# Augmenting FarsNet with New Relations and Structures for verbs

**Mehrnoush Shamsfard**
Faculty of Computer Science and Engineering
Shahid Beheshti University,
Tehran, Iran.
m-shams@sbu.ac.ir

**Yasaman Ghazanfari**
NLP Research lab.
Shahid Beheshti University,
Tehran, Iran.
yasaman_ghazanfari@yahoo.com

## Abstract

This paper discusses the semantic augmentation of FarsNet -the Persian WordNet- with new relations and structures for verbs. FarsNet1.0, the first Persian WordNet obeys the Structure of Princeton WordNet 2.1. In this paper we discuss FarsNet 2.0 in which new inter-POS relations and verb frames are added. In fact FarsNet2.0 is a combination of WordNet and VerbNet for Persian. It includes more than 30,000 lexical entries arranged in about 20,000 synsets with about 18000 mappings to Princeton WordNet synsets. There ae about 43000 relations between synsets and senses in FarsNet 2.0. It includes verb frames in two levels (syntactic and thematic) for about 200 simple Persian verbs.

## 1 Introduction

The Persian language, also known as Farsi, is a member of the Iranian group of the Indo-Iranian sub-family of the Indo-European languages. It is the official language of Iran, Afghanistan and Tajikistan with more than 100 million speakers.
In Persian verbs are the main carriers of a sentence meaning like many other languages. They may appear in simple or complex forms. Simple verbs have simple morphological structure, the verbal constituent. Compound verbs, on the other hand, consist of a nonverbal constituent, such as a noun, adjective, past participle, prepositional phrase, or adverb, and a verbal constituent.
In this paper we focus on the new relations and structures added to Persian WordNet (FarsNet) for verbs.
In the rest of the paper we first have an overview on FarsNet, the Persian WordNet and its features in the last two versions. Section 3 talks about verb argument structures and frames. Section 4 discusses the developed corpus management system in which argument structures are extracted and tagged. Section 5 is dedicated to results and discussion and at last section 6 concludes the paper.

## 2 FarsNet: The Persian WordNet

FarsNet project was announced with the release of FarsNet 1.0 at 2008. FarsNet 1.0 included the lexical, syntactic and semantic knowledge about more than 17000 Persian words and phrases organized in about 10000 synsets of nouns, adjectives and verbs. It was a medium scaled WordNet like the Arabic one (at that time). Table 1 shows the statistics of FarsNet 1.0.

Table 1. FarsNet 1.0 Statistics

| POS Category | Word | Sense | Synset |
|---|---|---|---|
| Noun | 9488 | 14079 | 5180 |
| Verb | 4402 | 6028 | 2306 |
| Adjective | 3950 | 4363 | 2526 |
| Total | 17842 | 24480 | 10012 |

As it can be seen for each word in FarsNet 1.0 we have an average of 1.5 senses and each synset includes an average of .1.7 words.
FarsNet 1.0 was developed by a semiautomatic approach. The base concepts covered in FarsNet were chosen from the base concepts BCS1 and BCS2 of BalkaNet (Tufis, 2004) with an equivalent in Persian to achieve compatibility with other WordNets. And also from the most frequent words of two Persian corpora: Peykareh (Bijankhan, 2004) and PLDB (Assi, 1997) to preserve the Persian specific structures (Shamsfard, et. al, 2010).
FarsNet 1.0 had two main classes of relations defined: inner language and inter-language relations. Synonymy, hypernymy and hyponymy,

different types of meronymy, Antonymy and cause were among the inner-language relations. The second class included the relations equal-to and near-equal-to between FarsNet and WordNet 3.0 synsets as inter-language relations. All inner-language relations were inner-POS; which means that their domain and range were from the same POS category. In other words FarsNet 1.0 did not cover inter-POS relations.

At 2010 a major restructuring of FarsNet began which resulted in FarsNet 2.0. The main goals of the changes were enlarging the size (improving the quantity) along with enhancing the quality. The new version was supposed to include new PoS category, new types of relations and new structures.

