Developing the Persian WordNet of Verbs; Issues of Compound Verbs and Building the Editor

Masoud Rouhizadeh
NLP Research Laboratory
Shahid Beheshti University
Tehran, Iran
mrouhizadeh@gmail.com

Mahsa A. Yarmohammadi
NLP Research Laboratory
Shahid Beheshti University
Tehran, Iran
yarmohamadi@gmail.com

Mehrnoush Shamsfard
NLP Research Laboratory
Shahid Beheshti University
Tehran, Iran
m-shams@sbu.ac.ir

Abstract

In this paper we mostly focus on the behavior of Persian compound verbs and the way we propose to deal with them. Most of the Persian verbs are compound verbs and they are formed by two major patterns of combination and incorporation. In many cases the compound verbs are semantically transparent. This behavior of the verbs has some important consequences in the Persian semantic lexicon hence; we design an editor to fully support it. The system architecture is three-tier model and in analysis, design and implementation of this editor we used prototyping methodology. The database consists of 11 tables which are related to each other by definite relations and store Persian verbs, nouns, adjectives, adverbs, prepositions and synsets. The results are compatible to other WordNets and the information is exportable to XML.

1 Introduction

Persian is the official language of three countries and it is also spoken in more than six other countries. 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, the unavailability of domain-independent lexica with rich semantic information on lexical items. Such lexica could greatly improve the quality of current applications.

There have been some attempts for reaching this goal (Famian & Aghajaney 2006; Keyvan et al., 2006; Mansoori & Bijankhan, 2008); however, most of them are only considering design of the structure and, in practice, limited sets of words or lexemes are entered in the lexicon.

This paper is a report of an ongoing project of developing the Persian WordNet of verbs, persuading our previous work (Rouhizadeh et. al. 2007 and 2008). It is a part of a larger project of building a semantic lexicon for Persian called FarsNet (Shamsfard, 2008).

Here we mostly focus on the behavior of Persian compound verbs and the way we propose to deal with them. Then we will review the editor of the WordNet of Persian verbs which is designed to handle the compound verbs phenomena in Persian.

This paper is divided into two parts; first, we give some theoretical considerations about Persian compound verbs and then we briefly review the editor of the Persian WordNet of verbs.

2 Compound verbs in Persian WordNet

Persian verbs can be divided into two major morphological categories: simple and compound verbs. Compound verb formation is highly productive in Persian. The number of simple verbs in Persian today, is less than 200 verbs while the number of compound verbs is more than 4000. compound verb formation is highly productive in Persian today. Persian compound verbs show interesting semantic behavior and a good semantic lexicon of Persian should deal with such particular characteristics. In the following subsections we briefly review different types of compound verb formation in Persian and their semantic properties, then, we review the consequences of these properties in Persian WordNet.

2.1 Persian compound verbs and their semantics

According to Dabirmoghaddam (1997) there are two major types of compound-verb formation in Persian which are *Combination* and *Incorpora-*

tion. These two types of verb formation are described below.

2.1.1 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: *delxor-shodan* 'to become annoyed' 'annoyed-become'

Noun + Verb: bâzi-kardan *'to play' 'play-do'*

Prepositional Phrase + Verb: *be donya âmadan* 'to be born' 'to-world-come'

Adverb + Verb: dar yâftan *'to perceive' 'infind'*

Past Participle + Passive Auxiliary: *sâxte ŝodan* 'to be built' 'built-become'.

2.1.2 Incorporation

In Persian, the direct objects (losing its grammatical endings) can incorporate with the verb, to create a compound verb, which is a conceptual whole as shown in the following example:

1-a. mâ qazâ-y-e-m-ân- râ xor-d-im
we food-our-pl.-DO eat-past-we
'We ate our food'

1-b. mâ qazâ- xor-d-im
'We did food eating'

In direct object incorporation the argument structure of verb changes and the transitive verb changes to intransitive, as a result of incorporation.

Also, some prepositional phrases can incorporate with verbs. Here, the proposition disappears after incorporation:

2-a. ân-hâ be zamin xor-d-and
that-pl. to ground eat-past-they
'They fell to the ground.'

2-b. ân-hâ zamin xor-d-and
'They fell down.'

