Object ordering in Persian: a corpus-based study

Pegah Faghiri, Pollet Samvelian

▶ **To cite this version:**

Pegah Faghiri, Pollet Samvelian. Object ordering in Persian: a corpus-based study. Linguistic Evidence 2013 - Berlin Special , Apr 2013, Berlin, Germany. , 2013. halshs-01429405

HAL Id: halshs-01429405
https://shs.hal.science/halshs-01429405

Submitted on 17 Jan 2017

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.

Object ordering in Persian: a corpus-based study[1]

Pegah Faghiri & Pollet Samvelian
Université Sorbonne Nouvelle

Although a number of hypotheses have been proposed for preverbal object ordering in Persian, none of them have been investigated empirically. In other languages, corpus-based and experimental studies on word order variations, particularly on the relative order of nominal and prepositional objects in the postverbal domain in English, have identified a set of parameters to account for the preferential order: relative length and complexity, givenness, definiteness, pronominality and animacy (Wasow, 2002; Bresnan *et al*, 2007, among others). In line with these studies, this paper presents the first corpus-based study of word order variation between the nominal and the prepositional object in the preverbal domain in Persian, which is head final in the verbal domain (SOV)[2].

Persian is known for its Differential Object Marking (DOM, cf. Lazard, 1982 among others). Roughly speaking, a definite and/or specific DO is always marked with the enclitic *=râ* while an indefinite non-specific DO is not. Moreover, because of the limited number of simplex verbs (about 200), "verbal concepts" are expressed by Complex Predicates (CP), i.e. combinations of a verb and a preverbal element which can be a noun syntactically comparable to an object but which displays some degree of collocationality and semantic idiosyncrasy (cf. Samvelian 2012 among others).

Several studies (Karimi, 2003; Karimi, 2005; Ganjavi, 2007, among others) have related the DO position to DOM. They claim that a) non- *râ*-marked (non-specific) DOs are adjacent to the verb; b) *râ*-marked (specific) DOs can be non-adjacent to the verb (i.e. precede indirect objects). However, these generalizations are mostly based on intuition and construed data. They have mainly been used to support different "deep" syntactic positions for DOs and to account for their "movement" in terms of scrambling. Word order variations *per se* have however never been systematically investigated.

In order to verify these observations empirically we have conducted a corpus study on the Bijankhan corpus (Bijankhan, 2004), a freely available corpus of about 2,6m tokens manually annotated for the POS. We lemmatized inflected verbs, and extracted the potentially ditransitive ones (42.5k tokens, 122 types[3]). Out of a random sample of 2k tokens, we manually identified 541 sentences (82 verb types), in which the verb is preceded by a subcategorized NP and PP with no other constituent in between –regardless whether the verb and its object formed a CP or not.

As a first step, we annotated each sentence for ORDER (NP-PP vs PP-NP) and DOM (*râ*-marked vs non-*râ*-marked).We observed an average preference of 59% for NP-PP order. DOM and order are strongly related ($\chi 2$=258.2552, df=1, p<.001), however while 96% of marked DOs are in NP-PP order only 74% of unmarked DOs show the inverse order. Upon this observation and the fact that unmarked DOs do not constitute a homogenous class, we considered a more fine-grained annotation for unmarked DOs, distinguishing four DO realization types (DOREAL) with respect to their degree of determination: a) Bare noun (BARE , ex. 1), b) Bare modified noun (BAREMOD, ex. 2), c) determined/quantified NP with an indefinite/non-specific reading (UNMRKD-DET, ex. 3) and d) *râ*-marked NP with a definite/specific reading (MARKED, ex. 4).

(1) Maryam ketâb xarid
Maryam book bought
'Maryam bought a book/books.'

(2) Maryam ketâb=e târix xarid
Maryam book=EZ[4] history bought
'Maryam bought a history book/history books.'

(3) Maryam čand ketâb=e qadimi xarid

[1] This study is part of a project on word order effects across languages in the Labex Empirical Foundations of Linguistics; we would like to thank Barbara Hemforth for helpful discussions.

[2] This canonical order can however be easily modified and almost all variations are possible (OVS, SVO, VSO...).

