Accessibility and Word Order:
The case of ditransitive constructions in Persian

Pegah Faghiri[1,2], Pollet Samvelian[1], and Barbara Hemforth[2]

[1]Université Sorbonne Nouvelle - Mondes iranien et indien (CNRS)

[2]Université Paris Diderot - Laboratoire de la Linguistique Formelle (CNRS)

Introduction

In this paper, we present a corpus-based and experimental study of the preferred ordering of complements in ditransitive constructions of Persian and propose an account based on the (referential) *Givenness Hierarchy* (Gundel *et al*. 1993). We furthermore show that Persian data help tease apart two competing theories that are proposed to explain the "long-before-short" preference in OV languages. Namely, Hawkins's (1994) *Early Immediate Constituent* (EIC) principle, which elegantly accounts for mirror-image preferences in OV and VO languages, fails to explain Persian data, while, Yamahista and Chang's (2001) approach, which claims that sentence production is sensitive to language-specific features, provides a valid account of our Persian data. Moreover, the latter can be directly integrated into our analysis based on the referential givenness hierarchy and hence allows to account for the preferential ordering of the preverbal complements by a unified principle, that is, through the conceptual accessibility.

The preferred order of constituents, including in ditransitive constructions, has been a focus of interest in empirical and experimental linguistics, which have shown that word order is independently influenced by relative length, givenness, animacy, collocationality, verbal lemma, etc. (see e.g Wasow 1997, Arnold et al. 2001, Bresnan 2007). These findings, mainly built on data from English and other Germanic languages, have reinforced most prevailing accounts of sentence production, which attribute constituent order preferences to accessibility-based incremental production (e.g. Bock 1982): More accessible constituents (that is, given and short vs. new and long) tend to be produced earlier in the sentence. However, as highlighted by Yamashita and Chang (2001), longer constituents have conflicting properties. While they are less accessible in the formal arena, they are lexically richer hence more salient and more accessible in the conceptual arena. In English (Stallings et al. 1998, Arnold et al. 2001), French (Thuilier 2012) and German (Behaghel 1909), sentence production seems to be more sensitive to formal factors, hence the "short-before-long" preference. Japanese, on the other hand, is more sensitive to conceptual factors, hence the "long-before-short" preference.

Both processing-oriented and production-oriented accounts have been proposed to explain these opposing tendencies. On one hand, Hawkins's EIC, a distance-minimizing dependency-based principle grounded in parsing constraints and sensitive to the direction of the head, correctly predicts mirror-image tendencies in (strictly) head-initial vs. (strictly) head-final languages. On the other hand, Yamashita and Chang (2001) propose a production-oriented account which takes into account language-specific features, arguing that since both conceptual and formal factors have been found to influence word order preferences (Bock 1982), the sensitivity of the production system to these factors can be viewed as being language-specific (see also Chang 2009). English has a fairly fixed word order and requires all arguments to be overtly realized. Moreover, the ordering happens in the postverbal domain, where it is shown that the verb exerts a strong influence (see Stallings et al. 1998). The authors claim that because of the syntactic rigidity of English, speakers are more sensitive to formal factors and consequently prefer to

postpose longer constituents, which are formally less accessible. Japanese, on the contrary, has a fairly free word order and does not require all arguments to be overtly realized. Moreover, the ordering occurs in the preverbal domain, hence speakers, more sensitive to conceptual factors, prefer to put longer constituents before shorter ones.

Studying word order preferences in Persian is of great interest in this debate, since Persian, like Japanese, is a *pro*-drop OV language with a fairly free word order. However, unlike Japanese, it has mixed head-direction behavior. In this paper, we study the relative order of the direct (DO) and the indirect (IO) objects in the preverbal domain in Persian. Section 2 presents the basic properties of Persian relevant for the issue at stake and the existing hypotheses on word order preferences. Our corpus and experimental data are briefly presented in Sections 3 and 4. We discuss our results in Section 5.