2.1.3 Compound verbs semantics

As far as the semantic behavior of the different types of compound verbs are concerned, statistical findings show that the verbal constituent has *transparent* meaning in "Direct Object Incorporation" and "Adjective + Auxiliary" combination. In other word, in these types of compound verb, the meaning of the compound unit is the summations 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 as the summation of its units. Interestingly, in "Noun + Verb" compounding the verbal constituent is lexicalized to serve as an aktionsart marker and it shows how the action of the whole verb is performed. A detailed study of this behavior in Persian WordNet can be found in Mansoori and Bijankhan (2008).

2.2 New relations for Persian compound verbs

According to the properties mentioned above, we define the two new relations for Persian compound verbs. The first one is TRANSPARENT_COMPOUND relation which exists 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.

We also define the TRANSAPRENT_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.

We defined also collocation relation between verbal and non-verbal constituents of all compound verbs regardless of their compounding process. This is not a semantic relation but it is very effective relation for detection of the verbs in syntactic/semantic processing.

2.3 Compound verbs and the structure of Persian WordNet

This classification of Persian verbs has five consequences in the structure of Persian WordNet of verbs.

The first consequence is the possibility of defining Hyponymy/Hyperonymy relation between the verbal constituent and the whole compound verb in all transparent compounds. In other words the compound verb is the Hyponym of its verbal constituent.

The second consequence shows itself in the relations between different parts of speech. The transparent compound verbs have direct semantic relations to their non-verbal constituents. Every relation and characteristic of these non-verbal

constituents can be transferred to their compound verbs.

The third consequence is about the compound verbs which are formed by combination of adjective and auxiliaries budan (to be), šodan (to become) and kardan (to do). Once we have entered their semantic information in our database, we can predict systematically, the meaning of their compound verbs – a function of an adjective and an auxiliary.

The fourth consequence is about the verbs formed by direct object incorporation. Their non-verbal consequent was in fact the direct object of the simple verb. Thus, their hyperonyms, hyponyms and co-hyponyms are very good candidates for the direct object (without 'râ', the direct object marker in Persian) of that simple verb.

The fifth consequent is about the verbs formed by the combination of noun + verb. In this kind of verbs the simple verb is lexicalized to serve as an aktionsart marker. The real or metaphoric viewpoints are stored in Persian our database and in the case of their combination to nouns this information is transferred into the whole compound verb.

These kinds of relationship are special characteristics of the Persian verbs lexicon. All these special features of Persian verbs are completely supported in our editor. In the following section we briefly review the editor and its technical aspects.

3 The editor for Persian WordNet of verbs

Considering the above-mentioned properties of the Persian compound verbs we required an editor to fully support those special characteristics and semantic behavior.

We were using VisDic, the BalkaNet multilingual editor (Tufis et. al. 2004) for a while but it was not supporting compound verbs special characteristics and semantic behavior. It was also not appropriate for Persian scripts in the sense that since it could not support the right-to-left direction and some encodings of Persian scripts. On the other hand, the information in our verb lexicon is not limited to WordNet relations and we are going to put some more information like semantic restriction, verbs frames and so on.

On the base of these facts, we started to design the editor from scratch to fulfill our needs. We focus on the WordNet information of verbs since the FarsNet project concerns this at the first phase. However, the editor and its database are extendible to support many other kinds of relations and information such as those in FrameNet and VerbNet.

What will be discussed in the rest of this sections is the first[1] editor that we developed for WordNet of verbs. The editor is designed based on relational database model and the results can be exported to standard XML format. This feature makes our results compatible to the WordNets of the other languages. In this section we discuss about the design methodology of the editor, its graphical user interface and different parts of it and the editor's database model. Finally there is a note of compatibility of the editor's output with the other WordNets.

3.1 Methodology

In the analysis, design and implementation of this editor we used prototyping methodology. The programming language is Visual Basic 6. The software architecture is based on three-tier model, which are: 1) Data Access layer, 2) Business Logic layer and 3) Presentation layer. We use an adaptor between the first and the third layers to keep the adaptability to other prospective DBMSs. Each of these three layers has a dedicated and special responsibility. As a result, experts can make different changes in any layer with no interference to the other layers. To produce the web version of the editor, we simply used the first and the second layers of the desktop version. Then we design only the third layer.

3.2 The Graphical User Interface (GUI)

The graphical user interface of this editor is designed to make the manual lexicography task easy and straight-forward. It has two forms to edit verb entries and synsets separately. This provides the facility for the lexicographer to edit individual verb or the verbs within the synsets independently. When launching the editor, the main menu appears and one can select to go for "Verbs" or "Synsets". This directs the lexicographer to the "List of Verbs" or the "List of synsets" forms which are described below in details.