[3] We consider Particle-Verbs as distinct lemmas from the simplex verb (*dar-âvardan* 'to take out' vs. *âvardan* 'to take'), hence a larger number of verbal types in our data.

[4] EZ stands for the *Ezafe*, realized as an enclitic, which links the head noun to its modifiers and to the possessor NP.

Maryam some book=EZ old bought
'Maryam bought some old books.'

(4) Maryam in ketâb=râ xarid
Maryam this book=DOM bought
'Maryam bought this book.'

We observed that DOREAL and ORDER are strongly related ($\chi2$=336.6, df=3, p<.001). However, contrary to the above-mentioned claims relating the DO position to its markedness, we observe that not all unmarked DO types show the same DO position preference: as expected, BARE DOs have a strong preference (90%) for the PP-NP order, but UNMRKD−DET DOs have a clear preference (70%) for the inverse order, thus grouping with marked DOs. Hence, the relevant criteria with regard to DO position seems to be bareness or determination (determined vs. non-determined) rather than definiteness or specificity. In other words, the more an object is determined - *râ*-markedness being the highest degree of determination - the more it is likely to be separated from the verb by the indirect object. DOREAL by itself does already provide a prediction accuracy of 88% for the relative order.

Once the relevance of the DO type in determining the relative order preference between DO and IO established, we tried to determine the relevance of other variables by using logistic regression modeling. So far we have taken into account the logarithmic relative word-length (NP – PP (RELLEN)) and the frequency of the sequence P-N-V or N-V (automatically extracted) in the whole corpus (COLLMEAS). This variable is conceived to provide a measure of the collocational relation between a verb and its complements, identified as highly pertinent to the relative order (Wasow 2002 among others). Automatic extraction was estimated to be superior to manual annotation which is doomed to be subjective and easily influenced by the order.

After verifying that both COLLMEAS and RELLEN are relevant variables (each provided prediction accuracies above the baseline: 69% and 78% respectively), we ran a mixed-effect model including all three variables, and the verbal lemma as random factor. We then used likelihood ratio tests, leaving out one variable each time, to determine if they contribute significantly to the fit. COLLMEAS, which presents a negative coefficient, votes for the PP-NP order as expected, contributes to the fit ($\chi2$=3.93, df=1, p<.05) though improving its prediction accuracy by only 0.18%. RELLEN contributes significantly to the fit ($\chi2$=27.005, df=1, p<.001) and improves its prediction accuracy by 1.5%. It presents a positive coefficient and votes for the NP-PP order when the NP is longer than the PP, and for the inverse when the NP is smaller than the PP. The relative order preference thus conforms to the mirror-image of "short-before-long" principle in VO languages (Hawkins, 2008 among others). Nevertheless, we observe that DOREAL and RELLEN are related in our data and we will have to control for them in order to study their effect independently. We are planning a series of controlled experiments to study the effect of these variables independently along with the effect of other potentially relevant variables such as givenness and animacy.

Bijan Khan, M. (2004). The role of the corpus in writing a grammar: An introduction to a software, *Iranian Journal of Linguistics*, 19(2).

Bresnan, J. , *et al.* (2007). Predicting the dative alternation. In Boume, Kraemer, and Zwarts (eds.) *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science.

Ganjavi, S. (2007). *Direct objects in Persian*, University of Southern California.

Hawkins, J. (2008), An asymmetry between VO and OV languages. In Corbett & Noonan (eds.) *Case and Grammatical Relations: Essays in Honor of Bernard Comrie*. Amsterdam: John Benjamins.

Karimi, S. (2005). *A Minimalist Approach to Scrambling: Evidence from Persian*.qqwqw Studies in Generative Grammar, Mouton De Gruyter.

Karimi, S. (2003). Object positions, specificity and scrambling. In Karimi, S. (ed.) *Word Order and Scrambling*. Blackwell Publishers, 91-125.

Lazard, G. (1982). Le morphème râ en persan et les relations actancielles. *Bulletin de la Société de Linguistique de Paris*, 77(1), 177-208.

Samvelina, P. (2012). *Grammaire des prédicats complexes : les constructions nom-verbe*. Paris : Lavoisier.

Wasow, T. (2002). *Postverbal behavior*. CSLI lecture notes. CSLI.