FarsNet 2.0 extends FarsNet 1.0 in the following dimensions:

- Size: FarsNet 2.0 includes more than 30,000 lexical entries organized in about 20,000 synsets with about 43,000 relations and 18000 mappings to Princeton WordNet 3.0. The size is approximately doubled comparing to FarsNet 1.0. In FarsNet 2.0 Princeton base concepts are included in addition to the base concepts of BalkaNet.
- POS categories: FarsNet 2.0 adds the adverb category to FarsNet 1.0. It includes nouns, verbs, adjectives and adverbs now.
- Number and type of relations: FarsNet 2.0 includes inter-POS as well as inner-POS relations. 'Derivational form', antonymy, 'verbal part of' and 'non-verbal part of' are relations between word senses. 'Verbal (non-verbal)-part-of' is a new relation between a compound verb and its verbal (non-verbal) component.

  From the synset relations, in addition to hypernym (as between peach and fruit), hyponym (as between food and hamburger) , various types of meronym (as between apple and apple juice) and holonym (as between car door and car ) entailment (as between snore and sleep) and cause (as between kill and die) which were all present at FarsNet 1.0 as well as at Princeton WordNet (Fellbaum, 1998), FarsNet 2.0 includes the following relations:

  - Has-attribute / is-attribute-of: the relation between a quantitative adjective and the attribute whose value is the adjective. For example the relation between heavy and weight or between warm and temperature
  - Domain / is-domain-of: the relation between a domain specific term and its corresponding domain. For example between Carbid and chemistry or between arthritis and medicine.
  - Agent/ Is-agent-of: the relation between a predicate (verb) and the potential agent of it. For example between author and writing or chef and cooking.
  - Patient/ Is-patient-of: the relation between a predicate (verb) and the potential patient or theme of it. For example between eat and edible thing or write and letter.
  - Instrument/ Is-instrument-of: the relation between a predicate (verb) and the potential instrument of it. For example between eat and spoon or write and pen.
  - Corresponding adjective: The relation between an adjective and the noun it often/ mainly describes. For example the relation between Stale and bread.
  - 'Related to'- the relation between any two synsets which has a semantic relation other than the previous named relations. For example the relation between author and book or between school and teaching.

  The above relations except the "domain/is domain of" and "has attribute/is attribute of" are new to both FarsNet and Princeton WordNet. Their creation is motivated by various NLP tasks. For example the relations between a predicate and its arguments such as agent, patient and instrument help semantic role labelers, word sense disambiguation (WSD) modules and information/ knowledge extraction systems to better find the corresponding relations and do their jobs.

  "related-to" relation is used to relate any two synsets which has a sort of relation not included in the above named relations. Although the relation between some of the related concepts could be extracted by traversing the links in Princeton WordNet or FarsNet 1.0, the new relation specifies the important ones explicitly. It is mostly used in information retrieval and also in finding similarity between text components for example in text summarization.
- New structure- FarsNet 2.0 is actually a combination of Persian WordNet and Persian VerbNet. It includes the verb frames (argument structure) of about 200 Persian simple verbs along with the selectional restrictions of their arguments. In the rest of the paper we discuss this new feature in more details.

## 3 Augmenting FarsNet with Verb Frames

### 3.1 Argument structures and Verb frames

FarsNet 2.0 includes the information about the argument structure of verbs and their selectional restrictions. In this part it is somehow similar to resources like VerbNet (Kipper, et al., 2006) developed for English language.

When talking about the semantic relationships among different entities within a sentence, the most relevant term is proposition. The core semantic content of every sentence is called a proposition which in turn consists of a predicate and one or more arguments (Brinton & Brinton, 2010). The arguments may appear in the form of a noun phrase, a propositional phrase, an adjective or adverb phrase or a sentence.

The argument structure (or frame) of a verb can be defined as the representation of that verb regarding the nature and number of participants it requires. In other words, it is considered as the kind of semantic relationship which holds among verb and other obligatory constituent within a sentence [Ghazanfari, 2014]. Other expressions in the sentence whose existence are optional are called adjuncts. The number of arguments of a verb makes its valency. Verb valence may be from zero to 4 (Dixon, 2000).

In many NLP applications, knowing the verb arguments can help parsers and analyzers to process and disambiguate the text. The arguments are the constituents of a sentence which complete the meaning of its verb.