[1] We have ported and expanded this editor to the second one in Java working on XML data to be able to work platform free. We have also developed a web based version of the editor to enable collaborative lexicography over the internet. The theoretical foundations and design criteria of the first editor is inherited to its successors too.

3.2.1 Verbs entries

The first window of the verb editor is a "List of Verbs". Here one can view all the existing verbs of the database, edit a verb by double clicking on it, delete a verb and add a new verb. If the lexicographer chooses to add a new verb he/she will be redirected to another window of "Add New Verb". This window is quite similar to "Verb Properties" window which will be appeared if he/she selects a verb to edit (Figure 1).

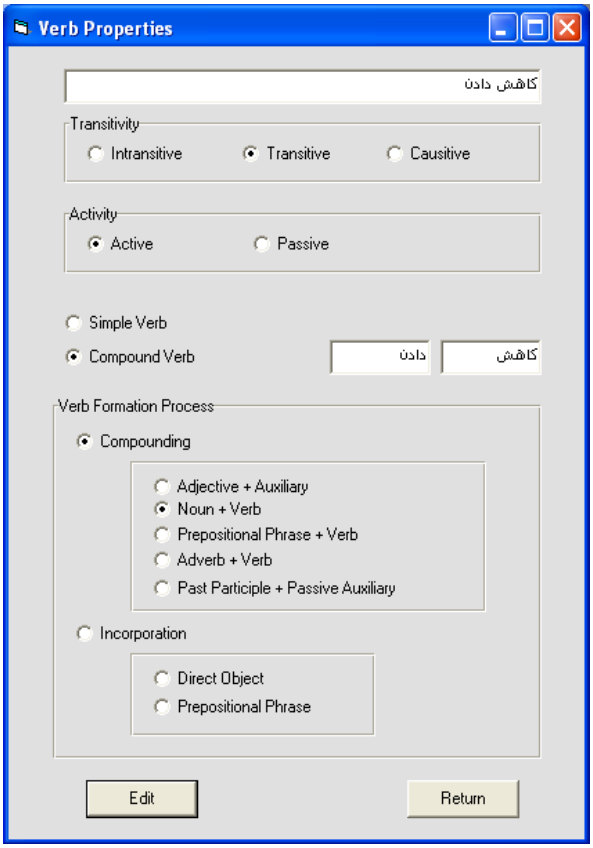

Figure 1: The window of "Verb Properties"

In this window all the necessary information should be presented for a single verb entry. At first the lexicographer selects the transitivity of a verb that is weather the verb is "Intransitive", "Transitive" or "Causative". In the next part he/she selects if the verb is "Active" or "Passive". The most important part is defining the morphological structure of the verb. He/she can select if the verb is "Simple" or "Compound". If he/she selects the verb as a compound verb, the two verbal and non-verbal constituents are automatically separated. No information will be saved in the database about the constituent structure of the verb until he/she selects the type of "Verb Formation Process". If he/she chooses the "Compounding Process", then it is necessary to select whether the verb is composed of "Adjective + Auxiliary", "Noun + Verb", "Prepositional Phrase + Verb", "Adverb + Verb" or "Past Participle + Passive Auxiliary". If he/she chooses the "Incorporation" process then it is necessary to choose whether the verb is formed through "Direct Object" or "Prepositional Phrase" incorporation.

As the verb formation process is selected, all the non-verbal constituents are saved separately in their related tables i.e. the tables of Nouns, Adjectives, Prepositions, etc.

As mentioned before, verbal constituent has transparent meaning in "Direct Object Incorporation" and "Adjective + Auxiliary" verb formation processes. The meaning of the resulting compound verb in these cases is the function of the meaning of its verbal and non-verbal constituents. So following information will be saved about such kind of entries: a) the compound verb, b) the non-verbal constituent, c) the verbal constituent (if this constituent is not entered previously, a new window will appear to add it) and d) a direct link between verbal and non-verbal constituent to show transparency of the meaning of the compound verb. The TRANSPARENT_COMPOUND and TRANSAPRENT_INCORPORATION relations are defined automatically as we select one of these types of compounding process. A collocation relation between the verbal and non-verbal constituent is also saved here.

The abovementioned two types of verbs inherit also the relationships which belong to their non-verbal counterparts. This is the way in which we connect lexicon of Persian verbs to the lexicon of Persian adjectives (in compounding process of Adj.+ V.) and the lexicon of Persian nouns (the direct objects incorporation process).

