International Journal of Lexicography, Vol. 25 No. 1, pp. 50–67
doi:10.1093/ijl/ecr022 Advance access publication 11 August 2011

COMPOUND VERBS IN PERSIAN WORDNET

Niloofar Mansoory: *Payame Noor University, Tehran, Iran (nmansoory@gmail.com)*
Mehrnoush Shamsfard: *Shahid Behesti University, Tehran, Iran (shams@sbu.ac.ir)*
Masoud Rouhizadeh: *Shahid Behesti University, Tehran, Iran (masoud@csee.ogi.edu)*

Abstract

This paper discusses some linguistic issues in developing the Persian WordNet of verbs with a special focus on Persian compound verbs. It begins with describing different types of compounding mechanisms in verbs and the grammatical structure and semantic properties of each type. It then continues with discussing the lexical and conceptual relations between compound verbs in the Persian WordNet and, finally, talks about the way that properties are used in the semi-automatic extraction of compound verbs and their relations from dictionaries and text corpora.

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 (Mahootian 1997). It is the official language of Iran, Afghanistan and Tajikistan with more than one hundred million speakers[1] and also spoken in more than six other countries. As a result, there are millions of Persian written materials such as online pages, newspapers, and books. There is no doubt in the necessity of constructing basic language processing resources and tools for it, like many other less-studied languages. On the other hand, one of the most urgent problems in language technology is the lexical semantics bottleneck, that is the unavailability of domain-independent lexicons with rich semantic information on lexical items. Such lexicons could greatly improve the quality of current applications.

One of the well known semantic lexicons which meet such objectives is WordNet. In 1986, George Miller started the development of this lexical database based on semantic relations. His main goal was to simulate the systematic patterns and relations of the mental lexicon in the database in order to feed the computational linguistics community with a store of lexical knowledge as extensive as human lexical storage. WordNet covers words from four POS

(part-of-speech) categories: nouns, verbs, adjectives, and adverbs. The database is organized around the notion of synset (synonym set) between which semantic relations are expressed (Miller 1995, Fellbaum 1998). A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. Synsets are interrelated by means of lexical (word-to-word) relations (such as Antonymy) and conceptual-semantic (synset-to-synset) relations (such as Hypernymy). 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 forms).

The result is a large lexical network structured around meaning similarity. The English WordNet, also known as Princeton WordNet (PWN[2]; Fellbaum 1998) is now a mature lexical Ontology which is applied efficiently in a variety of NLP tasks such as word sense disambiguation (Agirre and Edmonds 2006), machine translation, question answering, information retrieval (IR) and so on (Morato et al. 2003).

Figure 1 shows some examples of synsets and relations in Princeton WordNet 3.0. It demonstrates a verbal synset including three synonyms; *'help'*, *'aid'* and *'assist'*. This synset has some direct troponyms (three of which are shown in the figure) and a hypernym synset. It is also related to a nominal synset {helper}, by a derivationally related form relation. Each synset has a gloss (description) and some examples to improve the readability and understandability.

- (v) help, assist, aid (give help or assistance; be of service) *"Everyone helped out during the earthquake"; "Can you help me carry this table?"; "She never helps around the house"*
  - *Some direct troponyms*
    - (v) facilitate, ease, alleviate (make easier) *"you could facilitate the process by sharing your knowledge"*
    - (v) serve, attend to, wait on, attend, assist (work for or be a servant to) *"May I serve you?"; "She attends the old lady in the wheelchair"; "Can you wait on our table, please?"; "Is a salesperson assisting you?"; "The minister served the King for many years"*
    - (v) benefact (help as a benefactor) *"The father benefacted his daughter in more ways than she was aware of"*
    - (v) help out (be of help, as in a particular situation of need) *"Can you help out tonight with the dinner guests?"*
    - (v) subserve (be helpful or useful)
  - *direct hypernym*
    - (v) support, back up (give moral or psychological support, aid, or courage to) *"She supported him during the illness"; "Her children always backed her up"*
  - *derivationally related form*
    - (n) helper [Related to: help] (a person who contributes to the fulfillment of a need or furtherance of an effort or purpose) *"my invaluable assistant"; "they hired additional help to finish the work"*

Figure 1: A part of PWN 3.0

Inspired by the success of PWN many languages started to develop their own WordNets, taking PWN as a model. Today, 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[3].

Several researchers, such as Famian and Aghajaney (2006), Keyvan et al. (2006), Rouhizadeh et al. (2007, 2008) and Mansoory and Bijankhan (2008), tried to develop a WordNet for Persian. While all of their works include noteworthy theoretical framework for developing a WordNet for Persian, they generally did not address the practical side of the problem. As a result, they were not successful in developing an *actual* WordNet for the language.

In 2008, a large-scale project on building a lexical ontology for Persian called FarsNet was started. It combined the previous works and utilized its own methodologies in developing a wide coverage semantic lexicon. The first phase of FarsNet contains the Persian WordNet for about 18,000 words and phrases organized in about 10,000 synsets (Shamsfard 2008a and b, Shamsfard et al. 2010).

