated verbs and availability of this collection.

A. *Collecting Stems*

In this step we tried to gather stems with two methods: grabbing stems from available dictionaries and extracting verbs from some prominent Persian corpuses.

According to Dr. Bateni [15] Persian language has 277 simple verbs which some of them are deprecated in today's writing and speaking. Among them 115 verbs are common in both modern writing and speaking. About 150-200 verbs estimated as Persian Live Verbs with respect to less frequent modern spoken verbs but still usable in modern writing. On the other hand, compound verbs have an infinite nature.

1) Grabbing Stems from Persian Verb Dictionaries: We grabbed 297 simple verbs and 992 compound verbs from an online Persian verb list for an online course developed by National Middle East language resource centre of Brigham Young University[1]. Then we used Jahanshiry's[2] online PVC (Persian Verb Conjugator) to stem grabbed verbs.

2) Extracting Stems from Corpus: We tried to use large available Persian corpuses as a resource to collect more Verbs. Hence, we processed Bijankhan [16], Hamshahri [17] and TebCorp[3] [18] collections for this reason. Among them Bijankhan is a tagged collection and has no need to verb detection. For detecting verbs in Hamshahri and

[1] http://sartre2.byu.edu/persian/courses/PRS506/PVClist.html

[2] http://www.jahanshiri.ir

[3] http://tebcorp.sourceforge.net

TebCorp we matched 554 patterns of all forms of cojugated verbal structure which some of them shown in appendix I with this collections. In this step 663 distinct stems were detected from more than 75,727 verbs which about 53,000 were duplicated. Finally 58 new stems (in spite of verbs acquired from previous step) were acquired from these corpuses.

B. Conjugating Verbs using Structures

In this step, each acquired stem from previous steps will generate variety of verbs with Persian verbal inflection system by language specific structures. As appendix I shows, in brief view, we collect 89 structure with 554 conjugated forms for Persian verbs. Figure I demonstrates a detailed view of Persian verb conjugation variables in Persian Verb Collection. Combination of these variables with predefined structures using large list of acquired stems (about 1300 infinitives) were made a very large collection of Persian verbs.

Finally, in its first version, Persian Verb Collection was contained more than 860,000 conjugated verbs with this generative approach.

