

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/228529533>

Semi Automatic Development Of FarsNet: The Persian Wordnet

Article · January 2010

CITATIONS

24

READS

698

6 authors, including:



Mehnoush Shamsfard

Shahid Beheshti University

73 PUBLICATIONS 649 CITATIONS

SEE PROFILE



Akbar Hesabi

University of Isfahan

6 PUBLICATIONS 32 CITATIONS

SEE PROFILE



Niloofer Mansoory

Payame Noor University

4 PUBLICATIONS 28 CITATIONS

SEE PROFILE



Ali Famian

Payame Noor University

7 PUBLICATIONS 30 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Farsnet, A Wordnet for Persian [View project](#)



Linguistic landscape in Iran [View project](#)

Semi Automatic Development of FarsNet; The Persian WordNet

Mehrnoush Shamsfard

Computer Engineering
Dept., Shahid Beheshti
University, Tehran, Iran
m-shams@sbu.ac.ir

Akbar Hesabi

Linguistic Dept.,
Allameh Tabatabaiee
University, Tehran, Iran
a.hesabi11@yahoo.com

Hakimeh Fadaei

Computer Engineering.
Dept., Shahid Beheshti
University, Tehran, Iran
Shafagh4@yahoo.com

Niloofer Mansoori

Payam Noor University
Iran
nmansoori@gmail.com

Ali Famian

Payam Noor University
Iran
famianali@yahoo.com

Somayeh Bagherbeigi

Allameh Tabatabaiee
University, Tehran, Iran
sb_4715@yahoo.com

Elham Fekri

Shahid Beheshti University,
Tehran, Iran
elham.fekri
@gmail.com

Maliheh Monshizadeh

Shahid Beheshti University
Tehran, Iran
monshizadeh
@ce.sharif.edu

S. Mostafa Assi

Institute for Humanities and
Cultural Studies,
Tehran, Iran.
s_m_assi@ihcs.ac.ir

Abstract

This paper describes the development process of FarsNet; a lexical ontology for the Persian language. FarsNet is designed to contain a Persian WordNet with about 10000 synsets in its first phase and grow to cover verbs' argument structures and their selectional restrictions in its second phase. In this paper we discuss the semi-automatic approach to create the first phase: the Persian WordNet.

1 Introduction

WordNet (Miller 1995, Fellbaum 1998) is an electronic lexical database originally designed for English and replicated in several other languages. WordNet organizes words into sets of cognitively synonymous sets, called synonym sets or synsets. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. Actually, each synset represents a distinct concept, which can be expressed by its members in a range of different contexts. Synsets are interrelated by means of lexical (word-to-word) and conceptual-semantic (synset-to-synset) relations. The relations may relate words within a POS category (such as Synonymy, Antonymy, Hyponymy, Meronymy) or between different categories (such as Attributes and Derivationally related form.).

WordNet presently contains approximately 155287 different word forms organized into some 82115 word meanings synsets.

Nowadays WordNet is developed for more than 40 languages around the world. EuroWordNet, BalkaNet, AsiaNet and WordNets for Dutch, Italian, Spanish, German, French, Czech and Estonian are among them. Unfortunately some languages such as Persian (Farsi) lack such a semantic resource for use in NLP works.

There have been some efforts to create a wordnet for the Persian language too (Keyvan, et al, 2007; Mansoori & Bijankhan, 2008; Famian & Aghajani, 2007) but no available product have been announced yet. The only available lexical resources for Persian are some lexicons containing phonological and syntactic knowledge of words (such as Eslami, et al. 2004). This paper describes the semi automatic construction of FarsNet 1.0; the first WordNet for the Persian Language (Shamsfard, 2008). In this paper after an introduction to Persian Wordnet and its features and general construction methodology and resources, we will describe the language specific issues and our semi-automatic methods for building the WordNet.

2 Persian WordNet

FarsNet is designed to contain a Persian WordNet with about 10000 synsets in its first phase and grow to cover verbs' argument structures and their selectional restrictions in its second phase. FarsNet has also inter-lingual relations connecting Persian synsets to English ones (in Princeton WordNet 3.0). Persian WordNet goes closely in the lines and principles of Princeton WordNet,

EuroWordNet and BalkaNet to maximize its compatibility to these WordNets and to be connected to the other WordNets in the world to enable cross lingual tasks such as MT, multilingual IR and developing multilingual dictionaries and thesauri.

