A Semantic Template for Light Verb Constructions

Karine Megerdoomian
University of California, San Diego
karinem@ling.ucsd.edu

1. Introduction

Multiword Expressions (MWEs) raise an important problem for the development of large-scale NLP systems. Sag et al (2002) define MWEs as “idiosyncratic interpretations that cross word boundaries (or spaces)”. In an English lexicon such as WordNet 1.7 (Fellbaum 1999), 41% of the entries are multiwords. In Persian, this number is much larger since most verbs are formed by putting together a preverbal element (such as a noun, adjective, or adverbial) with a light verb. In addition, many adverbial or prepositional elements are listed as MWEs in the language.

The main problem posed by MWE processing or generation arises from the fact that these multiwords display both lexical and phrasal properties. The lack of compositionality of the semantics of these expresseions has led some researchers to simply list them in the lexicon. However, the components of many MWEs can be separated from each other by intervening material in syntax and are often missed by systems that treat these expressions as single units.

In this paper, we will focus on one genre of multiword expressions, namely the light verb constructions (LVCs) in Persian. LVCs display an idiomatic reading yet can undergo a number of syntactic operations such as scrambling, relative clause formation, and internal modification. In natural language processing systems, an enumerative lexicon that simply lists the MWEs will not only be faced with lexical proliferation given the productivity of these constructions, but will also be unable to deal with the wide range of syntactic patterns they may appear in. On the other hand, a fully compositional approach would suffer from the idiomatic interpretations of these phrasal structures and will have trouble in generating correct meanings or translations. Sag et al (2002) and Abeillé (1988) claim that LVCs are highly idiosyncratic in the sense that it is quite difficult to predict which preverbal element can combine with a given light verb. However, a number of recent approaches in linguistics – in various fields ranging from lexical semantics and generative semantics to computational linguistics and formal syntax – have been able to detect patterns and restrictions on the combinations of these elements.

We argue that the latest developments in analyzing these complex verbal predicates along with the linguistic research performed on Persian light verb constructions can shed some light on a successful computational modeling of multilingual complex predicates in general, and on Persian light verb constructions in particular. The approach proposed here is based on the lexical semantic representation of verbal predicates by providing a template that reflects the combination of the various primitive components and the realization of the verb’s argument structure in syntax. By extending the research described in Fong et al (2000) to Persian verbal constructions, we propose to map the semantic templates as an interlingual representation. It is argued that this approach, which is based on the recent linguistic research, can provide a more efficient NLP system in the long run and will facilitate multilingual computational applications.

2. Persian light verb constructions and machine translation

Persian light verb constructions consist of a preverbal element such as a noun, adjective or adverbial element, that combines with a light or “semantically bleached” verb to form a single predicate in terms of argument structure and semantic interpretation. However, these constructions are very productive in Persian and their components are visible to syntactic processes. In this section, we review some of these dual properties and point out some of the main problems caused by these MWEs for a computational application such as Machine Translation.

2.1 Idiomatic meaning

Most approaches to machine translation involving Persian choose to list the light verb constructions as a unit in the lexicon since it is usually not possible to translate these expressions word for word as shown in the English translation below:

دست داریوش شروع کرد درد گرفتن
word for word translation: hand Dariush beginning did pain catching
correct translation: ‘Dariush’s hand begain hurting.’

2.2 Parsing ambiguity

Determining noun phrase boundaries is a difficult task in Persian since the *ezafe* morpheme linking the noun phrase components is often not written in text, thus giving rise to ambiguity in parsing phrases. This ambiguity can be increased if the LVC is not analyzed as a single unit since the preverbal noun or adjective could be bracketed incorrectly as part of the preceding noun phrase:

[بگفته آسوشیتد پرس] [شمار بیکاران افزایش] یافته است.

Thus to obtain a correct parse, it may be advantageous for the NLP system to list the light verb constructions in the lexicon along with their subcategorization information.

2.3 Lexical proliferation

Listing the LVCs in the lexicon with the appropriate translations and subcategorization information can therefore allow for a more effective parsing and translation results. However, this would need to be done for each language pair and thus will increase the time and cost of development for the NLP system[1]. In addition, a number of verbal predicates will be missed since these constructions are highly productive and can combine with loan or newly coined

[1] There are about 6,000 light verb constructions in a generic computational lexicon of Persian.

words to create new verbs (this is specially prevalent when processing online colloquial documents). Examples found in online corpus sources are: کلیک کردن – ایمیل زدن – سیگنال انداختن.