We also mentioned that in the other compound verb formation processes there is a metaphorical extension and/or semantic bleaching of the verbal constituents and this constituent does not have a transparent meaning. As a result we would save the following information: a) the compound verb, b) the non-verbal constituent and c) the verbal constituent (if not exists, a new window will appear to add this verb) but *no* direct link between the two constituents. Instead, we save the type of verbal constituent in Noun + Verb compounds. This constituent serves as the aktionsart marker. A collocation relation is saved for these kinds of verbs too.

Once the lexicographer entered the information of each verb in this form, the verb will be

ready to be a part of a synset (group of synsets). To define the synsets there are separate forms in the editor which will be described in the next subsection.

3.2.2 Synsets

The editor goes to list of the whole synsets if you select the "Synsets" option from the main menu. Is the lexicographer selects to bottom of "Synsets" he/she can view all the existing synsets of the database, edit a synset by double clicking it, delete a synset and or a new synset. If he/she chooses to add a new synset he/she will be redirected to another window of "Add New Synset" (Figure 2). This window is quite similar to "Synset Properties" window which will be appeared if he/she selects a synset to edit.

For each synset it is possible to enter the following information:

Synset ID: This ID can be defied automatically or manually. Our lexicon could be connected to other existing WordNets via the links of Persian synsets IDs to their equivalent synset IDs in Princeton WordNet 3.0.

Definition: A definition of a synset is given in this field which is very similar to existing definitions in mono-lingual dictionaries.

Usage: An instance of the synset (or particular word(s) of it) usage is given in this filed.

Synonym words: Here the lexicographer should enter all the synonym words which form a synset. He/she can add every verb by typing it in the blank input box. If the verb was not defied before, the editors opens "Add New Verb" window and the new verb is added after definition process. In addition he/she can select the synonym verbs from the list of existing verbs. The list of verbs appears as he/she presses this bottom and it is possible to select one or more verbs from the list. Finally, he/she can delete any verbs of the synonym sets, or edit it by double clicking on it.

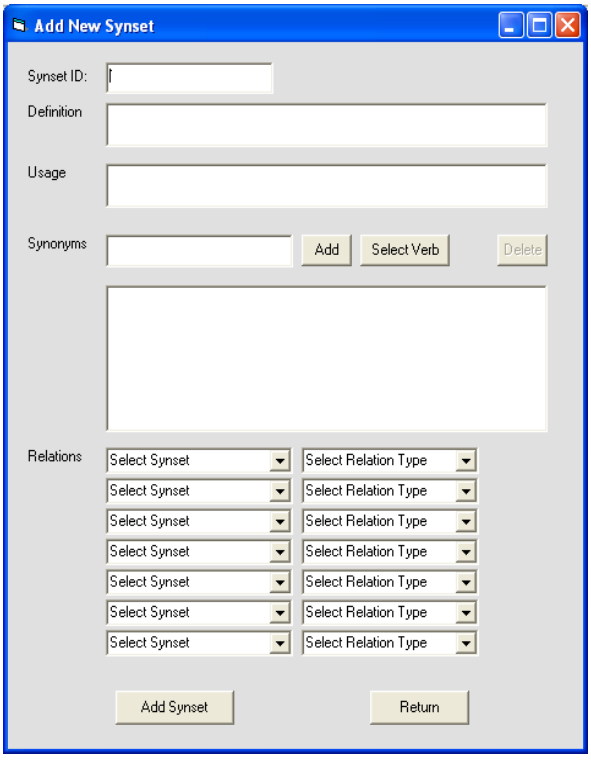

Figure 2: The "Add New Synset" form

Relations: it is possible to establish a relation among the current synset and the other synsets via different kinds of relations. These relations, which are mainly derived from EuroWordNet, include:

HAS_HYPERONYM
HAS_HYPONYM
ANTONYM
NEAR_ANTONYM
CAUSES
IS_CAUSED_BY
HAS_SUBEVENT
IS_SUBEVENT_OF

These relation exist among *verb* synsets, however, there are some other relations which exist among verbs and other parts of speech:

XPOS_NEAR_SYNONYM
HAS_XPOS_HYPERONYM
HAS_XPOS_HYPONYM which exists between a verb and a noun and:
XPOS_NEAR_ANTONYM between a verb and an adjective and:
IN_MANNER and
MANNER_OF which exist between a verb and adverb and vice versa.