In this section after a brief description of Persian Language, we talk about the general methodology and lexical resources we used for FarsNet 1.0.

2.1 The Persian Language

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 millions speakers. As far as the lexicon is concerned, Persian has borrowed many loanwords mostly from Arabic, English, French, and the Turkish languages. The Arabic influence has been great in both number of borrowed words and their frequency in use. Syntactically, Persian is primarily an SOV language with many ordering exceptions which makes it almost a free word order language. We will mention some specific features of Persian language for different parts of speech in the following sections.

2.2 Features and Coverage

FarsNet 1.0 is going to include the lexical, syntactic and semantic knowledge about more than 15000 Persian words and phrases organized in about 10000 synsets of nouns, adjectives and verbs. The size of this WordNet is more than some small scaled wordnets such as Hebrew and less than large scaled ones such as Princeton or EuroWordNet. According to size, it falls in the same category as medium scaled WordNets like the Arabic one.

- *Concepts*

The base concepts covered in FarsNet are classified into two groups: (a) Language independent base concepts which are those counted as base or important in many languages (b) Persian base concepts which are the most frequent words or common or important concepts among Persian speakers.

From the first group, FarsNet covers most of the base concepts BCS1 and BCS2 of BalkaNet to achieve compatibility with other WordNets. The base concepts of European languages which do not have equivalent or are not common in Persian are eliminated from this set.

The second group is extracted from the most frequent words of two Persian corpora: Peykareh

(Bijankhan, 2004) and PLDB (Assi, 1997) to preserve the Persian specific structures. Many of the members of this group are present in the first group as well.

- *Relations*

There are two main types of relations defined in FarsNet: inner language and inter-language relations. As FarsNet is mapped to WordNet 3.0 there are two inter-language relations; equal-to and near-equal-to between FarsNet and WordNet synsets. Inner language relations which are held between different senses and synsets of FarsNet are the same as the relations in WordNet 2.1. They include synonymy, hypernymy and hyponymy, different types of meronymy, Antonymy and cause. FarsNet 1.0 does not cover inter-POS relations. So the domain and range of all relations are from the same POS currently.

2.3 General Methodology

The fact that already there are no taxonomies or ontologies that provide a formalized description of concepts of Persian language and also the lack of machine readable Persian (mono- or bi-lingual) dictionaries led us to choose and adopt the expand strategy in constructing the Persian synsets and hierarchy. Following the methods applied for creating EuroWordNet (Vossen, 1998), at the first step we manually develop the core WordNet of Persian concepts and then we will semi-automatically expand it by adopting a top-down process.

In other words, our general methodology consists of three major steps: (1) Construction of a Core-Wordnet for a set of common base concepts (2) enrichment of this set providing relational links and incorporating their direct semantic contexts and (3) top-down extension of this core-WordNet by new concepts and relations. To construct the core concepts of FarsNet we have translated the BalkaNet concepts sets BCS1 and BCS2 (Tufis, 2004). Also the most frequent and language specific concepts will be added in the next phases using electronic Persian corpora. Adding hyperonyms and the first level hyponyms to these Base Concepts will result to the core WordNet of Persian. This core WordNet has to be expanded (semi) automatically using specifically available resources, e.g. monolingual and bilingual dictionaries, lexicons, ontologies, thesauri, corpora, etc.

This approach maximizes compatibility across WordNets and at the same time preserves the language specific structures of Persian.

2.4. Lexical Resources

There are some lexical resources we use to construct the Persian WordNet. Some of them are paper copies which are used in manual creation of the lexicon and some others are electronic resources which are used in employed automatic methods as well. Our main lexical resources are as follows.