2. An overview of Persian

2.1 Head-direction and Word order

Persian exhibits a mixed head-direction behavior: All phrasal categories other than the VP, that is, NP, PP, and CP are head-initial as illustrated by (1). Furthermore, while the canonical word order is SOV, all other variations are also possible, namely, the dependents of a verb (subject, DO and IO) can occur in a post-verbal position. Note however that written Persian is more conservative with respect to the canonical order. The focus of our study is the (S)OV order. Note that our corpus data are extracted from a written corpus where the word order variations are expected to be limited and the canonical SOV order to be dominant.

(1) dar in ketāb=e jāleb ke diruz xānd-am
in this book= EZ[1] interesting that yesterday read-1SG
'In this interesting book that I read yesterday'

2.2 *NPs in the DO position*

In formal Persian, there is no overt marker for definiteness, only indefiniteness is overtly marked by the enclitic *–i*, as in (4), by the cardinal *yek* 'one' or by both, as in (5). Furthermore, Persian has what Corbett (2002) calls a general number, expressed by the singular form. Consequently, a bare noun, or a bare modified noun for that matter, is not specified for number and can have a mass reading whether a mass noun or not, as in (2) and (3). An indefinite NP, on the other hand, whether formed by a quantifier or by the indefinite marker, is always specified for number. Moreover, Persian displays Differential Object Marking (DOM), triggered by definiteness (roughly) (Lazard 1982) and realized by the enclitic *-rā.* A definite and/or specific DO is always marked, as in (6). Consequently, a non *rā*-marked DO has necessarily a non-definite and/or non-specific reading (2-5). Note that the enclitic *-rā* can also act as a topicalizer for other non-subject constituents (see e.g. Dabir-Moghadam 1992)

On the basis of these facts, the following hierarchy based on increasing degree of determination can be traced for NPs in the DO position: Bare < Bare modified < Indefinite < *rā*-Marked. Furthermore, we can assume that on the (referential) Givenness Hierarchy, *rā*-markedness corresponds to the highest status (from "uniquely identifiable" to "in focus") and bareness to the "type identifiable" status, that is, the lowest degree of givenness on the hierarchy.

(2) Sārā be Nimā ketāb dād Bare
Sara to Nima book gave
'Sara gave a book/some books to Nima.'

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

(3) Sārā be Nimā ketāb=e tārix dād Bare-modified
Sara to Nima book=EZ history gave
'Sara gave a history book/some history books to Nima.'

(4) Sārā ketāb=i be Nimā dād Indefinite
Sara book=INDEF to Nima gave
'Sara gave a book to Nima.'

(5) Sārā yek ketāb(=i) be Nimā dād Indefinite
Sara a book(=INDEF) to Nima gave
'Sara gave a book to Nima'

(6) Sārā (in) ketāb=rā be Nimā dād Marked
Sara (this) book=DOM to Nima gave
'Sara gave (this/) the book to Nima.'

2.3. *The DOM criterion*

It is generally assumed by Persian grammars as well as by more recent studies in the generative framework (see e.g. Karimi 2003) that DOM determines the (unmarked) position of the DO: A marked DO can be separated from the verb, while an unmarked DO should be adjacent to it. However, this claim has remained mostly theoretical and lacks systematic empirical underpinning.

3. The corpus study

We conducted a study on the Bijankhan corpus[2], a freely available corpus of about 2.6m tokens gathered from daily news, manually annotated for POS. Our dataset contained 905 sentences of DO-IO-V or IO-DO-V patterns, identified manually out of a semi-random sample extracted from the corpus. We observe that marked DOs, as predicted by the DOM criterion, have a very strong preference (95%) to be separated from the verb. However, unmarked DOs display more variation: Bare DOs have a strong preference (84%) for adjacency but bare modified DOs have a more moderate preference (67%) for this position and, surprisingly, indefinite DOs have a clear preference (77%) for the opposite order. As illustrated by Figure 1, the position of the DO is strongly related to its degree of determination, that is, to its position on the hierarchy established previously: The higher the DO on the hierarchy, the more likely it is to precede the IO.