Arguments can be defined in different levels: syntactic (such as NP, PP,…), grammatical (such as subject, object, …) and thematic or semantic (such as agent, patient, theme, …). In syntactic level, arguments are represented by their POS categories. For example the verb خندیدن (khandidan) 'to laugh' has one NP argument while دیدن (didan) 'to see' has two NP arguments regardless of their grammatical or semantic relations to the verb. Syntactic arguments can be used by syntax parsers to resolve the ambiguities.

On the other hand arguments may be defined at grammatical level showing grammatical roles such as subject and object of a verb. In the above example the verb 'to see' has two grammatical arguments, a subject and an object. These argument structures may be used by dependency parsers for disambiguation. We don't consider this level in our work.

The third level is semantics. Semantic arguments known as semantic roles, thematic roles or Θ-roles (theta roles) are used for semantic processing of texts. The verb 'to see' has two thematic roles; agent and theme as semantic arguments.

By these considerations, we define the argument structure or the frame of a verb in two levels: syntactic and semantic.

Syntactic tags include NP, VP, PP, Sentence, ….

For more than half a century, linguists have been trying to come up with a neat comprehensive set of universal semantic roles; however, there has not been a general agreement regarding the inventory of them yet. In this paper we use the role list proposed by Ghazanfari (2014). She has modified the list of Brinton & Brinton (2010) in order to fit the requirements to be used in different wordnets and especially to be applied in a convincing manner in FarsNet. Her list consists of the following roles [Ghazanfari, 2014] (in each case the role holder is shown in *italic*):

1. Agent: the human initiator, causer, doer or instigator of an action who acts by will or volition. The *logger* felled the tree. The tree was felled by the *logger*.
2. Actor: the animate entity who or which acts or causes an action. The *boy* broke the window accidentally. The *dog* barks.
3. Force: the inanimate cause of an action and its direct cause. The *wind* felled the tree. The window was broken by the *wind*.
4. Instrument: the means by which an event is caused or the tool generally inanimate used to carry out an action. The tree was felled with an *axe*. He used an axe to fell the tree.
5. Stimulus: The entity which causes a kind of psychological effect in another entity, the experiencer. The *noise* frightened the students.
6. Experiencer: the animate being affected inwardly by a state or action. *Mina* feels lonely. *I* like apple. The noise frightened the *students*. The news is pleasing to *me*.
7. Source: the place-from-which or person-from-whom an action emanates. I got the book from the *library/* my *friend*.
8. Goal: the place-to-which an action is directed, including indirect objects and directional adverbs. She reached the *coast*.
9. Recipient: an animate or some kind of quasi-animate entity, the person who gets or receives something. My *mother* was sent a gift. A new idea came to *me*. Daniel wrote a letter to the *bank*.

10. Path: the path taken in moving from one place to another in the course of an action. Hannibal travelled *over the mountains*. The package came *via Tehran*.
11. Location: the place-at/in-which an action occurs. The cat is in the *room*/ under the *table*. The *room* has many people in it.
12. Temporal: the time at which something happens or an action occurs. I will call on *Tuesday*/ at *noon*.
13. Possessor: the possessor of a thing, *He* has/ owns/possesses a house. The bag belongs to *minoo*.
14. Benefactive: the person or thing for which an action is performed or the person derives something from the actions of another. He ordered the book for *me*.
15. Patient: the person or thing affected by an action or the entity undergoing a change. I baked the *chicken*. He ate the *cake*.
16. Theme: The person or thing which undergoes an action or that which is transferred or moved by an event otherwise unchanged. I put the *book* on the table. The *paper* flew out of the window.
17. Neutral: The person or thing which is not changed or even acted upon but is simply present at an action. The *house* costs a lot. The *table* measures three feet by three feet.
18. Range: The specification or limitation of an action. The dress costs *a hundred dollars*. We drove *ten miles*.
19. Role: a person playing a role or part in an action or state. We made *Lise* treasurer of the club. *Hilda* is the principal of the school.
20. Associate: the entity having an equal status (role) with another argument in the sentence. They made Reza *the head of department*. She calls her doll *Juju*.
21. Reason: This refers to the reason or purpose for which an action takes place. Robin called the police for *help*. She returned to class to *take her book*.
22. Accompaniment: the entity which participates in close connection with the agent, actor, force, patient or theme but has a secondary role in the event. I went to the movies with *my friends*.
23. Manner: the qualification of an event, the way in which an action is performed or an event takes place. He lived out his life happily. Tom left in a hurry.