In this paper, we discuss the development of the Persian WordNet of verbs within the FarsNet project. We describe different types of Persian compound verbs, as well as syntactic and semantic properties of each type. We then talk about the way we address the specific characteristics and behaviors of these types in order to develop a semantic lexicon. Finally, we present our method of using such linguistic properties in the automatic extraction of compound verbs and their relations from large text corpora and dictionaries to enrich the Persian WordNet of verbs.

Following this introductory section, the paper continues with four further sections. Section 2 presents some theoretical consideration about compound verbs and their important features. Section 3 introduces the structure of the Persian WorNet of verbs, designed in accordance with the Persian verbal system, and also discusses the reflection of the particularities of Persian compound verbs in a relational semantic framework. In Section 4, we introduce some consequences of the above discussions to be applied to extraction of compound verb synsets and relations in the Persian WordNet. Finally, some conclusions are drawn from the ideas presented in the paper.

2. Compound verbs in Persian

Persian verbs can be divided into two major morphological categories: simple and compound verbs. The number of compound verbs is much larger than simple verbs. According to Mohammad and Karimi (1992), the maximum number of simple verbs in today's Persian is 115, while Dabir-Moghaddam (1997) registered 2500-3000 compound verbs. Some other researchers such as Megerdoomian (2002) listed even more compound verbs.

Persian compound verbs (also known as complex verbs or light verb constructions) are combinations of a non-verbal element and a verbal (light) verb. The non-verbal elements that precede the light verb, called the preverbal elements, range over a number of lexical and phrasal categories such as noun, adjective, adverb and prepositional phrase.

One of the highly discussed issues in the literature about compound verbs in Persian is the dual nature of these constructions as lexical and syntactic elements. As these verbs have a single word stress and also undergo adjective formation and nominalization processes, some linguists have suggested that they are lexical units. On the other hand, the fact that the verbal and non-verbal elements in these constructions can be separated by adverbs, negation and other inflectional affixes or by auxiliaries, has led some researchers to consider them to be phrasal categories. A detailed study of lexical and phrasal features of Persian compound verbs can be found in Megerdoomian (2002). Regarding this duality, some scholars like Karimi-Doostan (2005) have studied them in complete syntactical frameworks, and others like Barjeste (1998) have tried a lexical approach. Barjeste considers compound verb formation in Persian to be a productive lexical phenomenon that is the outcome of various operations in the lexicon.

In addition to these two trends, and also following the fact that dealing with 'complex predicate' constructions in the frameworks that make a distinction between the 'lexicon' and 'syntax' has posed many problems (Ramchand 2008), one can find other studies which have adopted a syntactic-semantic approach regarding the Persian compound verbs (Tabaian 1979, Vahedi-Langrudi 1996, Megerdoomian, 2002) and have tried to relate the different features of these verbs to both syntax and the lexicon module and even some have tried to eliminate the syntax-lexicon borderline to describe the dual feature properly. In any case, although the research done on this subject has chosen different theoretical frameworks and perspectives, most of them had one thing in common, that is finding a way to explain the productivity of these constructions and explain the semi-compositional semantic nature observed in them.

The dual and transitional nature of compound verbs is not specific to Persian. As some languages from different language families are reported to have the same feature. For example, Chinese as an isolating language (Yin 2010), Japanese as an agglutinating one (Kageyama 1993), Korean (Sup Jun 2007), and also Hindi (Chakrabarti et al. 2008), which is more similar to Persian in origin and linguistic features, are all of this kind and the syntactic or lexical nature of their compound verbs have been subject to much controversy among linguists from both theoretical and practical perspectives.

Besides the studies done with a theoretical perspective on Persian compound verbs, one can find others which have been conducted in recent years with practical purposes such as Persian machine translation and construction of

semantic web or WordNet (Megerdoomian 2004, Mansoory 2008, Shamsfard et al. 2010).

In the following subsections, we briefly review the structure of Persian compound verbs and the different types of compound verb formation and their semantic properties. Also, we discuss the semantic contribution of verbal and the preverbal elements in these constructions. Then the following sections will discuss the consequences of these properties in developing the Persian WordNet.

2.1 *The structure of Persian compound verbs*

As mentioned, Persian compound verbs are a combination of two elements: preverbal and verbal. The verbal elements are usually called 'light verbs' since their semantic content is bleached and their meaning is mostly unpredictable when they are used in a compound verb. *kærdæn* 'to do', *zædæn* 'to hit', *gereftæn* 'to take', *shodæn* 'to become', *dadæn* 'to give' and *aværdæn* 'to bring' are among the most commonly used light verbs in these constructions. Also, the preverbal elements are within the lexical categories noun, adjective, adverb, and the phrasal category prepositional phrase (Karimi 1997). Some combination patterns for making compound verbs are as follows:

1-	Noun + Verb:	*atæsh zædæn*	(fire- hit)	'to burn down'
2-	Adjective + Verb:	*tælx kærdæn*	(bitter-do)	'to make bitter'
3-	Adverb + Verb:	*dær gozæshtæn*	(off-pass)	'to die'
4-	Prepositional Phrase (PP) + Verb:			
		æz yad bordæn	(of-memory-take)	'to forget'

As the examples show, the meaning of the whole verb is not the sum of the meanings of its parts, and the meaning of the verbal element is not clearly represented in the whole.