2.4 Intervening elements

The components of LVCs in Persian are often separated by intervening morphology as shownare: درد نمي گيرد – درد بگيرد – درد نخواهد گرفت.

If the LVC is listed as درد گــرفــت and/or درد گــيــر in the lexicon, the system may not be able to detect the constructions above since they contain intervening imperfective, negation, subjunctive and future morphology. This is specially problematic for linear morphological analyzers such as a two-level morphological approach, although less of an issue for unification-based systems.

Furthermore, the subparts of many Persian LVCs can be separated by syntactic chunks as illustrated below for خــواســتار شــدن, in which a noun phrase and prepositional phrase intervene between the components of the complex verb, yet the verb is to be translated with the single English verb “request”.

کشورهاي اسلامي **خواستار** نقش فزاينده سازمان ملل در عراق **شدند**.
word for word translation: countries islamic **requester** role increasing United Nations in Iraq **became**.
correct translation: “The Islamic countries requested an increasing United Nations role in Iraq.”

2.5 Internal modification

Another instance of the dual (lexical and phrasal) behavior of the Persian light verb constructions can be seen in the example below where the adjective appears on the nominal element and separates the preverb and light verb, yet it modifies the whole verbal event and not just the preverbal noun:

ماني کتک **بدي** خورد.
word for word translation: Mani beating bad-Indef ate.
correct translation: “Mani was beaten badly.” (Lit: Mani ate a bad beating.)

3. A Semantic Template analysis

Recently, a number of researchers working on various subfields of linguistics have begun decomposing verbal predicates in order to analyze the primitive features of meaning that combine to form the verbs. These approaches have taken advantage of the compositionality of these predicates and of the regular alternating patterns across languages to argue for a combinatorial analysis of verbs (cf. Pustejovsky 1995, Levin and Rappaport Hovav 1995, Hale and Keyser 1993, Fong, Fellbaum and Lebeaux 2001).

Starting from a lexical semantic description of event structure, Levin and Rappaport Hovav (1995) develop a system in which the templates can be augmented by the addition of subevent templates thus giving rise to various verbal alternations. Fong et al (2001) extend this system by adopting the notions of primary and secondary predicate templates, where the primary template expresses the core meaning of the verb and the conjoined secondary template represents the secondary predication in syntax. In this system, the formation of verbal predicates are constrained by restrictions on template co-occurrences.

Similar approaches have been taken up in the study of the Persian LVCs in formal linguistics, whereby the researchers have attempted to decompose the complex verbal predicates in order to determine the features that combine to give rise to the distinct aspectual, semantic and syntactic properties of LVCs. More recently, there has also been work on predicting the lexical selection which determines which light verb is to be combined with a given preverbal element (cf. Vahedi-Langrudi 1996, Dabir-Moghaddam 1997, Karimi-Doostan 1997, Haji-Abdolhosseini 2000, Megerdoomian 2001, 2002, Folli et al 2003).

In this section, we discuss how some of these analyses can be used to develop a Semantic Template analysis for Persian verbs.

3.1 Persian LVC templates

Change of state alternation verbs

Change of state verbs are perhaps the single verbal alternation category that has been analyzed most in the literature. The semantic templates that have been proposed for these verbal predicates are shown below, where BECOME and CAUSE represent the primitive semantic features for the change of state and causative verbs, respectively:

Inchoative: y BECOME <state>
Causative: x CAUSE y BECOME <state>

The analysis for these change of state verbs proposed in Megerdoomian (2002) directly parallels the semantic template above, where باز شدن would be analyzed as

y BECOME <باز>

while the causative version باز کردن will be represented as shown, where it is argued that the verb کردن in the sense of 'make' is in fact the causative version of شدن.

x CAUSE-BECOME y <باز>

Activity verbs

A study of the activity verbs (or unergatives) in Persian can shed some light on the semantic template of these predicates. In Fong et al (2001), activity verbs such as 'cry', 'swim' or 'think' are analyzed simply as: x ACT