- monolingual Dictionaries
 - Sokhan dictionary (Anvari, 2004) is our main reference for lexical definitions, examples, and other lexical information. This monolingual dictionary in eight volumes (available in paper format only) is widely acknowledged as the most complete, reliable Persian dictionary.
 - Sadri Afshar Dictionary (Sadri Afshar, et al. 2008) with 50000 entries is another resource to ensure the satisfying coverage.
- Corpora
 - Persian Linguistic Database (PLDB)¹, (Assi, 1997) is an on-line database for the contemporary (Modern) Persian. The database contains more than 50 million words of all varieties of the Modern Persian language in the form of running texts. Some of the texts are annotated with grammatical, pronunciation and lemmatization tags. PLDB has been our main resource for determining words' frequency.
 - Peykareh (Bijankhan, 2004): is a collection gathered from Ettela'at and Hamshahri newspapers of the years 1999 and 2000, dissertations, books, magazines and weblogs. Written and spoken texts were collected randomly from 68 different subjects in order to cover varieties of lexical and grammatical structures. The version of Peykareh (also known as Bijankhan corpus) which we use contains about 10 millions manually tagged words with a tag set that contains 109 Persian POS tags. A subset of this collection tagged with smaller tag set is also prepared and distributed by DBRG at Tehran University.
- Bilingual Dictionaries
 - English-Persian Farhange Mo'aser Dictionary (Batani, 1992)
 - English-Persian Millennume Dictionary (Haghshenas, 2007)
 - Aryanpour (2008) English-Persian electronic Dictionary containing more than 200000 entries.

- Others
 - Khodaparasti (1997) is a dictionary of Persian synonyms and antonyms (in paper format).
 - Persian Thesaurus (Fararoyy 2008) is an electronic thesaurus used for automatic mapping between Persian and English synsets.

3 Nouns

To make the noun part we move from two sides as well; translating English synsets and choosing Persian concepts. For the first side, we manually translate the PWN Synsets of selected base concepts, using our linguistic knowledge of English and Persian and English-Persian Farhange Mo'aser Dictionary (Batani, 1992) and then enter the Persian equivalences in our WordNet editor. Then we refer to Anvari (2004), to check out the consistency and correctness of our equivalences. For the second side, we use the most frequent nouns from Persian Linguistic Database (PLDB) and Bijankhan corpus, and add the nouns which are missed in the first side to the lexicon. In the next step, the suggestions are defined, some glosses and examples are added and some relations are held between synsets.

To complete the relations between nouns or their synsets and also doing the mapping between Persian and English synsets (especially for the cases in which we move from Persian side) we use semi-automatic methods to suggest new relations and mappings.

3.1 Automatic Mapping Suggestion Between English and Persian Words and Synsets

Our goal is finding the most appropriate mapping between Persian words (or synsets) and English synsets. We used the adaptation of Farreres' approach [Farreres, 2005] for Persian [Dehkharghani & Shamsfard, 2009] which needs bilingual Persian-English and English-Persian dictionaries, monolingual Persian-Persian dictionary, Persian thesaurus and English WordNet as resources. The results show 72% precision in mapping Persian words to English synsets and 69% precision in mapping Persian synsets to English synsets

3.2 Semi-automatic Extraction of Conceptual Relations

Conceptual relations are classified in two categories: Taxonomic and Non-taxonomic relations, both of which could be learned by our relation learning system. These relations are extracted from either raw or tagged texts of two sources:

¹ <http://www.pldb.ihcs.ac.ir>

Bijankhan corpus and Wikipedia articles using the following approaches.

3.2.1. Pattern based approach

We exploit pattern based approaches to extract both taxonomic and non-taxonomic relations from Persian texts.

To extract taxonomic relations we define a set of 24 patterns containing the adaptation of Hearst patterns [Hearst, 1992] for Persian and some other new patterns. We have also extracted some patterns for some well known non-taxonomic relations such as "Part of", "Has part", "Member of" and "synonymy". The translations of some of these patterns are shown in table 1 (TW stands for target word).

Table1: Some patterns for extracting relations

	Pattern	Relation
1	TW is (a) X.	Hypernymy
2	TW is considered as X	Hypernymy
3	TW is known as X	Hypernymy
4	TW is called X	Hypernymy
5	TW is named as X	Hypernymy
6	TW is a part of X	Part of
7	TW includes X	Has part
8	TW means X	Definition
9	TW is defined as X	Definition
10	TW1 or TW2 or ... are	Synonymy
11	TW has X	Has

Pattern-based approaches in extracting relations are usually of high precision but low recall in a corpus. Searching Wikipedia articles thus is much simpler as Wikipedia articles are high informative and in the first section of these articles we can usually find some occurrences of our patterns. To start the pattern matching phase we extracted the 1000 most frequent Persian Nouns and extracted the Wikipedia articles related to these words. For each word the related article is searched for the phrases matching any of the patterns. The translations of some of the extracted relations by this method are mentioned in table2. It should be mentioned that searching the patterns need some text processing tools (e.g. chunker) to find the constituents of sentences. While there is no efficient chunker for Persian, we did some post-processing to eliminate incorrectly extracted relations. This phase includes eliminating the stop words, applying some heuristics such as matching the head of the first noun phrase in the sentence with the head of the extracted TW in copular sentences, eliminating

prepositional phrases for taxonomic relations, replacing long phrases with their heads and so on.

