A Statistical Approach to Persian Light Verb Constructions
Kim Gerdes, Pollet Samvelian

To cite this version:
halshs-00668547

HAL Id: halshs-00668547
https://halshs.archives-ouvertes.fr/halshs-00668547
Submitted on 9 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A Statistical Approach
to Persian Light Verb Constructions

Kim Gerdes¹, Pollet Samvelian²
Sorbonne Nouvelle

Abstract
This article presents the linguistic bases of Persian light verb constructions and shows the corpus based construction of lists of collocates for some common Persian verbs. The proposed methods of corpus construction are language independent and the good results on a relatively small corpus of 20 million words confirms the power of association measures based on the hypergeometric distribution. The resulting lists show a graduation of lexicalization and the semantic homogeneity of some light verb subcategorization schemes which could be the reason for their wide usage.

Keywords: Corpus, Persian, Light Verb Constructions, Statistics

1. Introduction
Persian light verb constructions have been extensively discussed in the literature, because of their abundance, productivity and particularities, but until recently all quantitative measures have been hampered by the lack of sufficiently large and sufficiently annotated corpora for this language. First, this paper shows why these constructions are particularly interesting for the description of Persian in particular, and for the study of collocation in general. Then we briefly describe the different steps taken to access Persian multiword expressions, using the GrosMoteur software: the construction of the corpus and of the conjugation tables and the choice and application of the association measure. Before concluding, we show and comment some of the abundant results and describe possible usage.

2. Persian Light Verb Constructions
Persian is well known for having very few simple verbs and for making a great use of multiword predicates, referred to as Complex Predicates (CPs) or Light Verb Constructions (LVCs), composed of a verb and a preverbal element. The number of simple verbs can be estimated around 200, but many of them do not display a high frequency of usage and are in competition with an LVC. Furthermore, despite the fact that morphological derivation is still available in Persian for creating new verbs, the verbal lexicon is mainly enriched by LVC creations and not by simple verb formation. This high degree of productivity entails that LVCs are not necessarily lexicalized and can be freely created on the basis of some semantic

¹ ILPGA, LPP (CNRS), Signes (Inria), kim.gerdes@univ-paris3.fr
² ILPGA, Mondes iranien et indien (CNRS), pollet.samvelian@univ-paris3.fr
and syntactic analogy with already lexicalized LVC patterns. Consequently, although the LVC phenomenon is not specific to Persian, the scope, the frequency and the productivity of these constructions in Persian raise some specific problems not necessarily encountered in other languages. This is the reason why Persian LVCs have received a great deal of attention during the last fifteen years in syntactic literature (Karimi-Doostan 1997, Samvelian 2001, Megerdoomian 2002, among others). However the main topic addressed in those studies concerns theoretical and representational issues. Thus, despite lexicographers’ attempts to enrich Persian dictionaries with LVC entries (see for instance Sâdeqi 2000), a comprehensive description and semantic classification of Persian LVCs is still to be done and this task cannot be accomplished without a large scale corpora investigation.

This paper will focus on Noun-Verb combinations. Here is a non-exhaustive list of verbs participating in LVC formation: āmadan ‘to come’, andâxtan ‘to throw, âvardan ‘to bring’, bardâštan ‘to take’, bastan ‘to attach’, ‘to close’, bordan ‘to carry’, dâdan ‘to give’, dâštan ‘to have’, gereftan ‘to take’, ‘to receive’, gozâštan ‘to put’, kardan ‘to do’, ‘to make’, kesidan ‘to pull’, ofîâdan ‘to fall’, raftan ‘to go’, šodan ‘to become’, xordan ‘to collide’, yâftan ‘to find’, zadan ‘to hit’, resîdan ‘to bring’, ‘to make arrive’, resîdan ‘to arrive’. Some of these verbs, e.g. kardan ‘to make’, ‘to do’ or šodan ‘to become’, are prototypical light verbs (LV), in that they have a very poor semantic content and are always used as LVs. Many others, however, have also non-light uses. This is the case of zadan ‘to hit, to fix, to apply’:

\[(1)\] N1 N2 Verb
maryam omid-râ zad
Maryam Omid-DDO HIL.PAST
‘Maryam hit Omid.’

\[(2)\] N1 N2 Prep-N Verb
maryam tâblo-râ be divâr zad
Maryam picture-DDO to wall fix.PAST
‘Maryam fixed the picture on the wall.’

