Constituent Ordering in Persian and the Weight Factor

Pegah Faghiri
Pollet Samvelian

Studies on constituent ordering have pointed out the tendency to postpose heavy constituents. However, head-final languages seem to display the mirror-image tendency. In this paper, we present corpus data on the relative order between the direct object (DO) and the indirect object (IO) in Persian, an SOV language. Our study shows a similar effect in Persian; however, relative length plays a secondary role, since the position of the DO mainly depends on its degree of determination.

Keywords: word order, heaviness, differential object marking (DOM), givenness, Persian

1 Introduction

The "end-weight" principle in constituent-ordering preferences was first proposed by Behaghel (1909) based on observations of German. More recently, several studies, mainly on English, have highlighted the tendency to postpone heavy constituents (e.g. Wasow 1997, Stallings et al. 1998, Arnold et al. 2000). This weight effect is either accounted for in terms of processing or in terms of planning and production. Incremental models of sentence production (e.g. Bock and Levelt 1994, Garrett 1980, Kempen and Harbusch 2003) claim that the ordering of constituents depends on their required processing time. Short simple constituents can be processed and formulated faster and thus become available for production sooner than long and/or complex ones. Since this explanation is grounded in general principles of cognition, it has sometimes been suggested that the "short-before-long" principle is universal. However, investigations on some (strictly) head-final languages have undermined the (inferred) universality of this preference. The opposite tendency has been reported for Japanese (Hawkins 1994, Yamashita and Chang 2001) and Korean (Choi 2007).

Based on extensive data from typologically different languages, Hawkins (1994, 2004) highlights an asymmetry between VO and OV languages. The latter display the mirror-image tendency, placing long constituents before shorter ones. Hawkins proposes a theory of word-order preferences in terms of processing constraints which is sensitive to the direction of the head and consequently correctly predicates the asymmetry between strictly head-initial and head-final languages. Yamashita and Chang (2001, 2006), on the other hand, provide a production-oriented account for the "long-before-short" preference in Japanese. They revisit the availability-based account of ordering preferences in sentence production highlighting the necessity to consider language-specific features.

In this study we investigate the relative order between the direct object (DO) and the indi-

This work is supported by a public grant funded by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083). We are grateful to Barbara Hemforth for helpful discussions. We would also like to thank the audience at CSSP 2013 (Université Paris-Diderot) for their comments, as well as Christopher Pinon and an anonymous reviewer for reading through the first version of this manuscript and making valuable suggestions.

Empirical Issues in Syntax and Semantics 10, ed. Christopher Piñón, 215–232
http://www.cssp.cnrs.fr/eiss10/

rect object (IO) in the preverbal domain in Persian. Data from Persian is of special interest for the issue at stake, since Persian is an SOV language, but, contrary to Japanese, it is not strictly head-final. It is largely assumed that in Persian, the position of the direct object depends on its markedness and relative length or heaviness have never been mentioned to be relevant. Meanwhile, no systematic data-driven study on the subject has ever been conducted to support this hypothesis.

The remainder of this paper is organized as follows. In the next section, we present an overview of Persian focusing on properties relevant for this study, and in section 3, the existing hypothesis on the position of direct object. Our corpus study is presented in section 4. We present available accounts of "long-before-short" in OV languages in section 5, and in section 6 our account of the data.

2 An Overview of Persian

2.1 Word Order

Persian exhibits mixed behavior with regards to head-direction. The unmarked (neutral or canonical) word order is uncontroversially SOV. Meanwhile, all phrasal categories (other than the VP), namely, NP, PP, and CP are head-initial, as illustrated by (1). Even the verbal domain is not strictly head-final. Clausal complements are strictly postverbal, as in (2), and goal arguments are systematically postverbal in colloquial speech, as in (3).

(1) dar in ketāb=e jāleb ke diruz xānd-am
in this book=EZ[1,2] interesting that yesterday read-1SG
'In this interesting book that I read yesterday.'

(2) (man) goft-am (ke) in ketāb jāleb ast
(I) said-1SG (that) this book interesting is
'I said that this book is interesting.'

(3) (mā) diruz raft-im sinema.
(we) yesterday went-1PL movies
'Yesterday, we went to the movies.'

While SOV is the canonical order, all other variations are possible. Although the written language is conservative with regards to the canonical SOV order, the colloquial register exhibits a fair amount of variation. It should be noted, however, that these variations are not all equally frequent and some imply a special prosody. In this study, we only focus on verb-final constructions.

2.2 Persian NPs

As mentioned previously, the relative order of objects in Persian has generally been linked to the differential object marking (DOM) (see section 2.3 below), which in turn is related to definiteness and/or specificity. This section provides an overview of Persian NPs in this respect.