Table2: Some extracted relations

1	Isa (blood, liquid)
2	Isa (newspaper, publication)
3	Isa (prison, location)
4	Isa (heart, organ)
5	Isa (pen, tool)
6	Isa (representative, person)
7	Isa (organization, collection)
8	Has part (Tehran, Tajrish)
9	Has (Greece, history)
10	Synonym (thought, idea)

3.2.2. Structure based approach

In Wikipedia pages Structures such as tables, bullets and hyperlinks are informative pieces of text. For example in many Wikipedia documents we can find some information given via bullets. This information usually shows some taxonomic relations. In these cases the title of the section which only contains bulleted text is considered as the domain of the relation and each bullet forms the range of the relation.

Hyperlinks are other sources of information. In the whole document each important word which has an article in Wikipedia is linked to its related article. These linked words especially the ones locating in the first section of the text are usually related to the title of the document. We use this fact to extract some taxonomic and non-taxonomic relations. That means for a given word we search the first section of its related article in Wikipedia and extract the linked words. These linked words are usually related to the original word but to reduce the error rate, for each linked word in this section we search its related article to see if we can find a link to the article of original word or not. If such a link exists it means that these two documents are inter-related and there is a high probability that the two words are related.

This method is used over the articles of 1000 most frequent nouns. The types of the extracted relations are not known in this method but they can be mostly found by using some extra searches on the web. Some learned relations from this approach are shown in table 3.

Table 3: Extracted relations from Wikipedia structures

	First word	Related word
1	fire	flame
2	water	earth
3	professor	university
4	marriage	wife
5	face	human
6	path	street
7	tree	wood
8	North sea	sea
9	heart attack	disease
10	Brighton	city in England

3.2.3. Statistical approach

Statistical methods are widely used in extracting relations in many systems. In this system we use this approach to extract co-occurrence relations. To extract these relations, for each pair of words within the 500 most frequent nouns of Persian we searched a 100,000 word subset of Bijankhan corpus to find in how many sentences these two words co-occur. If this number is above a certain threshold, these two words are considered as co-occurrent. Experimental investigations show that 19 would be a proper threshold in this method. Some of these relations and their frequencies are shown in table4.

Table 4: Statistically extracted related words

	First word	Second word	Freq.
1	interest	sale	19
2	price	merchandise	23
3	market	Price	29
4	manufacture	product	36
5	manufacture	merchandise	70
6	stock	merchandise	142

4 Adjectives

In Persian, adjectives are either simple or are formed through adding a number of affixes to other lexical categories, especially nouns. These affixes range from highly productive [-i:irani (Iranian)], to fairly productive [-mand: servatmand (rich)], and nonproductive [-nâk: dardnâk (painful)] .

In building the adjective part of FarsNet, we use three Persian dictionaries (Sokhan, Sadri Afshar and Khodaparasti) and the PLDB database as our main lexical resources.

Following a modified version of GermaNet adjective classification², we have organized Persian

adjectives in 12 main semantic categories and 57 sub-classes. At this stage, just the following 12 general categories have been implemented. For each category some examples are provided.

- 1- Conceptual: torS (sour); rošan (light); narm (smooth)
- 2- Temporal: dir (late); râyej (popular)
- 3- Spatial: nazdik (near); čap (left); xâli (empty)
- 4- Movement moteharrek: (moving); sâken(fixed); sâbet (stable)
- 5- Material: felezi (metal); sangin (heavy); garm (warm)
- 6- Body: gorosne (hungry); bimâr (ill); mo-zakkar (male)
- 7- Emotion: xošhâl (happy); qamgin (sad); afsorde (depressed)
- 8- Intelligence: bâhuš (intelligent); ahmaq (dull); âgâh (aware)
- 9- Behavior: (tanbal: lazy), dust (friendly); mâher (skilled)
- 10- Social: (melli: national); servatmand (rich); xosusi (private)
- 11- Quantity: do(two); kam (few,little); arzân (cheap)
- 12- Relational (saxt:difficult); sâlem (safe); mohem (important)

About 1500 Persian adjective synsets have been put to the system to date. This covers about 1800 adjective word forms.