2.2 *Compound verb formation processes in Persian*

According to Dabir-Moghaddam (1997), there are two major types of compound-verb formation in Persian named *Combination* and *Incorporation*. These two types of verb formation are described below.

(a) Combination

In this type of compound-verb formation, the non-verbal and the verbal constituent are combined in the following patterns. The Persian examples are shown in front of each item.

Adjective + Auxiliary[4]:	*delxor shodæn*	(annoyed-become)	'to become annoyed'
	delxor budan	(annoyed-be)	'to be annoyed'
	delxor kærdæn	(annoyed-make)	'to annoy'

Noun + Verb:	*zæmin xordan*	(ground hit)	'to fall'
	dærd gereftæn	(pain-take)	'to feel pain'
	næfæs keshidæn	(breath-draw)	'to breath'
	dæst dashtæn	(hand-have)	'to be involved'
Prepositional Phrase + Verb:			
	*æz beyn bordæ*n	(from-between-take)	'to destroy'
Adverb + Verb:	*dær yaftæn*	(in-find)	'to perceive'
Past Participle + Passive Auxiliary:			
	saxte shodæn	(built-become)	'to be built'

(b) Incorporation

This type of compound-verb may be formed by the following syntactic patterns:

Noun + Verb
Prepositional Phrase + Verb

But it differs from the previous category in the grammatical role of the noun in the first pattern and omitting the preposition in the second pattern.

In Persian, the direct objects (losing their grammatical marker such as the postposition 'ra') can incorporate with the verb, to create a compound verb (Dabir-Moghaddam 1997). The resulting verb is a syntactic-conceptual whole as shown in the following example (modified from Dabir-Moghaddam 1997):

(1)(a) *ma qæza -y-e-m-an- ra xord-im*
(we food-our-pl.-DOM[5] eat-past-we)
'We ate our food'

(1)(b) *ma qæza- xord-im*
'We ate food'

*(1)(c) *ma qæza -ye-man xord-im*
(we food-our-pl. eat-past-we)

(1)(d) *ma qæza - ra bimowqe xord-im*
(we food-DOM untimely eat-past-we)
'We ate food untimely'

*(1)(e) *ma qæza bimowqe xord-im*
(we food untimely eat-past-we)

These examples show that we can incorporate the direct object in (1)(a) to make an incorporated compound verb as in (1)(b) and the incorporated noun is part of the new compound verb and cannot be separated either by possessive pronouns (compare (1)(a) and (1)(c)) or by an adverb (compare (1)(d) and (1)(e)). In direct object incorporation, the argument structure of the verb changes and the transitive verb changes to an intransitive one, as a result of

incorporation. In addition to direct object incorporation, some prepositional phrases can also incorporate with verbs. Here, the proposition disappears after incorporation:

(2)(a) an-ha be zæmin xord-ænd
(that-pl. to ground hit[6]-past-they)
'They fell to the ground.'
(2)(b) an-ha zæmin xord-ænd
(that-pl. ground hit-past-they)
'They fell down.'

Since the focus of this paper is mainly on the Persian WordNet of verbs, the reader can refer to Dabir-Moghaddam (1997) for more detailed phonological, syntactic and semantic arguments for the incorporation process in Persian verbs.

2.3 *Compound verbs semantics*

As far as the semantic behavior of the different types of compound verbs is concerned, there is a wide disagreement among scholars. Dabir-Moghaddam (1997) shows that the verbal constituent has *transparent* meaning in one type of Incorporation -'Direct Object Incorporation'- and in one type of Combination -'Adjective + Auxiliary' combination. In other words, in these types of compound verbs, the meaning of the compound unit is the summation of the meanings of its verbal and its non-verbal constituents. In the other processes, however, there is a metaphorical extension and/or semantic bleaching of the verbal constituent of the compound, that is, the meaning of the whole compound *cannot* be considered to be the summation of its units. While some scholars, such as Mohammad and Karimi (1992), have suggested that the verbal elements are semantically empty, others like Vahedi-Langrudi (1996), Karimi-Doostan (1997), and Barjeste (1998) have stated that the light verbs contribute the aspectual meaning of the compound verbs but the argument structure of the compound verb is not based on the semantic content of the light verb. On the other hand, the thematic structure of the whole verb is the outcome of the semantic content of both the verbal and non-verbal element (Karimi 1997). That is why the compound verbs which are created by combination are not semantically considered to be merely non-compositional, but semi-compositional.[7] What is important is that in the case of transparent compound verbs although the meaning of the whole verb is actually the combination of its parts, as Dabir-Moghaddam (1997) stressed, this group of verbs is just like other non-transparent compound verbs in that they semantically

constitute a conceptual whole. For example, the sentence (3)(a) below contains an independent direct object which is incorporated in (3)(b).