As an LV, zadan occurs in various LVC constructions. In some cases, the valency pattern is identical to those already observed in the above examples, as in (3), but there are also patterns specific to zadan in LVCs, as in (4):

\[(3)\] maryam be omid sili / labxand / telefon zad
Maryam to Omid slap / smile / telephone HIL.PAST
‘Maryam slapped Omid.’; ‘Maryam smiled to Omid.’; ‘Maryam called Omid’

\[(4)\] tâmâšâgar-ân kaf zad-and
spectators-PL palm HIL.PAST-3.PL
‘The spectators applauded.’

Given the fact that it is generally assumed that the noun-verb combination in an LVC is idiomatic, it seems impossible at first sight to predict: first, which nouns combine with zadan and second, which is the appropriate construction when combining a given noun with zadan? Indeed, there are some examples like labxand zadan which satisfy different criteria suggested to identify LVCs and where the noun-verb combination is idiomatic: a) the noun phrase headed by the predicative and the LVC construction are synonymous; b) the subcategorization frame of the LVC is inherited from the noun; c) the first argument of the predicative noun is coreferent with the subject of the LVC. However, a more systematic investigation of the data reveals the two following points:

---

1. The enclitic –râ, which attaches to the right edge of noun phrases, marks the definite and/or specific direct objects in Persian.

27th conference on Lexis and Grammar, L’Aquila, 10-13 September 2008
1) In many cases there seems to be regularities in noun-verb combinations, in the sense that nouns with certain semantic and valence properties preferentially combine with a specific verb. Here are a few examples with zadân:

a) «Scream nouns»: faryâd, jâr, jiq, arbade
b) «Kick nouns»: čâk, kešide, kotak, lagad, sili (kick, slap, beat)
c) «Prejudice nouns»: latme, sadame, zarbe, xesârat (loss, suffering, damage)
d) «Instrument nouns»: bros, jâru, mesvâk, otu, šâne (brush, broom, toothbrush, iron)

This provides evidence in favour of a semi-compositional approach for some Persian LVCs.

2) The nouns occurring in Persian LVCs are not necessarily predicative nouns and many ‘concrete’ nouns participate in LVC formation. This entails not only that Persian LVCs are much more various than LVCs stricto sensu, but also that there is no clear-cut distinction between LVCs and ordinary transitive and intransitive constructions. Compare:

(5) maryâm in wah-râ be kaft-hâ-yašt zad
Maryam this polish-DDO to show-PL-3.SG apply.PAST
‘Maryam applied this polish to her shows.’

(6) maryâm kaft-hâ-yašt-râ wah zad
Maryam show-PL-3.SG-DDO polish hit.PAST
‘Maryam polished her shows.’

These facts suggest that there is a continuum ranging from lexicalized LVCs to non-LVCs and that in intermediate cases, the ‘light’ verb is in fact a “semi” LV.

3. Corpus

As for all corpus work on multiword expressions, corpus based lexicon construction and quantitative measures on LVs are hampered by the lack of sufficiently large, and sufficiently annotated corpora, and the situation is particularly severe for Persian as virtually no free resources of considerable size are available with the notable exception of the Hamshahri Corpus (Darrudi et al. 2004). So why not work directly on the web with a search engine?

1. The main reason is that Persian is a language with relatively free word order and search engines basically offer research of co-occurrence of terms in a page, or adjacent strings. None of these choices can give any interesting results for a free word order language.

2. Moreover, Persian is an inflectional language which has to be accounted for in our statistical measures: it is possible to search individually for each of the nearly 60 forms per verb and the plural, possessive, and Ezafe forms of nouns, but to get relevant measures, one needs to cross all possibilities, unfeasible for any list of verbs.