4.1 Semi-automatic extraction of adjectives and their features

In semi-automatic development of adjective part we extract the antonymy relations by applying morphological rules and testing on a corpus.

On the other hand we automatically cluster adjectives. The goal is to put adjectives that are defining different degrees of the same attribute in one cluster. For example words {hot, warm, cool, cold, chilly} describe temperature attribute with different intensity, and so they must be put into the same class. To cluster adjectives we compute dissimilarity between them. Our system employs known linguistic and statistical methods for adjective clustering. In linguistic side we use a pattern based approach and search for co-occurring adjectives in noun phrases. If two adjectives are co-occurring in an Ezafe-construction, they may not be in a cluster while if they occur in a positive or negative conjunction they probably belong to a cluster. For example, adjectives "سرد" [sard, cold] and "گرم" [garm, hot] which belong to one cluster, usually cannot be used in one

² <http://www.sfs.uni-tuebingen.de/lsd/>

Ezafe-construction ("آب سرد گرم" [äb - e sard - e garm: cold hot water]) because one thing cannot be hot and cold at the same time. While they can occur in a conjunction such as ("نه سرد و نه گرم" [na sard va na garm: neither cold nor hot]).

On statistical side we assume that similar adjectives appear with common set of nouns. Suppose that frequency of occurrence of adjective i with noun j is F_{ij} . For each two adjective, A and B and nouns X and Y If $F_{Ax} < F_{Ay}$ and $F_{Bx} < F_{By}$, or, $F_{Ax} > F_{Ay}$ and $F_{Bx} > F_{By}$ the two adjectives are concordant and otherwise they are discordant.

Similarity is define as: $\text{Similarity} = P_c - P_d$, where P_c is the probability of being concordant, and P_d is the probability of discordance, so it's range is between -1 (dissimilar) and 1 (similar).

Then we cluster adjectives according to their dissimilarity value by minimizing the following objective function by hill climbing approach.

$$\varphi(p) = \sum_{i=1}^R [1/|C_i| \sum_{\substack{x,y \in C_i \\ x \neq y}} d(x,y)]$$

In which R shows the number of classes, C_i shows the i th class, $|C_i|$ is the total number of elements in i th class. $d(x, y)$ is dissimilarity parameter calculated for adjectives x and y .

The best results of evaluation for some groups of test data show %54.50 precision, %74 recall and 60.50% F-measure.

5 Verbs

The construction of the verb hierarchy in FarsNet also follows a top-down strategy on a expand methodology to achieve a high level of overlapping between English and Persian, at least in the highest levels of the hierarchy.

In our current project we are linking our verbal synsets with basic relations such as synonymy, hypo/ hypernymy, antonymy and the cause relation. As the hypo/ hypernyms are constructed along with the structure of PWN and its verbal hierarchy, it is clear that in most of the cases there is a one-to-one correspondence between the two languages. But regarding the antonymy and cause relations there are some language specific features which affect the structure of Persian WordNet of verbs.

Beside the manual process, we have used some semi-automatic techniques for extraction of verbal synsets. In the following two sub-sections first we will discuss some specific Persian verbal

features and their effects on determining the semantic relations among verbs and then will explain the semi-automatic method used in synset extraction.

5.1. Particularities of Persian verbal system

One of the significant characteristic of the Persian verbal system is its small number of simple verbs. Actually most of the verbal concepts are expressed by compound verbs in this Language. The syntactic and semantic features of Persian compound verbs have been the subject of interest for many linguists and some authors (Mansoori & Bijankhan 2008; Rouhizade, et al. 2008) have discussed the issue in WordNet framework or actually from relational semantic perspective.

Each compound verb in Persian is the combination of a nonverbal element and a light verb. The non-verbal elements which come before the light verb and in this sense are called the preverbal elements range over a number of lexical and phrasal categories such as noun, adjective, adverb and prepositional phrase.

For the construction of WordNet, this morphological information can help the lexicographer to determine some semantic relations among the compound verbs and also to predict the relations that may connect two verbal synsets.

In determining the antonymy relations among the verbal synsets we found that in most of the cases when the verbs are compound and their preverbal elements are adjectives and nouns, the existence of the antonymy between the two adjectives or nouns will lead us to connect the two verbs with the same lexical relations (e.g. *dorugh goftæn* (to lie) vs. *rast goftæn* (to tell the truth)).