Persian activity verbs however, such as گریه کردن, شنا کردن, کار کردن or فکر کردن, clearly show a more detailed decomposition. Based on the semantic and syntactic properties of these verbs, it has been argued by Megerdoomian (2002), Folli et al (2003) and Karimi-Doostan (to appear) that the preverbal noun in these instances is in fact a verbal or eventive noun, which we represent by incorporating the primitive semantic feature ACT within the nominal template. The resulting semantic template for these verbal predicates should therefore be represented as the more complex structure illustrated below for گریه کردن. In this analysis, the eventive noun گریه is built in the semantic template by combining the root form گري (shown in angle brackets) with the nominalizing morpheme – ه.

x ACT [ه - [ACT <گري>]^{vp}]^{np}

Thus, Persian LVC properties for activity verbs such as ‘cry’ suggest that the semantic template for these constructions should be more complex than what has currently been suggested on the basis of English verbs.

Instrument verbs

Finally, an analysis of instrument verbs in Persian such as جارو زدن – رنگ زدن – چکش زدن and شانه زدن indicate that these verbs all denote a repetitive action, have agentive subjects and in each instance, the preverbal noun is interpreted as an instrument of the event. The semantic template proposed for these constructions is as follows:

x ACT$_{<\text{REPETITIVE}>}$ y with <instrument>

For Persian, the decomposition between the *instrument* primitive and the main verbal feature is maintained, as exemplified. We assume that زدن is the overt realization of the combination of the features ACT$_{<\text{REPETITIVE}>}$ and with.

x زدن y <چکش>

However, in English, the *instrument* primitive is merged with the verbal feature thus producing the verb ‘to hammer’ or ‘to sweep’.

3.2 Structure of the Lexicon

The semantic template analysis provided suggests that the primitive verbal templates can be represented in a lexicon that contains the underlying and universal semantic and syntactic templates. This ***Template Lexicon*** is in essence an interlingua for verbal representations that all languages could be mapped to. On the other hand, we are proposing a secondary lexicon, named the ***Vocabulary***, which includes the words of the particular language and maps them to the subparts of the semantic-syntactic templates.

For instance, the semantic template for change of state verbs in the *template lexicon* for both Persian and English may consist of the following: x CAUSE y BECOME <state> but the specific *vocabulary* of the languages would differ as shown[2]:

Persian:	شدن=BECOME
	کردن=CAUSE BECOME
	کردن=ACT
English:	open=BECOME <open>
	open=CAUSE BECOME <open>

3.3 Feature structure represenation

In order to implement these proposals within a computational model, we adopt a feature structure representation where the semantic templates as well as the vocabulary items are

[2] Note that this distinction between *lexicon of primitives* and *vocabulary items* has also been proposed in formal linguistics within the framework of Distributed Morphology (Halle and Marantz 1993).

presented as types with defined values. The verbal predicates are then formed through a unification process, by combining the compatible constructions.

The representation of the change of state verbs discussed in the previous section is given below where the main verbal primitives are defined as the EVENT type. As can be seen from this feature structure model, the subevent consists of the BECOME event indicating a change of state; the causative is then formed on top of this smaller predicate by the addition of the CAUSE event which takes the subevent as its predicate. In essence, this is similar to the notion of tree structures used in formal syntax in the analysis of these verbal constructions.

```
┌                                   ┐
│ vP                                │
│ EVENT      cause                  │
│ ARG        x                      │
│ PRED      ┌                     ┐ │
│           │ VP                  │ │
│           │ EVENT      become   │ │
│           │ ARG        y        │ │
│           │ STATE      <state>  │ │
│           └                     ┘ │
└                                   ┘
```

Feature structures are also used to represent the change of state verbs in the *vocabulary* section of the lexicon, as illustrated below for Persian.