3. Uncontrolled web research is generally error-prone but the problem is particularly important in Persian as the writing is less normative than many European languages, in particular concerning word spacing.

3.1 Corpus Construction

The present work is done with a specialized web crawler for linguistic corpus research, named «Grosmoteur». The goal of Grosmoteur is to give access to corpora without prior linguistic data and moreover, providing interfaces accessible to the common linguist without special computer training. Grosmoteur is not specialized for the Persian language. So the same procedure can be applied to any language with sufficient Web resources. Grosmoteur, mainly written in Python and Qt, allows for

- Crawling a specific site (or crawling freely with trainable language testing, which has proven useful for other languages and corpus needs)

27th conference on Lexis and Grammar, L’Aquila, 10-13 September 2008
Cleaning and normalizing the text: different file encodings, HTML special characters etc. are standardized into a MySQL database in Unicode encoding and special word corrections (based on regular expressions) can be performed (spelling errors, ...).

Cutting up the text into sentences based on language specific criteria.

Simple searches in the collected data.

Easy Configuration within a graphical user interface.

Preliminary tests showed the lack of normativity of the Persian writing system and we decided to keep to one website only. Although we chose to work on a quite well proof-read corpus, the official government newspaper Keyhan (کیهان in Persian, meaning universe), we still had to normalize the word spacing and unify some common spelling variations. The results are of course biased by this choice of corpus resulting in some "political" cooccurrences like the high scores of collocates like war, America, Israel...

3.2 Corpus Properties

We spidered the complete webpage of Keyhan with their online archive resulting in nearly 6000 different web pages containing 748,966 sentences, 21,266,372 tokens, nearly 800 MB of disk usage including the MySQL indices. The average length of the sentences is 28 tokens and the average number of sentences per page is 130. Our subsequent work is carried out on a transcribed and correctly tokenized version of the corpus, cut up in unique sentences but the example sentences can always been shown in their original version, too.

3.3 Mastering Inflection

Persian has an important verbal inflection of about 60 simple forms per verb. Many of the verbs are partly regular in the sense that they have different past and present radicals. Grosmoteur proposes an automatic morphology learning algorithm that can compute complete morphological paradigms based on a list of affixes (see Goldsmith 2001 for comparable approaches). With a handwritten file containing the Persian conjugation, based on the corpus, it can compute the likeliness of a word being a (e.g. verbal) radical by comparing the distribution to those of a high attestation rate. The user has to cut off the proposed verbs by a threshold. We only kept 125 very common verbs, which was sufficient for our work, but it is easy to build more complete verb lists with more manual interaction. The resulting morphological full form lexicon has about 7000 verbal forms.

---

4 These transformations of the corpus are optional but they allow for easier handling and higher coherence. The Grosmoteur can directly transcribe the corpus in Latin letters, provided a transcription scheme. We choose a non standard simple letter by letter transcription scheme, which is not one to one, so the original spelling cannot be reconstructed directly from the Latin transcriptions (homophones and initial and final letters are transcribed equally). This is a line from the transcription file indicating the Unicode hexadecimal encodings of the letter š: "0634 FEB5 FEB6 FEB7 FEB8" In order to correct tokenization and to split the text into sentences, Grosmoteur makes use of lists of language specific punctuation symbols like "." or ";" and of typical incoherencies in word segmentation.

5 With 5 times 60 lines like "3sgPreSubj: b|bi RAD d" indicating that the 3rd person singular present subjective of one of the five conjugation schemes is formed with either b or bi as a prefix and d as the suffix.

27th conference on Lexis and Grammar, L'Aquila, 10-13 September 2008
All these steps can be controlled from Grosmoteur’s graphical user interface. Grosmoteur also automatically creates an optimized word index giving direct access to all possible occurrences for each of the occurring words, which makes corpus access very fast in spite of its size.