The other interesting issue about the Persian compound verbs is the cause relation among them. Like English, Persian has lexicalized causative pairs but in contrast with English, the number of Persian causative pairs is very high. This fact results from a morpho-semantic pattern among the Persian simple and Compound verbs. Regarding the simple verbs, Persian has the suffix "*-andæn*" which can be replaced with the infinitive maker suffix "*-idæn*" and change a intransitive, anticausative verb to a simple transitive, causative one. The pair *lærzidæn/lærzandæn* (shake/ shake) is of these kinds. Referring to PWN you will find no cause relation between the two senses of the first verb (shake). Actually these two meanings are fused in one synset and the definition "*move or cause to move back and forth*" shows that both causative and anticausative meanings are referred to the same

lexical element and same synset respectively. But regarding the corresponding Persian concepts, because we have two different lexical items we must construct two different synsets and relate the causative one to the other by means of the cause relation.

The other productive pattern in making the causative/anticausative pairs in Persian is the replacement of one light verb with the other in Persian compound verbs. The replacement of *kærdæn* with *shodæn* in *?ævæz kærdæn* (change: cause to change) / *?ævæz shodæn* (change: undergo a change) and also the replacement of *dædæn* with *kærdæn* in *ta?ghir dædæn* (change: cause to change) / *ta?ghir kærdæn* (change: undergo a change) are of these kinds.

One interesting point which causes a clear difference between English verbal synsets and Persian one with respect to cause relation is that because in most of the cases in English there is no morphological realization for causation, this semantic relation is ignored and both causal and non-causal meaning are presented with one verb or synset. For example {close1} is defined as "*cease to operate or cause to cease operating*" in WordNet 2.0. So in construction of their equivalent synsets in FarsNet because there are two different lexical entries for both causative and non-causative meanings, we have made two different synsets and linked one to the other by means of the cause relation.

5.2. Semi automatic extraction of compound verb's synsets and relations

According to wide usage of compound verbs in Persian, we present a new methodology to semi-automatic enriching of Persian verbs WordNet by using Persian WordNet of nouns and adjectives.

Generally the semi-automatic extraction of compound verb synsets involves using synsets of their preverbal elements. To achieve this goal first the most important (36) Persian light verbs were selected. Then, the process of constructing verbs was done in two phases: first using noun synsets and second using adjective synsets. As these two categories are the most common preverbal elements, it was proposed that adding the common light verbs to members of each nominal or adjectival synset will result in well constructed verbal synsets.

Some resources containing monolingual and bilingual dictionaries along with a Persian corpus were used to evaluate the validity of the constructed compound verbs.

To test the idea we first add the light verbs to each noun or adjective and count the frequency of occurrence of the created compound verbs in the corpus to find the valid common compound verbs. Then we selected three types of structures to form a synset: (1) same preverbals plus synonym light verbs (2) synonym preverbals plus same light verbs (3) synonym preverbals plus synonym light verbs. The created synsets are evaluated both manually by a lexicographer and automatically by looking at the English–Persian dictionary. If there is an entry for which all elements of this synset are within the translation then the synset is accepted. For example combining synonym light verbs 'kardan', 'nemudan' and 'sakhtan' with the same preverb 'Ashkar' can form a valid verbal synset while the same light verbs combining with 'laneh' do not make a synset. Using this method and adding 6467 nouns (organized in 3625 nominal synsets) to 36 light verbs we have constructed 232812 compound verbs from which 4270 were accepted. These verbs were involved in the synset construction method and 3271 verbal synsets were built and accepted by compounding synonym nouns and same LVs and 2822 synsets from same or synonym nouns plus synonym LVs. The same process was performed on one hundred adjectives which resulted in 180 accepted verbal synsets.

6 The Developed Tools

We have developed two sets of tools for FarsNet. The first set consists of a browser (for users) and an Editor (for lexicographers) to view and edit the content manually. They are developed as both local and web-based applications. The second set contains some tools for automatic extraction of lexical and semantic knowledge from resources and proposing them to lexicographers for confirmation before inserting to the lexicon. FarsNet is both stored in XML files and in a database. In both formats, for each word, its POS category, its different forms of writing (orthography) and its phonetic, syntactic and morphological information are stored. Also different senses of the word with the synsets they occur in, along with their gloss and examples are represented. For each sense and also each synset the relations are stored as well. For each synset, there is also a link to its equivalent or near equivalent synset in WordNet 3.0.