To extract the argument structures of verbs and the selectional restrictions of arguments, we used a corpus driven approach. For this reason we developed a corpus management system called Samp. First we tagged the arguments of various occurrences of the candidate verbs in the corpus by both syntactic and semantic roles. Then using the developed tool the argument structure and also the selectional restrictions are concluded semi-automatically and confirmed by linguist before adding to FarsNet.

Next sections discuss the corpus management system and the process of extracting the argument structures for FarsNet in more details.

## 4 The Corpus Management System

To extract the verb argument structures we developed a corpus management system (CMS) called Samp [Shahriyari, et al., 2014]. Samp like other corpus management systems (such as BNCweb) is able to receive a corpus as input, search in it and find and show all occurrences of a word along with its surrounding words in the corpus and prepare various types of reports about it.

Besides the above ordinary capabilities of a corpus management system, Samp has the following features:

- Samp accepts any Persian corpus, and changes its format to the desirable standard.
- Samp is a web based system capable of handling multiple synchronous users enabling cooperative corpus tagging. It creates a log of users' activities over the net.
- Samp is able to tokenize a raw corpus and tag it by POS categories either automatically or help to tag manually.
- Samp helps users to tag the corpus by senses provided by FarsNet or user. In fact, Samp provides a cooperative environment to let users tag the corpus semantically by FarsNet senses or by new user defined senses.
- Samp is able to search for a word and all of its inflections, derivations and also multi part words in which the search keyword is involved. For each search the word within its surrounding context is returned. The size of the surrounding window can be determined by user.
- Samp helps users to tag sentences by their verb's syntactic and thematic arguments.
- Samp helps the linguist to extract the verb frames and determine the selectional re-

strictions of arguments. Actually, it recommends the verb frames by summarization and generalization (mining) of tags users created for verbs and their arguments and let the linguist to confirm or correct it (more details in the next subsection).

### 4.1 Extraction of verbs' argument structures

## Tagging the corpus

Tagging the corpus by senses and arguments of verbs has the following steps:

1- User enters the corpus to be tagged.
2- Samp reformats the corpus into its standard and makes it ready to be tagged.
3- User enters the word (verb) into the search pane.
4- Samp provides the list of sentences (evidences) in which the word (verb) or its inflections or its stem or its derivations are present by applying morphological analysis.
5- For the sentences in the list Samp asks the user to tag the verb by its meaning. It shows the list of senses provided by FarsNet. User can select the appropriate sense or add a new sense. User defined senses will be then evaluated to be added to FarsNet if necessary. This way we can complete the missing senses of FarsNet while tagging the corpus. Currently this task is performed manually. We are going to use WSD algorithms to tag word senses automatically in the future.
6- In the selected sentence, according to the determined sense, the arguments of the verb are found and tagged by syntactic (NP, PP, …) and semantic roles (Agent, Patient, …). More details are discussed in the next subsection.
7- Samp saves the tags and repeats steps 5 and 6 to complete the task for a verb.

After completing tagging the corpus, it's time to make a conclusion on the tags and extract the argument structure of a verb and the arguments' selectional restrictions. This task is discussed in following section.

**Determining Syntactic and Semantic Arguments**

For each evidence (sentence in which the desired verb is occurred) the verb arguments should be extracted. Then for each argument it is determined if it is obligatory or optional. Also the arguments are tagged by their selectional restrictions which show the properties of the filler of each argument slot.

For example suppose the verb بردن (bordan).
One of its meanings (senses) is 'to win' and the other one is 'to take'. For the first sense we may tag the following sentence in the corpus as follows:
Sentence: Iranian films won some prizes in the festival.
Force= Iranian Films and theme=prize
And for the second sense the following is an example.
Sentence: he took Reza from home to school at noon.
Agent = he, theme=Reza, source= home, goal= school, temporal= at noon.
As an instance the selectional restriction of the theme argument of this verb is 'to be portable".
Extracting syntactic and semantic arguments can be done in two modes; manual or semiautomatic.