(3)(a). *Mina zæhr-ra be Hæsæn dad*
(Mina poison-DOM to Hæsæn gave)
'Mina gave the poison to Hasan'

(3)(b). *Mina be Hæsæn zæhr-dad.*
(Mina to Hæsæn poison-gave)
'Mina poisoned Hasan'

As is clear in the English translation of each sentence, when the direct object is incorporated with the verb (3)(b), the outcome is a compound verb with the whole meaning (*to poison*) in which the combined nominal element no longer retains its definiteness and provides some semantic meaning.[8] In other words, it has a general meaning and does not refer to specific one (specific poison in this case); it has lost its referential feature as a noun becoming part of the verbal meaning of the compound verb. As shown in the translation, the compound verb in this case corresponds to a specific English verb (*to poison*), not to a phrase (*give the poison*).

3. The Persian WordNet of verbs

In the FarsNet project, as in every other WordNet project, the theoretical framework is a relational-semantic one with some lexicographic considerations. Our main goal is to find a way to represent the productive and semi-compositional nature of Persian compound verbs using the possibilities made available by relational-semantics. Also because the project aims at building a bilingual English-Persian semantic lexicon, we have tried to present some language specific semantic relations that represent the important differences existing in verbal systems of these two languages.

An important issue of concern to FarsNet developers, in particular in the construction of the verbal hierarchy, was the question of whether to include transparent compound verbs as lexical entries in the verbal synsets. It seems that the issue has not posed a challenge just to FarsNet developers. Chakrabarti et al. (2008) address a similar issue in the construction of the Hindi WordNet. In the process of automatic extraction of the Hindi compound verbs in their data base, they divide the Hindi V + V sequence into two groups: those that are formed in the syntax and those that are formed in the lexicon. They call the later group true Complex Predicates (CPs) because they function as single semantic units. As such, they are included in the Hindi WordNet. The former group is excluded from the lexical knowledge base because, as they

have stated, each verb in these constructions behaves as if it is an independent syntactic and semantic entity.

Addressing the issue in FarsNet, what makes the situation different with Persian compound verbs that are the outcome of direct object incorporation (as a syntactic operation), is that in contrast with syntactic compound verbs in Hindi, these Persian compound verbs function as a single semantic and syntactic unit and act as a semantic whole.

On the other hand, we can assume that 'Adjective + Auxiliary' produce 'transparent compounds' since the meaning of verbal and non-verbal constituent does not change. In other words, these kinds of compounds are directly derived from the combination of adjectives and auxiliaries. Since this is a rule-based mechanism, it would help us to define a general rule in the grammar level and avoid including such transparent compounds in the lexicon unless some specific conditions are met.

The meaning of the verbal and non-verbal constituents remain transparent also in the 'Direct Object Incorporation', however, as opposed to the 'Adjective + Auxiliary' compounds, this is not a productive process and we cannot always incorporate a direct object of a verb to make a compound verb.

There are some disagreements between researchers on assuming transparent compound verbs (incorporations) as lexical units or syntactic units. Entering these verbs in the lexicon enlarges it and affects the performance of management and retrieval. But on the other side, adding them to the lexicon improves the power of NLP systems which use this resource by providing more semantic knowledge. So we should reach to a practical tradeoff in this problem. The clue which guides us in selecting the most proper framework for our work is that we are using these theories to develop a Persian WordNet to be used in natural language processing applications. Entering the compound verbs into the lexicon and putting them in synsets and holding the lexical and conceptual relations among them facilitate semantic processing of the input texts by a machine. So we prefer to obey the theories which assume compound verbs as lexical units for those who can participate in a synset along with other compound or simple verbs. Following this fact, we have decided to include Persian compound verbs that are semantically transparent and constitute open sets (i.e. compounds formed through direct object incorporation and compounds that are the result of the combination of adjectives and auxiliaries) in the Persian WordNet of verbs as separate lexical entries and as members of verbal synsets in the database if one of the following conditions is applied:

(1) if they participate in a synset along with other compounds (non-transparents) or simple verbs;
(2) if the transparent compound of the kind 'Adjective + Auxiliary' has also a non-transparent, idiomatic meaning. For example the compound verb *siyah kærdæn* (black making) 'to make black' is transparent but it has a

second non-transparent meaning 'to deceive someone', so it is better to enter these kinds of verbs in the lexicon to show their different meaning by inserting them in different synsets;

(3) if their equivalent synset is present in the English WordNet and deleting this compound verb from Persian lexicon causes a gap between Persian and English WordNets.

For example, we have *mohtaj* 'needy' as an adjective which can be incorporated with *budan* 'to be' and make the compound verb *mohtaj budan* 'to be needy'. Although it is a transparent compound, we add it to the lexicon as it has synonyms from other constructions such as *niaz dashtan* (need have) meaning 'to be needy'. The whole synset will have an equality relation to the verbal synset {need} in PWN.

According to the above mentioned properties, we define two new relations for Persian compound verbs. The first one is TRANSPARENT_COMPOUND relation, which is held between the verbal and non-verbal constituent of the compound verbs which are combination of adjective or past participle and the auxiliary. Here the meaning of the compound verbs is transparent and it is the summation of the adjective or past participle and the following auxiliary. It must be noted that in cases in which the compound verb is a combination of adjective and auxiliary but has an idiomatic meaning, this relation is not defined between the adjective and the verbal element. So this relation can distinguish the transparent cases such as *siyah kærdæn* (black making) 'to make black' from other idiomatic meanings of the same verb like *siyah kærdæn* (black making) meaning 'to deceive someone'.