[FORM: کـردن
EVENT: caus
PRED.EVENT: become]

[FORM: شـدن
EVENT: become]

A sentential translation will then be obtained as follows: The sentences in the input language (e.g., Persian) are parsed and placed into a feature structure that directly represents the *vocabulary* item for the language. Thus for the sentence نـادر در را بـاز کـرد, the adjective بـاز is represented as the <state: *open*> and the light verb is mapped onto the feature structure containing [EVENT: caus; PRED.EVENT: become]. These feature structures are the partial representations of the *template lexicon* or universal representation. The resulting feature structure is then mapped onto the equivalent target language (e.g., English) component by trying to detect the most compatible feature structure in the English *vocabulary lexicon* using a unification process. For the example given, the feature structure that unifies most directly with the templates provided is the causative verb ‘open’ with the following feature structure:

[FORM: open
EVENT: caus
PRED.EVENT: become
PRED.STATE: open]

4. Conclusion

This paper presents a research proposal for modeling and processing Persian light verb constructions in NLP systems based on a semantic template analysis. This paper clearly describes only the first steps towards the implementation of this approach but we believe that it represents a worthwhile methodology that can provide more efficient multilingual applications in the long run.

In the approach put forth in this paper, the correct translations can be obtained for light verb constructions without needing to create new lexicons for each language pair under study. Furthermore, since each phrasal verb construction does not need to be listed separately in the computational lexicon, the *vocabulary* lexicon will be smaller in size. Instead, verbal predicates will be formed by the various combinations of the primitive elements. For instance, to analyze the change of state alternation verbs, one would only need to list all "state" elements (i.e., adjectives) and the two light verbs of کـردن and شـدن.

The analysis proposed makes use of an interlingual or universal lexical domain called the *template lexicon* which houses the possible combinations of primitive units in the verbal predicates. The main advantage of distinguishing between the *vocabulary* (i.e., language-specific) and *template lexicon* (i.e., universal) components is that the mismatches that arise in current systems between the surface form and the underlying meaning will not be a problem anymore. Hence, complex verbal predicates can be built compositionally, yet the system can also process LVCs that have undergone syntactic operations, are formed with novel usages, or which appear with intervening material.

References

Abeillé, Anne. 1988. Light Verb Constructions and Extraction out of NP in a Tree Adjoining Grammar. In *Papers of the 24th Regional Meeting of the Chicago Linguistics Society.*

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

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

Folli Raffaella, Heidi Harley and Simin Karimi, 2003. Determinants of event type in Persian complex predicates. In Marc Richards, ed., *Cambridge Working Papers in Linguistics*.

Fong, Sandiway, Christiane Fellbaum and David Lebeaux. 2000. Semantic Templates and Transitivity Alternations in the Lexicon. In *Proceedings of TALN 2000*, Lausanne, 16-18 October.

Haji-Abdolhosseini, Mohammad. 2000. Event Types in the Generative Lexicon: Implications for Persian Compound Verbs. In *Toronto Working Papers in Linguistics*, Proceedings of NLS 2000.

Hale, Ken & Samuel Jay Keyser. 1993. On Argument Structure and the Lexical Expression of Syntactic Relations. In *The View from Building 20*, ed. Kenneth Hale and S.Jay Keyser. MIT Press, Cambridge, 53-110.

Halle, Morris & Alec Marantz. 1993. Distributed Morphology and the Pieces of Inflection. In *The View from Building 20*, ed. Kenneth Hale and S.Jay Keyser. MIT Press, Cambridge, 111-176.

Karimi-Doostan, Mohammad-Reza. 1997. *Light Verb Constructions in Persian.* Doctoral dissertation, University of Essex.

Karimi-Doostan, Mohammad-Reza. To appear. Light Verbs and Structural Case. Manuscript, Kurdistan University.

Levin, Beth and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface.* MIT Press.

Mahootian, Shahrzad. 1997. *Persian*. Routledge.

Megerdoomian, Karine. 2001. Event Structure and Complex Predicates in Persian. In *Canadian Journal of* Linguistics 46(1/2):97-125

Megerdoomian, Karine. 2002. *Beyond Words and Phrases: A Unified Theory of Predicate Composition*. Doctoral dissertation, University of Southern California.

Pustejovsky, James. 1995. *The Generative Lexicon.* MIT Press.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, Mexico, pp. 1-15.

Vahedi-Langrudi, Mohammad-Mehdi. 1996. *The Syntax, Semantics and Argument Structure of Complex Predicates in Modern Farsi*.Doctoral dissertation, University of Ottawa.