A corpus study on identification and semantic classification of light verb constructions in Persian: the case of the light verb *xordan* 'to eat/collide'

CrossMark

Ramin Golshaie*

Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran

ARTICLE INFO

ABSTRACT

One of the challenging problems in the domain of Persian light verb constructions (LVCs) is to discover and classify a light verb (LV)'s multiple senses. This is important since productive use of LVs in Persian that leads to formation of novel LVCs can be explained in reference to these established LV senses. On the other hand, identification of LVCs in the first place is another problem which is a complex task given that not only no objective criterion exists for their identification but also the constituent elements of some LVCs can be split by interposing linguistic units that makes their identification difficult. This paper addresses these two issues using corpus methodology. To identify LVCs, the LV *xordan* with two unrelated meanings 'to eat/collide' was chosen for analysis and the corresponding LVCs were extracted from a sampled 50-million-word corpus based on a measure of collocational associations. The extracted LVCs consisted of frequent compositional and idiomatic noun-verb (N-V) patterns found in the corpus. Corpus examinations revealed that frequent compositional N-V sequences have constructional meanings and need to be recognized as LVCs. Finally, to discover the LV senses, 700 concordance lines of the extracted LVCs were studied and classified based on a behavioral profile analysis of their corpus usage patterns. According to the results of behavioral profile analysis, two constructional senses EAT and COLLIDE are coexistent under *xordan* each subsuming their own semantically-related LVCs. The findings while supporting the overall constructionist assumptions on polysemy network of LV senses necessitate a reconsideration of constructionhood criteria in Persian LVCs alongside the process of identification and classification of senses.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Persian light verbs constructions[1] (LVCs)[2] have been the focus of much controversial research in the recent decades. These multiword expressions are composed of a non-verbal element (NV), which is usually a noun or an adjective, and a light verb (LV) which is semantically impoverished compared to its 'heavy' counterpart. In studying Persian LVCs, various aspects

---

  * P.O. Box: 13185-1371, 1090 Enqelab Ave., Tehran, Iran. Tel.: +98 (0)21 66 4949 80.
    E-mail address: golshaie@irandoc.ac.ir.

[1] Also known as 'compound verbs' or 'complex predicates' in the literature.

[2] The following abbreviations have been used in the article: 1/2/3 = person marker; ACC = accusative marker; lit. = literal; LV = light verb; LVC = light verb construction; N = noun; NV = non-verbal element; PAST = past tense; PL = plural; POSS = possessive; PROG = progressive; SG = singular; V = verb. Other abbreviations have been defined in the text or footnote where necessary.

including syntactic properties, separability/non-separability, event structure, productivity, and semantic characteristics of LVCs have been investigated by researchers (e.g. Barjasteh, 1983; Dabir-Moghaddam, 1997; Folli et al., 2005; Haji-Abdolhosseini, 2000; Karimi, 1997; Karimi-Doostan, 1997, 2005, 2011; Megerdoomian, 2001; Mohammad and Karimi, 1992; Vahedi-Langrudi, 1996; Tabaian, 1979). Of these major themes, the topic of productivity (making new LVCs by adjoining a new NV to an already existing LV) continues to be of special interest for researchers. In the recent studies on Persian LVCs (Family, 2006, 2008), productivity is explained in the cognitive-constructional framework. LVCs are considered form-meaning pairings and an LV is said to have different constructions in which specific semantically-related NVs can appear. In other words, according to this approach, each LV in Persian forms networked clusters of polysemous meanings for which specific types of NVs are appropriate.

While the constructionist account seems to provide a solid scaffolding for explaining the productive behavior of Persian LVCs, some challenges still need to be addressed. One major problem is to empirically validate the constructionist account by finding a systematic method for discovering and classifying LV senses in Persian. For example, the LV x*ordan* (studied in this paper) has two unrelated meanings, namely 'to eat' and 'to collide'. These two meanings seem to be equally playing a role in the semantics of resulted LVCs, but in the literature these senses have been lumped together when studying LVCs and no clear semantic classification has been provided as to which sense might be associated with what LVCs. Identifying LVCs from N-V sequences is another challenge given the fact that some N-V sequences are ambiguous between verb phrases and LVCs. Furthermore, NV and LV in most LVCs can be separated by interposing linguistic elements complicating the process of identifying LVCs. It is proposed that these problems can be dealt with in a more effective and systematic way using corpus methods.

Corpus-linguistic methods have been used to study multiword expressions (including LVCs) in variety of languages such as English (Stevenson et al., 2004), Chinese (Huang et al., 2014; Lin et al., 2014), and Urdu (Ahmed, 2010; Ahmed and Butt, 2011). In Persian, except for some studies focused on NLP[3] applications (e.g. Gerdes and Samvelian, 2008; Rasooli et al., 2011), no significant corpus study has been carried out on LVCs. In the present study, corpus-based quantitative methods will be used to identify and semantically classify Persian LVCs containing the LV *xordan*. First, in order to identify and extract LVCs (including separable[4] ones) from corpus, mutual information measure (Church and Hanks, 1990; Stubbs, 1995) will be used as an indicator of collocational associations. It will be shown that frequent compositional LVCs have constructional meanings that distinguish them from phrasal N-V sequences. Second, based on a behavioral profile analysis (Gries, 2006, 2010) of the extracted LVCs, it will be shown that the LV *xordan* simultaneously subsumes two major constructional senses (namely EAT and COLLIDE) each with their own semantically-related LVCs. In the next section, a brief review of the relevant literature is presented. Section 3 introduces the methods, the corpus, and the procedures employed in the extraction and semantic classification of LVCs. Section 4 presents the results of the statistical analysis of the data. Section 5 discusses the main findings in comparison with previous research and implications for reconsidering constructionhood criteria in Persian LVCs. Finally, Section 6 concludes the article by pointing to the main findings of the study.

2. Background

Persian, contrary to languages like English, has fewer than 200 simple verbs (Sadeghi, 1993). For this reason, LVCs in Persian play an important role in expressing various verbal notions inexpressible by simple verbs. Although LVs may have a heavy or full semantic content in other contexts of use, they contribute lesser to the meaning of LVCs (see Jesperson, 1965). For example, compare the heavy and light verb usages of Persian *xordan* ('to eat') in the following sentences:

(1) a. *Maryam   sib   râ   xor-d.*
         Maryam apple ACC eat-PAST.3SG
         'Maryam ate the apple.'

     b. *Maryam   xeili   ghosse   xor-d.*
         Maryam   very   grief   eat-PAST.3SG
         'Maryam grieved very much.'

In sentence (1a), *xord* 'ate' is used as a heavy verb taking *Maryam* as its subject and *sib* 'apple' as its direct object. In sentence (1b), however, *xord* is not used in its heavy meaning. In this example, the LV *xord* together with the NV *ghosse* 'grief' is functioning as an LVC in the sentence which lexically means 'to grieve'. In sentence (1b), the semantic relation between the NV and LV is not transparent compared to the semantic transparency existing between the direct object and the verb in (1a). Cases like (1b) can be considered typical instances of LVCs in Persian since the meaning of the construction is idiomatic or non-compositional.[5]

---

[3] Natural Language Processing.

[4] In separable LVCs, NV and LV can be syntactically separated by intervening constituents.

[5] The terms 'non-compositional', 'idiomatic', 'constructional' and 'unpredictable' have been used interchangeably throughout the article to refer to a meaning that is not the compositional sum of the meanings of its component parts.