4. Cumulative Hypergeometric Distribution for discovering collocations

We can now compute the collocates of certain verbs, chosen for their frequent use as support verbs, see the table of results on the last page. For a given infinitive and a window extension (in items to its left and right), Grosmoteur can find all occurrences of any of the verb forms, and the system extracts from the database all occurrences of all the forms with their corresponding windows. These joint windows will be our subcorpus. Our goal will be to compute the specificity of this subcorpus compared to the whole corpus.

Given a word with all its surrounding windows (the places where we expect the collocate), which are the words that occur more often than we would expect if the words were distributed evenly? The purely statistical answer is the cumulative hypergeometric distribution (Fisher’s exact test). It avoids all expressive use of stop words or other ad-hoc measures, because it takes into account the total frequency of each word, the frequency of the word in the given windows and the size of the window. So if we stumble upon functional words in our result list, their presence is not an unwanted side effect but has to be explained syntactically, e.g. by a type of construction of the base verb. As Pedersen 1996 remarks, the computation of the hypergeometric distribution is prohibitively heavy and, at the time, was only applicable to small corpora. He shows that the results (at least for the dependent bigrams he's working on) are more reliable on sparse phenomenon like collocations. For our implementation, we developed a very effective computation of the cumulative hypergeometric distribution, with configurable thresholds. For each word in the subcorpus we apply the Fisher test, i.e. we compute the word’s frequency in the subcorpus k and in the whole corpus m, and we compute the mode of the hypergeometric distribution (i.e. of the function giving the probability that the word appeared x times in this subcorpus). If k is bigger than the mode, we compute the sum H of the probabilities of the word appearing k or more times, if k is smaller than the mode, we compute the sum H of the probabilities of the word appearing k or less times. To distinguish probable from improbable events we return a negative result in the second case. Very small probabilities show “abnormal” correlations, positive values mean that the word appeared more than expected, negative values mean that the word appeared less than one might expect from the overall frequency of the word. Our tables show inverted “specificity” values (1/H) (cf. Lebart & Salem 1998) to get smooth results from high positive values to low negative values.

5. Results

As Persian is a verb final language, we chose a window of size 10 preceding the verb. Subsequent work will attempt to investigate the effects of variations of the window size. The table of results shows the first 70 most specific words with andāstan ‘to throw’, gereftan ‘to catch’, zadān ‘to hit’ and xordan ‘to collide’ respectively. For each of these verbs, the majority of the most specific words are composed of nominal elements forming an LVC with the verb. However, some grammatical morphemes such as the definite object marker rā and a
Comments on *zadan* ‘to hit’: The great majority of the words in the provided table are nominal collocates of *zadan* the few exceptions include prepositions used in the grammatical constructions of *zadan*. The following 500 entries, not shown here for lack of space, contain similarly good results. Then noisy results start to flood in. One of the highest ranked nouns is *dast* ‘hand’. The combination of *dast* and *zadan* gives rise to two different LVs, meaning ‘to touch’ and ‘to accomplish’.

The first example is a simple LVC of *dast* and *zadan* and *be* *miz* is the argument of the construction. The second example shows the interesting compound LV formation: the combination of *dast* and *zadan* functions as a LV on its own, taking as a collocate a specific class of predicative nouns (e.g. *tazâhorât* ‘demonstration’). The fact that *dast* is high ranked is mainly due to its use as a compound LV. Indeed, among the other high ranked nouns of the list, some occur only as an argument of *dast zadan*. To obtain the complete list of this specific class, we constrained our window to cases containing both *dast* and *zadan*. The specificity computation of this subwindow provided an extensive list of this semantic class. Here are the first 20 entries: *tazâhorât* ‘demonstration’ , *be* ‘to’, *e’esâh* ‘strike’, *tahasson* ‘sit-in’, *aleyh* ‘against’, *gostarde* ‘ample’, *e’erêz* ‘protestation’, *râhpémâyî* ‘march’, *jenâyat* ‘crime’, *eqdâm* ‘action’, *koštâr* ‘killing’, *keşwersân* ‘countries’, *xowdkoši* ‘suicide’, *čonîn* ‘such’, *talâfî* ‘revenge’, *amaliât* ‘actions’, *xiâbân* ‘street’, *mosku* ‘Moscow’, *bozrog* ‘big’, *gall* ‘assassination’, *mânîov* ‘manoeuvre’, *ahmaqâne* ‘stupid’, *šahr* ‘city’, *šures* ‘rebellion’...