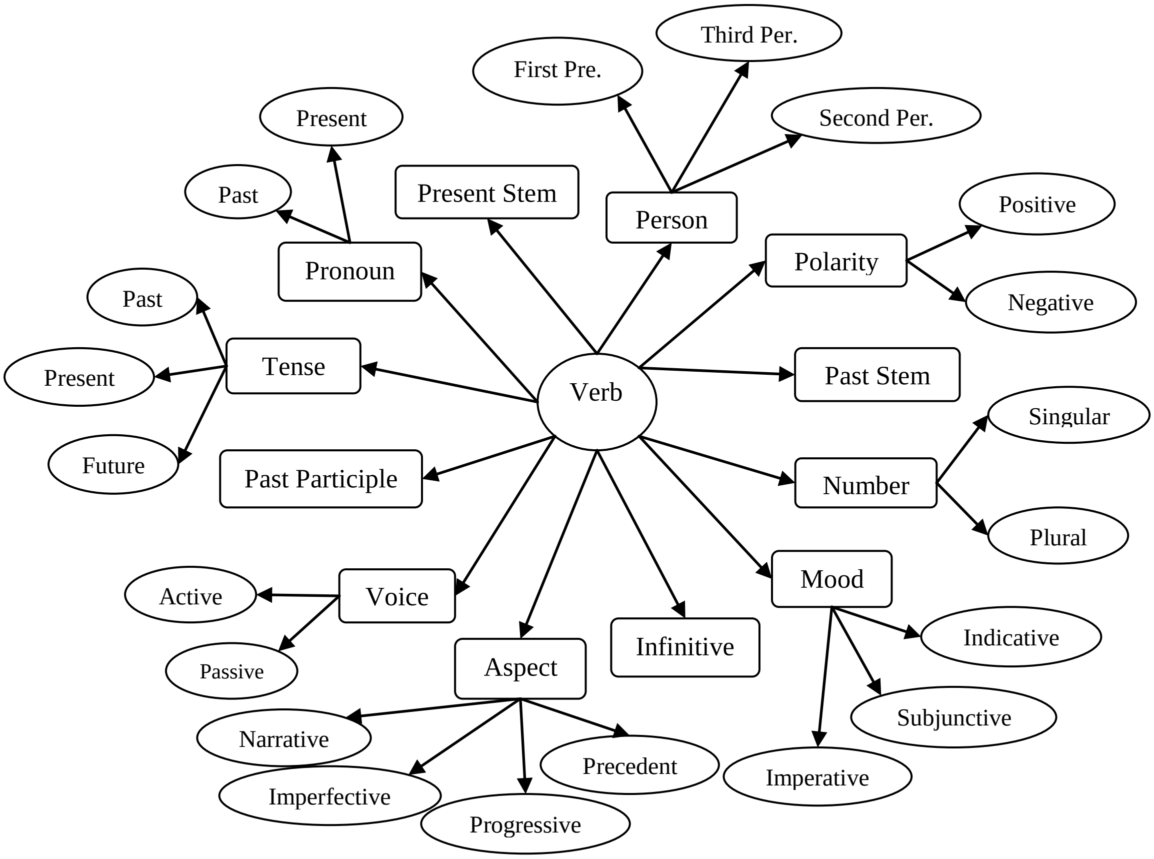

FIGURE I. PERSIAN VERB CONJUGATION VARIABLES IN PERSIAN VERB COLLECTION

```
<Verb>
        <Conjugate ></ Conjugate>
        <Infinitive></ Infinitive>
        <Transcription></ Transcription>
        <Past_Stem></ Past_Stem>
        <Present_Stem></ Present_Stem>
        <Past_Participle></ Past_Participle>
        <Number></ Number>
        <Person></ Person>
        <Polarity></ Polarity>
        <Pronoun></ Pronoun>
        <Tense></ Tense>
        <Mood></ Mood>
        <Aspect></ Aspect>
        <Voice></ Voice>
</Verb>
```

FIGURE II. STRUCTURE OF XML DISTRIBUTION

C. *Verb Disambigution and Translation*

The Persian language has six vowel phonemes and twenty-three consonant phonemes. Because the Perso-Arabic script is an abjad writing system (with a consonant-heavy inventory of letters), many distinct words in standard Persian can have identical spellings, with widely varying pronunciations that differ in their (unwritten) vowel sounds which can make ambiguity in our verb collection.

We made Transcription of each verb to disambiguate identical spelling verbs in our collection. Transcriptions of Persian attempt to straightforwardly represent Persian phonology in the Roman alphabet, without requiring a close or reversible correspondence with the Perso-Arabic script, and also without requiring a close correspondence to English-language phonetic values of Roman letters.

Moreover, to clearly define each verb, English translation of infinitives were included in Persian Verb Collection.

D. *Persian Verb Collection Advantage and Availability*

Even though major Persian corpuses contain a lot of conjugated verbs but they together have lesser verbs than the Persian Verb Collection Because of sparseness in their conjugation. Moreover, Persian Verb Collection consists of more stems gathered from other resources. Table I shows comparison of the Persian Verb Collection with major Persian corpuses.

Persian Verb Collection[1] was distributed as a XML file and a MySQL dump file. XML version of the Persian Verb Collection was generated with the structure that is demonstrated in Figure II. These files are downloadable as a package from Sourceforge website and can be used freely for noncommercial purposes.

TABLE I. COMPARISON OF THE PERSIAN VERB COLLECTION WITH MAJOR PERSIAN CORPUSES

Collection	Number of detected stems	Number of detected verbs
Bijankhan	107	6,862
Hamshahri	313	23,162
TebCorp	663	75,727
Persian Verb Collection	1,289	680,456

IV. CONCLUSION AND FUTURE WORKS

In this paper we tried to build a large conjugated verb collection using a generative approach for Persian language as a linguistic resource for NLP researches. The Persian Verb Collection can be used in some NLP researches like verb and sentence detection, POS tagging, Lexicography and etc. We plan to build a large Persian Lexicon from the Persian Web in future.

REFERENCES

[1] R. Hetzron, “The morphology of the verb in Modern Syriac (Christian colloquial of Urmi)”, Journal of the American Oriental Society, 1969.

[2] K. Hansson, U. Nettelbladt, L.B. Leonard, “Specific language impairment in Swedish: The status of verb morphology and word order”, Journal of Speech, Language and Hearing Research, 2000.

[3] A.M. Gimba, “Bole verb morphology”, 2000.

[4] G. Görz, D. Paulus, “A finite state approach to German verb morphology”, Proceedings of the 12th conference on Computational Linguistic, 1988.

[5] E. Pizzuto, M.C. Caselli, “The acquisition of Italian verb morphology in a cross-linguistic perspective”, Other children, other languages: Issues in the theory, 1994.

[6] A. Koutsoudas, “Verb Morphology of Modern Greek: a descriptive analysis”, Indiana University Research Center in Anthropology, 1962.

[7] F. Wouk, “The Impact of Discourse on Grammar: Verb Morphology in Spoken Jakarta Indonesian”, UMI Dissertation Services, Ann Arbor, Mich, 1989.

[8] M. Karine, “Persian Computational Morphology: A Unification-Based Approach”, NMSU, CRL, Memoranda in Computer and Cognitive Science, 2000.