In the manual mode (which is the main focus of this paper) Samp provides the environment for user to tag arguments and select their selectional restrictions in each sentence. The restrictions are recommended to the user by upward traversing the inclusion hierarchy of FarsNet from the argument node (finding its ancestors).

For semiautomatic mode we used a syntax parser to extract syntactic arguments and a semantic role labeler (SRL) (Jafarinejad & Shamsfard, 2012) to extract semantic arguments of the verb.

**Concluding the Structure**

In this part, the final argument structure and the most general selectional restrictions for its components are determined by Samp automatically. In the concluding subsystem, for each sense of a verb, Samp shows user a list of all of its assigned arguments in all sentences (evidences) with their selectional restrictions. This list shows the frequency of cooccurrency of each argument with the corresponding verb sense. It also shows the number of times each argument for a specific sense has been obligatory or optional.

According to this report Samp can suggest the final argument structure of a verb to be confirmed or corrected by user. This structure is built by getting union among all argument sets of the verb sense in all the evidences. In this task similar or identical sets are recognized and merged and different sets whose frequency of

occurrence is more than a threshold are added to the union set.

To determine the selectional restrictions, Samp finds the most general class among various classes introduced as the restriction of the arguments which are being merged.

For example suppose the verb خوردن (khordan). It is a polysemous verb for which 'to eat' and 'to hit' are two of possible meanings (senses). For the first sense we may tag the following sentences in the corpus as follows:

S1: To become healthy one should eat an apple a day.

S2: Babies eats milk as the main course before the age of 6 months.

S3: eating breakfast is important in having a successful day.

In S1 eat needs agent and patient as obligatory arguments and temporal (time) and reason as optional ones (adjuncts). In this sentence the selectional restriction of patient is being apple or its superclass: 'fruit'. Similarly in S2 the patient is milk and its selectional restriction can be 'drinks'. And in S3 the selectional restriction of the patient (breakfast) is meal.

In other words the patient of khordan in the meaning of 'to eat' may be a fruit, a dink[1] or a meal. Samp can infer from these evidences besides other sentences for this sense of khordan that the patient of 'khordan' may be an 'edible'.

It also concludes that 'khordan' ('to eat') has obligatory agent and patient and may have optional temporal, associate and reason.

In some cases more than one argument structure may be inferred for a unique sense of a verb. This may happen for one of the following reasons:

1- The argument sets may not be merged. For example for a unique sense, we may have agent and patient in some sentences and force and patient in some other sentences. In this case we may merge agent and force in a broader class as undergoer or keep the original structures and so have more than one legal argument structure.
2- The differences of two sets are in the obligatory arguments and have never co-occurred in the sentences. For example suppose a verb with agent, patient and source in some sentences and with agent and goal in some others but the patient and goal has never co-occurred for this verb in the corpus. In this case the two structures are kept separately to ask the user to see if they should be merged or not.
3- In case of having more than one argument set for a verb sense, the user may decide to split the sense into two more specific senses or add the argument sets 'as is' into FarsNet.

The final concluded argument structure is represented in a specific language and added to FarsNet

A sample of data added to FarsNet is following. (for verb bordan meaning to take). Anything within parenthesis is optional.

**Syntactic arguments:** NP&NP&(PP)&(PP)
(it means that the verb has 4 syntactic arguments , two obligatory noun phrases and two optional prepositional phrases)

**Thematic Roles :** Agent&Theme&(Source)&(Goal)
It means that the verb has 4 thematic arguments an obligatory agent, an obligatory theme and optional source and goal.

**Relations :**
NP&Agent
/NP&Theme/
(PP)&(Source)
(PP) & (Goal)

This shows the correspondence between the syntactic and the thematic arguments.

## 5 Results and Discussion

In this paper we talked about some new features developed in FarsNet 2.0. Table 2 shows the last statistics for FarsNet 2.0.

Table 2-some statistics on FarsNet 2.0

| | Noun | Adj. | Adv. | Verb | Total |
|---|---|---|---|---|---|
| Word | 16008 | 6560 | 2014 | 5679 | 30261 |
| senses | 19773 | 6904 | 2023 | 7438 | 36138 |
| Synset | 10954 | 4261 | 923 | 3266 | 19403 |
| Sense relation | 3096 | 345 | 22 | 3585 | 7048 |
| Synset relation | 31333 | 6733 | 1100 | 5492 | 36749 |
| Mapped synsets | 10108 | 4518 | 929 | 3023 | 18576 |

Besides extending the Persian WordNet we have had some studies (corpus based) on verbs.