We also define the TRANSPARENT_INCORPORATION relation between the verbal and non-verbal constituents of the compound verbs which are formed by direct object incorporation. The meaning of these verbs is also transparent.

Including Persian transparent compounds in the database and defining these two language specific relations in FarsNet, can help us achieve two important objectives. First, to define a formal means in the database that systematically distinguishes transparent and opaque compounds and also paves the way for the automatic extraction of other possible transparent compounds in a Persian Corpus and adding them to the lexical database. Second, because the project aims at building a bilingual English-Persian semantic lexicon, by inserting the transparent compound verbs as separate semantic verbal concepts in the Persian WorNet of verbs, we can achieve a maximum level of matching between English and Persian.

The construction of the verb hierarchy in FarsNet, like its other parts, follows a top-down strategy on an 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; Persian and English. However in the antonymy and cause relations there are some language specific features which affect the structure of the Persian hierarchy and the way these relations are determined among the Persian verbal synsets. To elaborate on these points in the following two subsections, we briefly discuss the causation and antonymy features in Persian compound verbs and their effect in structure and construction process of the Persian verbal hierarchy.

3.1 *Cause relation in the Persian WordNet of verbs*

In the literature on the causative construction in Persian, the relation between compound verbs and the concept of causation has been the subject of interest (Dabir-Moghaddam 1982, Golfam and Bahrami-Khorshid 2009). Based on Comrie's typological framework on causative constructions (Comrie 1992), Golfam and Bahrami-Khorshid (2009) have classified the Persian causative construction into three main classes: *morphological causative, lexical causative, and analytic causative*.[9] As our interest is in building the lexicon, and the last class concerns purely syntactic constructions, we just review the two first classes. Morphological causatives include two subclasses:

(a) simple morphological causatives, and
(b) compound morphological causatives.

In the first form, the causative predicate differs from its non-causative counterpart by the inclusion of the suffix '*-an*'. In other words the suffix '*-an*' is added to a non-causative simple verb and changes it to a causative predicate:

(4) *pæridæn* 'to jump' / *pær**an**dæn*[10] 'to make someone/something jump'

Compound morphological causatives which are the most productive causatives in Persian, utilize the auxiliary *kærdæn* 'to make' to mark a causative compound verb. This replaces the auxiliaries *budæn* 'to be' and *shodæn* 'to become' found in non-causative forms:

(5) *tælx* ***budæn*** (bitter being) 'being bitter' and *tælx* ***shodæn*** (bitter becoming) 'becoming bitter' / *tælkh* ***kærdæn*** (bitter making) 'making bitter'

In 'lexical causatives', which are divided into three subclasses, the relation between the causative verb and non-causative verb is unsystematic and cannot be handled with systematic lexical rules. Among the three kinds of

lexical causatives, which are 'identical lexical causatives', 'non-identical lexical causatives' and 'compound non-identical lexical causatives', the third one refers to those causative verbs which are the result of replacing a light verb with another one. But this substitution does not follow a regular and general pattern:

(6) *atæsh* ***gereftæn*** (fire catch) 'to catch fire' / *atæsh* ***zædæn*** (fire hit) 'to fire or to burn down'
(7) *yad* ***gereftæn*** (memory catch) 'to learn'/ *yad* ***dadæn*** (memory give) 'to teach'

As these examples show, it is possible to substitute the light verb *gereftæn* 'catch' with *zadæn* 'hit' to make the compound verb causative and in other cases you replace the same light verb with *dadæn* 'give' to reach the causative verb and a general pattern cannot be found.

So it can be seen that, as in English, Persian has lexicalized causative pairs. However, due to the above mentioned morpho-semantic patterns among Persian simple and compound verbs, the number of Persian causative pairs is very high and there is no one-to-one mapping between PWN and the Persian WordNet. For example, the pair *lærzidæn*/ *lærzandæn* 'shake/ cause to shake', which belongs to the category of simple morphological causatives, has no corresponding pair or cause relation in the PWN. Instead, just one English verb 'shake' with two senses (causative and non-causative) are fused in one synset, and the separate senses are available only in the definition '*move or cause to move back and forth*'. But regarding the corresponding Persian concepts, because we have two different lexical items we construct two different synsets and relate the causative one to the other by means of the cause relation.

'Compound non-identical lexical causatives', are also difficult to map directly to the PWN. For example, Farsi replaces *kærdæn* 'make or do' with *shodæn* 'become' in*?ævæz kærdæn* (exchange make) 'change: cause to change'/ *?ævæz shodæn* (exchange become) 'change: undergo a change' and also *dædæn* 'give' with *kærdæn* 'do' in *ta?ghir dædæn* (change give) 'change: cause to change' / *ta?ghir kærdæn* (change do) 'change: undergo a change' to make non-causative and causative forms. English synsets which represent the same concepts contain only a cause relation between the two synsets {change2} and {change1} without separate lexical items. The 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 {open1} is defined as *'cause to open or to become open'* in WordNet 0.3. So in the construction of its 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 the one to the other by means of the cause relation.