[9] K. Taghva, R. Beckley, M. Sadeh, “A stemming algorithm for the farsi language”, Proceedings of the International Conference on, 2005.

[10] G. Karimi-Doostan, “Light verb constructions in Persian”, iranianlinguistics.org, 1997.

[11] M. Dabir-Moghaddam, “Compound verbs in Persian”, Studies in the Linguistic Sciences, 1997.

[12] Iranpour Mobarakeh, Majid Minaei-Bidgoli, Behrouz, "Verb Detection in Persian Corpus", JDCTA: International Journal of Digital Content Technology and its Applications, Vol. 3, No. 1, pp. 58 ~ 65, 2009

[13] M. Karine, “Developing a Persian Part-of-Speech Tagger”, In Proceedings of the First Workshop on Persian Language and Computers, Tehran University, Iran, 2004.

[14] M. Karine, “A Semantic Template for Light Verb Constructions”, In Proceedings of the First Workshop on Persian Language and Computers, Tehran University, Iran, 2004.

[15] M. Bateni, “Towsif-e Sakhteman-e Dastury-e Zaban-e Farsi [Description of the Linguistic Structure of Persian Language]”, Amir Kabir Publishers, Tehran, Iran, 1995.

[16] M. BijanKhan, “The role of the corpus in writing a grammar: an introduction to a software”, Iranian Journal of Linguistics, Vol. 19, No. 2, 2004.

[1] http://persianverbs.sourceforge.net

[17] E. Darrudi, M.R. Hejazi, F. Oroumchian, “Assessment of a Modern Farsi Corpus”, In Proceedings of the 2nd Workshop on Information Technology & its Disciplines (WITID), ITRC, Kish Island, Iran, 2004.

[18] SN. Khalifehsoltani, A. Cholmaghani, A. Vahdani, R. Moallemi, “Towards Acquisition of a Thematic Persian Corpus from the Tebyan Portal: TebCorp”, The 2nd International Conference on Knowledge Discovery, Bali Island, Indonesia, 2010.