7 Results and Conclusion

In this article we had a review of the on-going project on building a Persian WordNet. The current statistics of FarsNet is shown in table 5. The numbers show the number of items entered to the editor and passed to evaluation phase.

Table 5-Statistics of FarsNet at current position

Category	Words	Synsets	Relations
Noun	8868	4081	8437
Adjectives	1691	1502	231
Verbs	2596	3683	391
Total	13155	9266	9059

To evaluate this WordNet: first, we have to compare the results with 3 reliable bilingual dictionaries; second, some human experts check and evaluate the synsets, third, when completed, we have to use the WordNet in some applications and evaluate the results.

Adding more entries to this lexicon, adding the argument structures and selectional restrictions of Persian verbs, adding inter-POS relations and mapping FarsNet to other general upper ontologies like SUMO are among our further works to complete the project. We are also working on enhancement of our automatic knowledge acquisition methods to enable faster and more reliable ontology construction.

Acknowledgement

This work has been funded in part by Iran Telecommunication Research center (ITRC) under contract no. T/500/19231.

References

- Hassan Anvari. 2004. *Sokhan Comprehensive Dictionary*. Sokhan Publishing Co.
- S. Mostafa Assi. 1997. *Farsi Linguistic Database (FLDB)*. International journal of Lexicography, V10, Euralex Newsletter.
- Mohammad R. Bateni. 1997. *Moasers's - English-Persian Dictionary*. Mazda Pub.
- Mahmood Bijankhan. 2004. *Role of language corpora in writing grammar: introducing a computer software*. Iranian Journal of Linguistics, No. 38.
- Rahim Dehkharghani., Mehrnoush Shamsfard.,2009. Automatic Mapping of a Thesaurus to WordNet, NLPCS workshop, Italy.
- Moharam Eslami, M. Sharifi Atashgah, L. S. Alizadeh, T. Zandi. 2004. *Persian Generative Lexicon*.

The first workshop on Persian language and computer. Tehran, Iran.

- Ali Famian, D. Aghajani. 2007. Towards Building a WordNet for Persian Adjectives. proceedings of the 3rd Global WordNet conference.
- Jamshid Fararooy. 2008. *Thesaurus of Persian Words and Phrases*. Ibex Publishers Inc.
- Javier Farreres. 2005. *Automatic Construction of Wide-Coverage Domain-Independent Lexico-Conceptual Ontologies*. PhD Thesis, Polytechnic University of Catalonia, Barcelona
- Christian Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Ali M. Haghshenas, H. Samei, N. Entekhabi. 2004. *Farhang-e- moaser English Persian Millennium Dictionary* (2Vol.), Sokhan Publishers, Tehran.
- Marti A. Hearst. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*, Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France.
- Farhad Keyvan, H. Borjan, M. Kasheff, C. Fellbaum. 2007. *Developing PersiaNet: the Persian WordNet*. proceedings of the 3rd Global WordNet conference .
- Farajollah Khodaparasti. 1997. *A Comprehensive Dictionary of Persian Synonyms and Antonyms*, Fars Encyclopedia.
- Niloofer Mansoory, Mahmoud Bijankhan. 2008. *The possible effects of Persian Light verb construction on Persian WordNet*. proceedings of the 4th global WordNet conference.
- George A. Miller. 1995. *WordNet: a lexical database for English*. Communications of the ACM archive. Volume 38, Issue 11. Pages: 39 – 41.
- Masoud Rouhizadeh, Mehrnoush Shamsfard, Mahsa A. Yarmohammadi. 2008. *Building a WordNet for Persian Verbs*. In: proceedings of the 4th global WordNet conf.
- Golam H. Sadri Afshar, N. Hakami, N. Hakami. 2008. *Farhang-e Moaser, Contemporary Persian Dictionary*.
- Mehnoush Shamsfard. 2008. *Developing FarsNet: A Lexical Ontology for Persian*, In proceedings of the 4th global WordNet conference.
- Dan Tufis. 2004. *Balkanet: Aims, Methods, Results and perspectives*. Romanian journal of Information Science and Technology. V7, pp.9-43.
- Peik Vossen , (ed). 1998. *EuroWordNet: A Multilingual Database with lexical Semantic Networks*. Kluwer academic Publishers.