For 'compound morphological causatives' a regular morphological rule may be used to determine the cause relation between the two synsets in FarsNet. We therefore apply a semi-automatic method to add these forms.

3.2 *Antonymy relation in the Persian WordNet of verbs*

In determining the antonymy relations among the verbal synsets in the Persian WordNet, we found that in most of the cases, when the verbs are compound and their preverbal elements are adjectives and nouns, existence of antonymy between the two adjectives or nouns will lead us to connect the two verbs with the same lexical relations. For example, since the nouns *dorugh* 'lie' and *rast* 'truth' are antonyms, the compound verbs *dorugh goftæn* (lie tell) 'to lie' and *rast goftæn* (truth tell) 'tell the truth' are linked with the same lexical relations. As our project proceeds and we have enough coverage for nouns and adjectives, using this morpho-semantic information will prove helpful in the semi-automatic extension of our verbal net.

4. Toward automatic extraction of verbs and their relations

In the previous sections, we discussed the properties of Persian compound verbs. In this section, we talk about the consequences of these properties which are used in the semi-automatic building of the Persian WordNet, particularly by generating possible candidates of verbs and their relations. The consequences are used in two major parts: extracting new verbs and extracting new relations.

Extracting new compound verbs from existing ones: In 'Direct Object Incorporation', since the non-verbal constituent is in fact the direct-object of the simple verb (i.e. the verbal constituent), the hypernyms, hyponyms and co-hyponyms of it are good candidates for being the potential direct object of that simple verb. We can use this property for generating a list of potential compound verbs from such potential objects, assuming that they *can* be incorporated with the verbal element. The generated verbs are not necessarily real compound verbs and we need to revise them either manually or by looking for them in dictionaries and text corpora. For example, the verb *qæza xordæn* (food eating) has the non-verbal component *qæza* 'food' which has a hyponym *nahar* 'lunch' and so the verb *nahar xordæn* (lunch eating) can be a candidate for being a compound verb. Finding nouns like *nahar-xori* 'place in which we eat lunch' from this combination indicates that this may be considered a compound verb. In some cases, the candidate is a compound verb with an idiomatic meaning. For example considering *nan* 'bread' as a hyponym of *qæza*

'food', we may build the verb *nan xordæn* (bread eating) which is a compound verb with a new idiomatic meaning 'to earn money or to spend money'.

Extracting new relations: In general, one can say that every relation and characteristic of the non-verbal constituents of compound verbs (which do not have idiomatic meaning) may be transferred to the whole compound verb as well. For instance, the compound verb c1 *isa* compound verb c2 (which means that c1 is the hyponym of c2 and c2 is the hypernym of c1) if there is an *isa* (hyper/hyponymy) relation between the preverbal part of c1 and the preverbal part of c2 and their verbal parts are equal (transferring hyper/hyponymy relationship). This is true if the relation between preverbal parts is synonymy or antonymy and sometimes for meronymy. Thus, we can infer the following relation discovery heuristic rules in the semi-automatic development of the Persian WordNet:

(a) *Synonymy discovery*: Considering two compound verbs c_1 and c_2 in which p_i and v_i are the preverbal part and the simple (light) verb of c_i respectively, we can infer that c_1 and c_2 are synonyms if p_1 is equal to or a synonym of p_2 and v_1 is equal to or the synonym of v_2. As an example for having synonym non-verbal components and equal light verbs, consider the combination of the nominal synset {*zærær, xesaræt* 'loss'}-which includes two synonym nouns- with the identical verbal component *zædæn* 'to hit, to make' which makes the verbal synset {*zærær zædæn, xesaræt zædæn* 'to damage'} including two synonym compound verbs. As another example for having synonymous nonverbal and verbal components, consider the verbal component synset {*kærdæn, Gozardæn, goftæn* 'to do'}, which can be added to the nominal synset {*Sokr, sepas* 'thanks'} and create the verbal synset {*Sokr kærdæn, sepas gozardæn, sepas goftæn* 'thanks giving'}.

(b) *Hypernym/Hyponym (troponymy) discovery*: There is a hypernym relation between two compound verbs if there is such a relation between their preverbal parts and equality or synonymy between their light verbs. For example, considering the nouns *hærekæt* 'movement' and *rej?æt* or *moraje?æt* 'return', which participate in a hypernym relation, leads to the compounds *hærekæt kærdæn* 'to move' and *rej?æt kærdæn* or *moraje?æt kærdæn* 'to return', which have a hypernym relation as well. There is also a hyponymy/hypernymy relation between the verbal constituent and the whole compound verb in all transparent compounds. In other words, the transparent compound verb is the Hyponym of its verbal constituent. For example the compound verb *qæza khordæn* 'to eat food' is a hyponym of its light verb *khordæn* 'to eat' and also *qæza dadæn* 'to feed' is a hyponym of its light verb *dadæn* 'to give'.