Moreover, mixed-model analyses showed that the relative length has a significant effect ($p<.01$) on the preferential order in the case of intermediate DOs, namely indefinite and bare modified DOs, corresponding to a long-before-short preference (see Figure 2)[3].

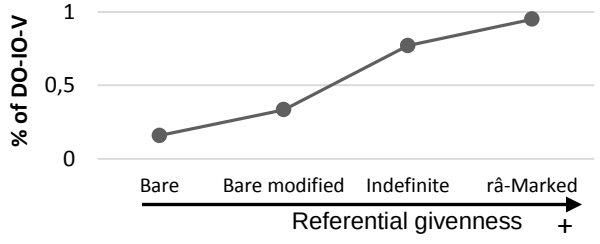


Figure 1: Degree of determination of the DO

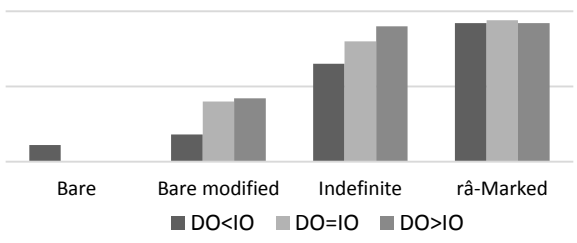


Figure 2: Relative Length

[2] The corpus was created in 2005 at the University of Tehran (http://ece.ut.ac.ir/dbrg/bijankhan/).
[3] For more details on the constitution of the dataset and the mixed-effect model see Faghiri & Samvelian (to appear).

4. The experimental study

Our corpus analysis contradicts existing claims regarding indefinite (unmarked) DOs. To verify our findings in a controlled experiment, we ran a web-based questionnaire, only with indefinite DOs, where we systematically varied relative length and givenness of the IO following a 2x2 design. We used a sentence completion task, where participants were given the choice of two DOs[4] and an IO to complete a preamble, see the English equivalent of a sample item in (7). 33 native speakers of Persian completed 20 target items interspersed with 30 fillers. It should be noted that consist with our corpus data, all items were in written (formal) register.

The experimental data were consistent with the corpus data. Mixed model analyses confirm the general preference (69%) of indefinite DOs for the DO-IO-V order as well as the significant effect of the relative length ($p<.001$) corresponding to the long-before-short preference (see Figure 3). The givenness of the IO did not turn out significant, however, there was a marginal interaction between the two variables ($p<.1$).

(7) Sara is a nice woman (*given IO:* with a niece she loves)
When her husband died, she gave…
[an apartment] [to her niece who recently had a baby]
[an apartment with 5 rooms] [to her niece]
2nd DO: [short: a necklace (*long:* with beautiful pearls)]

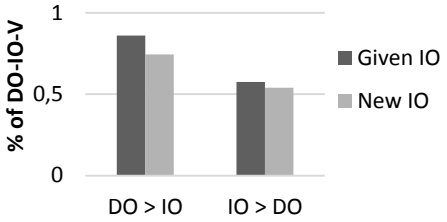
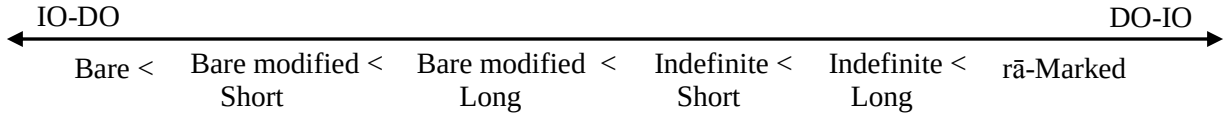