We see that most words belong to semantic class of public or negative actions, the other being part of the grammatical construction or semantically linked (crime, demonstration → street, city ...) or are themselves collocates of the nouns (demonstration → big). The results are interesting for two reasons. First, they enrich the list of Persian LVs with compound LVs, which have never been investigated. Furthermore, they show that although the compound LV *dast zadan* can be translated as ‘to accomplish’, it obviously does not combine with all nouns designating an action, but only with a subset of them. Further investigations on other corpora are necessary in order to get a clear idea about the class of nouns combining with *dast zadan*, however, at this stage of investigation, these nouns roughly fall into three semantic classes:

a) Nouns realizing the abstract concept of ‘action’: *amat* ‘acte’, *eqdâm* ‘action’...
b) Nouns designating public protest or rebellions: *tazâhorât* ‘demonstration’, *šures* ‘riot’...
c) Nouns designating crimes: *jenâyat* ‘murder’, *koštâr* ‘massacre’, *gall-e ām* ‘genocide’...

These facts constitute new evidence in favour of a semi-compositional approach to LVs: speakers do not need to learn all combinations of nouns and LVs extensively. Instead, once they know some cases of well established examples of usage with a more or less regular form-meaning pairing they extend it to new cases judged similar to them.

Comments on *xordan* ‘to collide’: There are two homonymous verbs *xordan* in Persian, meaning ‘to collide’ or ‘to eat’. Consequently, the specificity list contains both nouns forming an LVC with *xordan* and nouns designating comestible things. One of the highest ranked nouns is *šekast* ‘defeat’, which occurs in *šekast xordan* ‘to suffer a defeat’. This is probably a
side-effect of our corpus: a political newspaper devoting several articles to different wars and conflicts in the Middle East. After a list of about a dozen predictive nouns, the most specific food noun is qza ‘dish’, which is a borderline case between free combinations (like ‘eat pizza’) and LVCs like šekast xordan ‘to be defeated’.

6. Conclusion

We have shown how to obtain collocation lists only from large Web corpora with only minimal manual intervention even without prior linguistic resources. The corpus we collected is sufficiently large to provide extensive lists for some commonly used Persian verbs with the associative measure that we applied.

Work is in progress in various directions: 1. We are currently working on completing the morphological dictionaries of Persian, which will be provided as free resources. 2. We will have to diversify our corpus, maybe on blogs or internet forums in order to access more colloquial collocations. 3. We have already started to do reverse computations: Specificity computation of the windows for a given noun provides us with less used LV.

Our goal is the establishment of the productive schemes of Persian LVCs. We expect to contribute to a theoretical account of the gradual lexicalization of Persian LVCs, or more generally to the understanding of the widespread phenomena of collocations. A long term goal will be the elaboration of a linguistically well defined collocation dictionary of Persian in the style of Meľ’cuk’s Explicative and Combinatory Dictionary.

References


27th conference on Lexis and Grammar, L’Aquila, 10-13 September 2008
<table>
<thead>
<tr>
<th>Spec</th>
<th>Word</th>
<th>Spec</th>
<th>Word</th>
<th>Spec</th>
<th>Word</th>
<th>Spec</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
</tbody>
</table>

**Table of results**

<table>
<thead>
<tr>
<th>anzadiam 'to hit'</th>
<th>gérgeriam 'to catch'</th>
<th>azaniam 'to hit'</th>
<th>saraq fi 'to collide'</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
<tr>
<td>1e+16</td>
<td>nilafr</td>
<td>1e+16</td>
<td>nilafr</td>
</tr>
</tbody>
</table>

**27th conference on Lexis and Grammar, L'Aquila, 10-13 September 2008**