(c) *Causative relation discovery*: In many cases, causative relations are among synsets whose verbal components have causative relations with each other

(are causative/ non-causative alternations). We prepared a list of causative/ non-causative alternations of the verbal components, including for example, the following pairs:

/*zædæn*/	'to strike'	- /*xordæn*/	'to receive'
/*zædæn*/	'to strike'	- /*didæn*/	'to see'
/*resandæn*/	'to carry'	- /*didæn*/	'to see'
/*resandæn*/	'to carry'	- /*xordæn*/	'to receive'

Adding the same or synonymous non-verbal components to these pairs will result in compound verbs with causative relations. For example the following synsets (A) and (B) are made by adding the above alternations to the members of nominal synset {*sædæme, lætme, asib* 'hurt'}:

(A){/*sædæme zædæn*/, /*sædæme resandæn*/, /*sædæme vared kærdæn*/, /*lætme zædæn*/, /*lætme resandæn*/, /*lætme vared kærdæn*/, /*asib resandæn*/, /*asib zædæn*/, /*gæzænd resandæn*/ 'to hurt'}
(B) {/*sædæme xordæn*/, /*sædæme didæn*/, /*lætme xordæn*/, /*lætme didæn*/, /*asib didæn*/ 'to be hurt'}

(d) *Antonymy Discovery*: compound verbs which are made from an adjective + verb have antonymy relations if their adjective parts have the same relation. For example the antonym nouns *bala* 'up' and *pa?in* 'down' are merged with the same light verb *ræftæn* 'to go' to form antonym compound verbs *bala raftan* 'to increase, go up' and *pa?in ræftæn* 'decrease, go down'.
(e) *Other relations*: The transparent compound verbs have direct semantic relations to their non-verbal constituents. These relations sometimes extend to cover their neighbors (parents and children in the inclusion hierarchy). As an instance it can be seen that in direct object incorporation compounds, there is a *potential-object-of* relation between the hyponym and hypernym of the incorporated object and the verbal element. For example in *qæza xordæn* 'food eating', the non-verbal component *qæza* 'food' and all its hyponyms can potentially be object of the verb *xordæn* 'to eat' and so could be assumed in the selectional restrictions of the theme role (or object argument) for this verb.

5. Conclusion

In this paper, we discussed the structure, syntax and semantics of Persian compound verbs and described the Persian language specific features, which we need to consider in developing the Persian WordNet of verbs. In this process we obey the theoretical frameworks which consider a wide range of

compound verbs as lexical phenomena rather than syntactic ones. This helps us to have richer synsets in which compound and simple verbs are present. It facilitates the next processes needed for Persian NLP with fewer efforts.

The paper shows the differences between the Persian and English verbal system as well. These differences make it hard to build the Persian WordNet directly from translating English synsets by an expand approach.

We also introduced some consequences of the discussions and their application in semi-automatic building of the Persian WorldNet of verbs. Extending the test environment, extracting new relations, and applying the results on FarsNet are among our ongoing tasks.

Notes

1 http://en.wikipedia.org/wiki/Persian_language

2 The project is housed in Princeton's Cognitive Science Laboratory.

3 See 'WordNets in the world' at: http://www.globalWordNet.org.

4 Dabir-Moghaddam (1997) does not refer to the concept of light verb in Persian compound verbs and among the verbal elements of these constructions, classifies stative *budæn* 'to be', inchoative *shodæn* 'to become', and causative *kærdæn* 'to make' as auxiliaries and mentions the fact that the compound verbs formed by these auxiliaries, constitute an open set. He also calls the other light verbs just verbs.

5 Direct Object Marker

6 *xordæn* has different literal meanings. One of them is 'to eat' and the other as mentioned in this example is 'to hit'.

7 A detailed study of semi-compositionality of the Persian complex predicates can be found in Family (2006) and Mansoory and Bijankhan (2008).

8 There is no overt definite marker in Persian. The subject position with no noun marker is construed as definite but the direct object non-referential bare nouns are distinguished from definite bare nouns by the presence of 'ra'.

9 In addition to these three classes, they also introduce the existence of another class of causatives named Discoursal Causatives in this language.

10 In Old Persian 'i' was also present in causative verbs. For example the causative form of this verb was 'pæranidæn', but in modern Persian 'i' is usually omitted after the addition of the causative morpheme.

References

Agirre, E. and Ph. Edmonds, (eds) 2006. *Word Sense Disambiguation: Algorithms and applications*. Springer-Verlag, 1–28.

Barjeste, D. 1998. *Morphology, Syntax and Semantics of Persian Compound Verbs: A Lexical Approach*. PhD Thesis, University of Illinois.

Chakrabarti, D., M. Hemang, P. Ritwik, S. Vaijayanthi and B. Pushpak. 2008. 'Hindi Compound Verbs and their Automatic Extraction'. In D. Scott and H. Uszkoreit (eds), *Proceedings of International Conference on Computational Linguistics (COLING)*. 27–30.

Comrie, B. 1992. *Language Universals and Linguistic Typology: Syntax and Morphology*. 2nd edition. Blackwell.

Dabir-Moghaddam, M. 1982. *Syntax and Semantics of Causative Constructions in Persian*. PhD Thesis, University of Illinois.

Dabir-Moghaddam, M. 1997. 'Compound Verbs in Persian.' *Studies in the Linguistic Science*, 27.2: 25–59.