Figure 3: Relative Length and Givenness

5. General discussion

Our corpus and experimental data call into question the empirical validity of the existing claims on the position of the DO in Persian. Thus, the degree of determination, rather than markedness, turns out to be the key to the preferred position of the DO. Moreover, this revision is insightful in so far as it allows to develop a more fine-grained and gradual view of the influence of the accessibility of the DO and its position. We can trace a continuum, inspired by the Gundel's (1993) Givenness Hierarchy and in terms of the accessibility of the DO, to explain these ordering preferences. Starting from one end (high accessibility), exemplified by the very strong preference of *rā*-marked DOs to be separated from the verb, to the other end (low accessibility), that is the very strong preference of the bare DOs to be adjacent to the verb. Note that this tendency is compatible with the hierarchy of the grammatical roles (see Keenen and Comrie 1977). Interestingly, the long-before-short preference observed for the intermediate DOs can be integrated into this continuum, in line with the analysis provided by Yamashita & Chang (2001): Longer constituents gain in conceptual accessibility and Persian speakers are sensitive to conceptual factors more than to formal ones.



As for the long-before-short preference, note that Hawkins's processing-oriented EIC principle which provides correct predictions for strictly head-final or head-initial language, does not predict any length based preferred order for Persian, for example, in the case of indefinite DOs, as illustrated by (7).

[4] The idea of having an additional DO is to obtain ordering preference in a more indirect manner. The two DOs share formal properties and only differ lexically, hence the choice of one or the other is equal for us.

(7) [$_{VP}$[$_{NP}$ yek tup=e tenis=e no] [$_{PP}$ be Ali] dād vs. [$_{VP}$ [$_{PP}$ be Ali] [$_{NP}$ yek tup=e tenis=e no] dād
1 2 3 4 5 6 7 = 1 2 3 4 5 6 7
a ball=EZ tennis=EZ new to Ali gave
'(S)he gave Ali a new tennis ball.'

6. Conclusion

Beyond the interest of the data presented in this study for the empirical verification of the existing theoretical hypothesis for Persian, studying the word order preferences, in a mixed head-direction OV language, revealed to be cross-linguistically interesting: 1) Persian data confirm the long-before-short preference attested in other OV languages. 2) This preference is not predicted by Hawkins's (1994) processing-oriented principle, which provides correct predictions for English and Japanese. 3) It is predicted by the accessibility-based sentence production account proposed by Yamashita and Chang (2001) on the basis of the greater sensitivity of constituent ordering in OV languages to conceptual factors rather than to the formal ones (in contrast to VO languages).

Reference

Behaghel, Otto. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25:110-142.

Bock, J Kathryn. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 89:1.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, ed. G. Boume, I. Kraemer, and J. Zwarts, 69–94.

Chang, Franklin (2009), Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English, *Journal of memory and language*, 61:374-397.

Corbett, Greville G. (2000). Number. Cambridge University Press.

Dabir-Moghaddam, Mohammad. (1992). On the (in) dependence of syntax and pragmatics: Evidence from the postposition *-ra* in Persian. Cooperating with written texts 549–573.

Faghiri, Pegah, and Pollet Samvelian (to appear). Constituent ordering in Persian and the relative length, in Piñón, Ch. (ed.) *Empirical Issues in Syntax and Semantics 10 (EISS 10).*

Gundel, Jeanette K, Nancy Hedberg, & Ron Zacharski. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69 (2) : 274–307.

Hawkins, John A (1994), *A performance theory of order and constituency*. Cambridge University Press.

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

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

Keenan, Edward. L., and Comrie, Bernard (1977). Noun phrase accessibility and universal grammar. *Linguistic inquiry*, 8(1), 63-99.

Stallings, Lynne M., Padraig G. O'seaghdha, and Maryellen C. MacDonald. (1998), Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy NP shift. *Journal of Memory and Language*, *39*(3), 392-417.

Wasow, Thomas (1997). Remarks on grammatical weight. *Language Variation and Change*, 9(01):81–105.

Yamashita, Hiroko, and Chang, Franklin (2001), 'Long before short' preferences in the production of a head final language, *Cognition* 81(2): B45-B55.