In this study we selected 187 simple distinct Persian verbs. For these verbs, we extracted about 4118 distinct evidence sentences from the corpus and tagged them by the meaning (sense) of verb

[1] In Persian, It is usual to use the verb 'to eat' for drinks instead of 'to drink'

and its arguments. From these sentences we extracted 847 sets of verb-sense-argument structure which are all entered into FarsNet 2.0. In other words we completed the information of 187 verbs in FarsNet with their verb frames. considering that each verb has some senses and each sense may have more than one frame we entered 847 verb frames with their selectional restrictions into FarsNet.

To extract the arguments we considered the valency of verbs too. Valency refers to the capacity of a verb to take a specific number and type of arguments. Our study showed that there is no zero-valence verb in Persian. The statistics of the studied 190 simple verbs regarding their valence is shown in table 3.figure 1 is about the frequency of arguments in the test data.

Table 3-statistics on Persian simple verbs regarding their valence in the test set

| type | Percentage |
|---|---|
| 0-valence | %0 |
| 1-Valence | %17 |
| 2-Valence | %60 |
| 3-Valence | %23 |
| 4-valence | %0 |

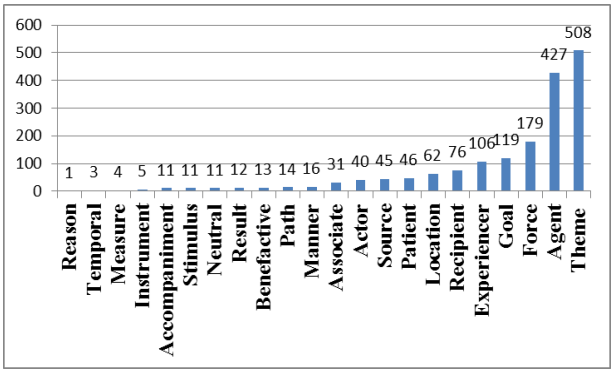


**Figure 1- frequency of the arguments in the selected set**

Enhancing the automatic part of our work especially in applying WSD algorithms to find the verb sense, SRL methods to extract semantic roles and the reasoning (concluding) part of extracting the argument structures besides using the extracted data in real world applications are among our further works.

## References


Bijankhan, M. (2000). Bijankhan Corpus, http://ece.ut.ac.ir/dbrg/Bijankhan.

Brinton, L. J. and D. M. Brinton. (2010). The Linguistic Structure of Modern English. Amsterdam and Philadelphia: John Benjamins publishing Company.

Dixon, R.M.W. (2000). A Typology of Causatives: Form, Syntax, and Meaning. In Dixon, R.M.W. & Aikhenvald, Alexendra Y. Changing Valency: Case Studies in Transitivity. Cambridge University Press.

Fellbaum, C. (ed.) 1998. WordNet: An Electronic Lexical Database. MIT Press.

Ghazanfari, Y. (2014) A survey on semantic roles for inclusion in Persian WordNet, International Journal of Language Learning and Applied Linguistics World (IJLLALW) Volume 6 (2), June 2014; 150-158

Jafarinejad, F.,Shamsfard M.,(2012) Extracting Generalized Semantic Roles from Corpus, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006) Extending VerbNet with Novel Verb Classes. Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy. June, 2006.

Scott, M. (2012). WordSmith Tools version 6, Liverpool: Lexical Analysis Software.

Shahriyarifard, A., Sharifzadeh A., Shamsfard M.,(2014), Introducing Samp, the corpus Management system, 3rd Iranian computational linguistics, Iran.

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., et al. (2010). Semi Automatic Development of Farsnet; the Persian Wordnet. Proceedings of 5th Global WordNet Conference (GWA2010). Mumbai, India.

Dan Tufis. 2004. Balkanet: Aims, Methods, Results and perspectives. Romanian journal of Information Science and Technology. V7, pp.9-43.