APPENDIX I

TABLE II. PERSIAN VERBAL STRUCTURE

Sample	structure	Voice	Tense	Mood
به دست آوردم	[پیش‌فعل] + [na] + ستاک گذشته + شناسه‌ی گذشته	معلوم	گذشته‌ی ساده Past simple	گزارشی Indicative
به دست آورده شدم	[پیش‌فعل] + [na] + ماده‌ی گذشته + گذشته ساده «شدن»	مجهول		
به دست می‌آوردم	[پیش‌فعل] + [ne] + می + ستاک گذشته + شناسه‌ی گذشته	معلوم	گذشته‌ی پاﻻ Past imperfective	
به دست آورده می‌شدم	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + گذشته پایای «شدن»	مجهول		
داشتم به دست می‌آوردم	«داشتن» به گذشته‌ی ساده + گذشته‌ی پایا	معلوم	گذشته‌ی روان Past progressive	
داشتم به دست آورده می-شدم	«داشتن» به گذشته‌ی ساده + گذشته‌ی پایا مجهول	مجهول		
به دست می‌آورده‌ام	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + شناسه‌ی گذشته	معلوم	گذشته‌ی نقلی پایا Past narrative imperfective	
به دست آورده می‌شده ام	[پیش‌فعل] + [ne] + ماده‌ی گذشته + گذشته نقلی پایا «شدن»	مجهول		
داشته‌ام به دست می‌آورده‌ام	«داشتن» به گذشته‌ی نقلی + گذشته‌ی نقلی پایا	معلوم	گذشته‌ی نقلی روان Past narrative progressive	
داشته‌ام به دست آورده می‌شده‌ام	«داشتن» به گذشته‌ی نقلی + گذشته‌ی نقلی پایا مجهول	مجهول		
به دست آورده بودم	[پیش‌فعل] + [na] + ماده‌ی گذشته + «بودن» به گذشته‌ی ساده	معلوم	گذشته‌ی پیشین Past precedent	
به دست آورده شده بودم	[پیش‌فعل] + [na] + ماده‌ی گذشته + گذشته‌ی پیشین «شده»	مجهول		
به دست می‌آورده بودم	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + «بودن» به گذشته‌ی ساده	معلوم	گذشته‌ی پیشین پایا Past precedent imperfective	
به دست آورده می‌شده بودم	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + گذشته‌ی پیشین پایا «شدن»	مجهول		
داشتم به دست می‌آورده بودم	«داشتن» به گذشته‌ی ساده + گذشته‌ی پیشین پایا	معلوم	گذشته‌ی پیشین روان Past precedent progressive	
داشتم به دست آورده می شده بودم	«داشتن» به گذشته‌ی ساده + گذشته‌ی پیشین پایا مجهول	مجهول		
به دست آورده بوده‌ام	[پیش‌فعل] + [na] + ماده‌ی گذشته + «بودن» به گذشته‌ی نقلی	معلوم	گذشته‌ی پیشین نقلی Past precedent narrative	
به دست آورده شده بوده‌ام	[پیش‌فعل] + [na] + ماده‌ی گذشته + گذشته‌ی پیشین نقلی «شدن»	مجهول		
به دست می‌آورده بوده‌ام	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + «بودن» به گذشته‌ی نقلی	معلوم	گذشته‌ی پیشین نقلی پایا Past precedent narrative imperfective	
به دست آورده می شده - بوده‌ام	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + گذشته‌ی پیشین نقلی پایا «شدن»	مجهول		
داشته‌ام به دست می‌آورده بوده‌ام	«داشتن» به گذشته‌ی نقلی + گذشته‌ی پیشین نقلی پایا	معلوم	گذشته‌ی پیشین نقلی روان Past precedent narrative progressive	
داشته‌ام به دست آورده می‌شده بوده‌ام	«داشتن» به گذشته‌ی نقلی + گذشته‌ی پیشین نقلی پایا مجهول	مجهول		
به دست آورم	[پیش‌فعل] + [na] + بن حال + شناسه‌ی حال	معلوم	حال ساده Present simple	
به دست آورده شوم	[پیش‌فعل] + [na] + ماده‌ی گذشته + حال ساده «شدن»	مجهول		
به دست می‌آورم	[پیش‌فعل] + [ne] + می + بن حال + شناسه‌ی حال	معلوم	حال پایا Present imperfective	
به دست آورده می‌شوم	[پیش‌فعل] + [ne] + ماده‌ی گذشته + حال پایا «شدن»	مجهول		
دارم به دست می‌آورم	«داشتن» به حال ساده + حال پایا	معلوم	حال روان Present progressive	
دارم به دست آورده می-شوم	«داشتن» به حال ساده + حال پایا مجهول	مجهول		
به دست خواهم آورد	[پیش‌فعل] + [na] + «خواستن» به حال ساده + ستاک گذشته	معلوم	آینده‌ی ساده Future simple	
به دست آورده خواهم شد	[پیش‌فعل] + [na] + ماده‌ی گذشته + آینده ساده «شدن»	مجهول		
به دست آورده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + «بودن» به حال ساده‌ی التزامی	معلوم	گذشته‌ی نقلی Past narrative	التزامی Subjunctive
به دست آورده شده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + گذشته نقلی «شدن» به حال ساده‌ی التزامی	مجهول		
به دست می‌آورده باشم	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + «بودن» به حال ساده‌ی التزامی	معلوم	گذشته‌ی نقلی پایا Past narrative imperfective	
به دست آورده می‌شده باشم	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + گذشته‌ی نقلی پایا التزامی «شدن»	مجهول		
به دست آورده بوده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + «بودن» به گذشته‌ی نقلی التزامی	معلوم	گذشته‌ی پیشین نقلی Past precedent narrative	
به دست آورده بوده شده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + گذشته‌ی پیشین نقلی التزامی «شدن»	مجهول		
به دست می‌آورده بوده باشم	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + «بودن» به گذشته‌ی نقلی التزامی	معلوم	گذشته‌ی پیشین نقلی پایا Past precedent narrative imperfective	
به دست آورده می‌شده بوده باشم	[پیش‌فعل] + [ne] + می + ماده‌ی گذشته + گذشته‌ی پیشین نقلی پایا التزامی «شدن»	مجهول		
به دست بیاورم	[پیش‌فعل] + [na] + be + ستاک حال + شناسه‌ی حال	معلوم	حال ساده Present	
به دست آورده بشوم	[پیش‌فعل] + [na] + حال ساده التزامی «شدن»	مجهول		
به دست آورده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + «بودن» به حال ساده‌ی التزامی	معلوم	گذشته‌ی نقلی Past narrative	خواستاری Imperative
به دست آورده شده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + گذشته نقلی «شدن» به حال ساده‌ی التزامی	مجهول		
به دست آورده بوده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + «بودن» به گذشته‌ی نقلی التزامی	معلوم	گذشته‌ی پیشین نقلی Past precedent narrative	
به دست آورده بوده شده باشم	[پیش‌فعل] + [na] + ماده‌ی گذشته + گذشته‌ی پیشین نقلی التزامی «شدن»	مجهول		
به دست بیاور	[پیش‌فعل] + [na] + be + ستاک حال + شناسه‌ی حال	معلوم	حال ساده Present	
به دست آورده بشوم	[پیش‌فعل] + [na] + حال ساده التزامی «شدن»	مجهول		