It is possible to select the above relations among the verb synsets and the synsets of other parts of speech.

3.3 The database model

The relational database of Persian WordNet of verbs consists of 11 related tables. The tables contain data for Persian Verbs, Nouns, Adjectives, Adverbs, Prepositions and Synsets. Table of VERB for instance stores verbs and their properties such as the verb's transitivity, verb's morphological structure and verb's compounding process. There are also three tables for storing the information of Persian synsets and their different relations. Figure 3 shows the database architecture of Persian WordNet of verbs.

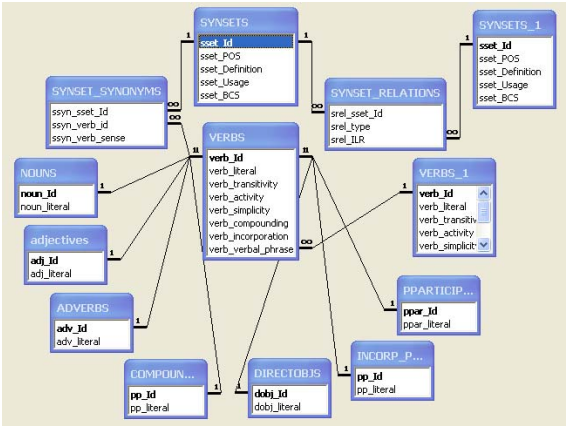

Figure 3: The database architecture of Persian WordNet of verbs

3.4 Compatibility

Our verb lexicon is structurally compatible to the other existing WordNets. All the information in the database is exportable to XML format. The data in SYNSETS table and its related tables i.e. SYNSET_SYNONYMS and SYNSET_RELATIONS could be exported to standard XML of WordNets. In this format each synset is defined within a <SYNSET></SYNSET> including some other inner tags such as <ID>, <POS>, <DEF>, <USAGE> and <SYNONYM>.

4 Conclusion and further work

In this paper we have discussed some linguistic issues of Persian compound verbs and their effects on developing the Persian WordNet of verbs. We also showed the features of the first editor which is specially designed for Persian verbs and their characteristics. Defining a web service for our verb lexicon to retrieve and publish data in XML format for other NLP softwares or websites is among our further works.

Once we have a larger lexicon, the editor may be set for deeper semantic processing. We can provide hierarchical definition of the verbal constituents to predict the meaning of the whole compound verbs. This feature reduces redundancy in the database to a great extent.

Adding the argument structures of verbs and their semantic restrictions are also in our future planning of this project.

Acknowledgments

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

References

Mohammad Dabir-Moghaddam. 1997. Compound Verbs in Persian. *Studies in the Linguistic Science*, 27(2), 25–59.

Ali Famian, Darioush Aghajaney. 2006. Towards Building a WordNet for Persian Adjectives. In *Proceedings of the 3rd Global WordNet conference*, pp. 307–308. South Korea.

Christiane Fellbaum (ed.), *Wordnet: an Electronic Lexical Database*, MIT Press, 1998.

Farhad Keyvan, (et. al.). 2006. Developing PersiaNet: The Persian Wordnet. In *Proceedings of the 3rd Global WordNet conference*, pp. 315-318. South Korea

Niloufar Mansoory, Mahmood Bijankhan .2008. The Possible Effects of Persian Light Verb Constructions on Persian WordNet. In *Proceedings of the 4th Global WordNet conference (GWC 2008)*, Szeged, Hungary.

Masoud Rouhizadeh, Mostafa Assi, Mahsa A.Yarmohamadi. 2007. Designing Persian Verbs WordNet. In *Proceedings of the 7th Iranian Conference on Linguistics*, Tehran, Iran.

Masoud Rouhizadeh, Mehrnoush Shamsfard, Mahsa A.Yarmohamadi. 2008. Building a WordNet for Persian Verbs. In *Proceedings of the 4th Global WordNet conference (GWC 2008)*, Szeged, Hungary.

Mehrnoush Shamsfard. 2008a. Developing FarsNet: A Lexical Ontology for Persian. In *Proceedings of the 4th Global WordNet conference (GWC 2008)*, Szeged, Hungary.

Mehrnoush Shamsfard. 2008b. Towards Semi Automatic Construction of a Lexical Ontology for Persian. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Morocco.

Dand Tufis, et. al. 2004. Special Issue on the BalkaNet Project. *Romanian Journal of Information Science and Technology*, Vol. 7, Nos 1–2.