Family, N. 2006. *An Interface Account of Light Verb Construction in Persian*. PhD Thesis, Ecole des Hautes Etudes en Sciences Sociales- Paris.

Famian, A. and D. Aghajaney. 2006. 'Towards Building a WordNet for Persian Adjectives'. In P. Sojka, K. Choi, Ch. Fellbaum and P. Vossen (eds), *Proceedings of the 3rd Global WordNet Conference*. South Korea: 307–308.

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

Golfam, A. and S. Bahrami-Khorshid. 2009. 'Causation as a Mental Process.' *Pazhuhesh-e Zabanha-ye Khareji*, 49: 125–139.

Kageyama, T. 1993. *Grammar and Word Formation*. Tokyo: Hitsuji Shobo.

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

Karimi-Doostan, G. 1997. *Light Verb Construction in Persian*. PhD Thesis, University of Essex.

Karimi-Doostan, G. 2005. 'Light Verbs and Structural Case.' *Lingua*, 115: 1737–1756.

Keyvan, F., H. Borjan, M. Kasheff and C. Fellbaum. 2006. Developing PersiaNet: The Persian Wordnet. In P. Sojka, K. Choi, Ch. Fellbaum and P. Vossen (eds), *Proceedings of the 3rd Global WordNet Conference*. South Korea: 315–318.

Mahootian, Sh. 1997. *Persian*. Routledge.

Mansoory, N. 2008. *Semantic Representation of Complex Verbs in Persian WordNet*. PhD Thesis, Tehran University.

Mansoory, N. and M. Bijankhan. 2008. 'The Possible Effects of Persian Light Verb Constructions on Persian WordNet'. In A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum and P. Vossen (eds), *Proceedings of the 4th Global WordNet Conference (GWC 2008)*. Szeged, Hungary: 297–303.

Megerdoomian, K. 2002. *Beyond Words and Phrases: A Unified Theory of Predicate Composition*. PhD Thesis, University of Southern California.

Megerdoomian, K. 2004. 'A Semantic Template for Light Verb Constructions' *Proceedings of First Workshop on Persian Language and Computers*. Tehran, Iran.

Miller, G.A. 1995. 'WordNet: a Lexical Database for English.' *Communications of the ACM archive*, 38.11: 39–41.

Mohammad, J. and S. Karimi. 1992. 'Light Verbs are Taken over Complex Verbs in Persian'. In J.A. Nevis and V. Samiian (eds), *Proceedings of WECOL5,* 195–212.

Morato, J., M. Marzal, J. Liorens and J. Moreiro. 2004. 'Wordnet Applications'. In P. Sojka, K. Pala, P. Smrz, Ch. Fellbaum and P. Vossen (eds), *Proceedings of the 2nd Global WordNet Conference (GWC 2004)*. Brno, Czech Republic: 270–278.

Ramchand, G. 2008. *Verb Meaning and the Lexicon: A First Phase Syntax*. Cambridge University Press.

Rouhizadeh, M., M. Assi and M. A. Yarmohamadi. 2007. 'Designing Persian Verbs WordNet'. In M. DabirMoghaddam, M. Assi, A. Golfam and Y. Modarresi (eds), *Proceedings of the 7th Iranian Conference on Linguistics*. Tehran, Iran: 518–530.

Rouhizadeh, M., M. Shamsfard and M. A. Yarmohamadi. 2008. 'Building a WordNet for Persian Verbs'. In A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum and P. Vossen (eds), Proceedings of the 4th Global WordNet conference (GWC 2008). *Szeged*. Hungary: 406–412.

Shamsfard., M. 2008a. 'Developing FarsNet: A Lexical Ontology for Persian'. In A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum and P. Vossen (eds), *Proceedings of the 4th Global WordNet conference (GWC 2008)*. Szeged, Hungary: 413–418.

Shamsfard., M. 2008b. 'Towards Semi Automatic Construction of a Lexical Ontology for Persian'. In N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis and D. Tapias (eds), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: 2629–2633.

Shamsfard, M, A. Hesabi, H. Fadaie, A. Famian, S. Bagherbeigi, N. Mansoory, E. Fekri, M. Monshizadeh and S.M. Assi. 2010. 'Semi Automatic Development of FarsNet; The Persian WordNet'. In P. Bhattacharya, Ch. Felbaum and P. Vossen (eds), *Proceedings of 5th Global WordNet Conference (GWA2010)*. Mumbai, India: 184–197.

Sup Jun, J. 2007. '*Co*-event Conflation for Compound Verbs in Korean'. In Proceedings of *Annual meetings of the Pacific Asia Conference on Language, Information and Computation (PACLIC)*: 202–209.

Tabaian, H. 1979. 'Persian Compound Verbs.' *Lingua*, 47: 189–208.

Vahedi-Langrudi, M. M. 1996. *The Syntax, Semantics and Argument Structures of Complex Predicates in Modern Farsi*. PhD Thesis, University of Ottawa.

Yin, H. 2010. 'The so called Chinese VV Compounds: a Continuum between Lexicon and Syntax'. In M. Heijl (ed.), *Proceedings of the 2010 annual conference of Canadian Linguistic Association*: 1–10.