[1]Glosses follow the Leipzig Glossing Rules (www.eva.mpg.de/lingua/resources/glossing-rules. php). The following non-standard abbreviations are used for clarity: DOM = differential object marking; EZ = Ezafe; RESTR = Restrictive.

[2]The *Ezafe*, realized as an enclitic, links the head noun to its modifiers and to the possessor NP (see Samvelian 2007).

In formal Persian there is no overt marker for definiteness; only indefiniteness is marked.[3] Furthermore, Persian has what Corbett (2000) calls a *general number*, expressed by the singular form. This means that in Persian the number is not specified for a bare singular noun. These properties have some bearings on the readings of NPs. In the remainder of this section, we will discuss the following NP types: bare nouns, bare-modified, indefinite/quantified and definite NPs.

It should be noted that since definiteness is not overtly marked, bare singular nouns, that is, nouns occurring alone in their bare singular form with no (overt) determiner or quantifier, may correspond to two different types of NPs, either a definite and/or an anaphoric NP, as in (4) and (5), or a bare noun, that is, a noun without any determination or quantification. By "bare noun" we only refer to the latter. As we will see in section 2.3, this possibility is excluded in the DO position, where only the bare noun reading is licensed for bare singular nouns.

(4) xoršid dar āsemān mi-deraxš-ad
sun in sky IPFV-shine-3SG
'The sun shines in the sky.'

(5) gorg zuze mi-kešid
wolf howl IPFV-pulled
'The wolf was howling.'

2.2.1 Bare Nouns Bare nouns are non-specified for number and have a nonspecific reading, which can be generic, as in (6), as well as existential (contra Karimi 2003), as in (7).

(6) gorg yek heyvān=e vahši va darande ast
wolf a animal=EZ wild and predator is
'The wolf is a wild and predator animal.'

(7) Maryam ketāb xarid
Maryam book bought
'Maryam bought a book/some books.'

Note that, contrary to Karimi's (2003:96–97) claim, bare nouns can introduce a discourse referent in Persian, which uncontroversially implies that they can receive an existential reading (Karttunen 1976), as illustrated by (8) (see Samvelian 2001 for a detailed discussion).

(8) (man) māšin dār-am vali tormoz=aš xarāb ast
(I) car have-1SG but brake=3SG broken is
'I have a car but its brake is broken.'

2.2.2 Bare-modified Nouns These nouns only differ from bare nouns by the presence of a (restrictive) modifier, as in (9) and (10), and have the same readings as the latter.

(9) ketāb=e qadimi nāyāb ast
book=EZ old rare is
'Old books are rare.'

[3]There is a suffix in the colloquial register which marks a noun as being discourse-given, which we present briefly when discussing definite NPs, see section 2.2.4.

(10) Maryam ketāb=e še'r xarid
Maryam book=EZ poetry bought
'Maryam bought a poetry book/some poetry books.'

2.2.3 Indefinite NPs These NPs can have either a specific or a nonspecific existential reading. In the DO position the two readings will be differentiated by DOM (see section 2.3). Contrary to bare nouns, indefinite NPs are always specified for number.

Indefiniteness is overtly marked in Persian. It can be realized by the enclitic *=i*, as in (11a), by the cardinal *ye(k)*[4] 'one', as in (11b), or by the combination of these two determiners, as in (11c).[5] It should be noted that these markers are not always interchangeable (see Ghomeshi 2003).

Indefinite NPs are also formed by numerals or other indefinite quantifiers, as in (12). In this case, the noun remains in the singular form, even when the NP denotes more than one entity, and it cannot take *=i*.

(11) a. gorg=i zuze mi-kešid
wolf=INDF howl IPFV-pulled
b. yek gorg zuze mi-kešid
a wolf howl IPFV-pulled
c. yek gorg=i zuze mi-kešid
a wolf=INDF howl IPFV-pulled
'A (certain) wolf was howling.'

(12) čand(=tā)/se(=tā) gorg zuze mi-kešid-and
few(=CLF)/three(=CLF) wolf howl IPFV-pulled-3PL
'A few/three wolves were howling.'

2.2.4 Definite NPs Definite NPs can either be formed by different definite determiners, like demonstratives, or by no overt determiner, as in (13). Furthermore, bare plural nouns[6] generally trigger a definite reading,[7] as in (14). Note, however, that the plural marking is not incompatible with the indefinite determination *=i* or *yek*, as in (15) (for a discussion of plural marking and definiteness, see Ghomeshi 2003).

[4]Pronounced *ye* in colloquial speech. We will use the formal form throughout this article.

[5]The use of the enclitic alone is restricted to the formal language.

[6]Persian disposes of several nominal plural suffixes, among them the suffix *-(h)ā* is universal and can systematically be added to any noun to form a plural (for a review of the nominal plural marking see Lazard et al. 2006 and Faghiri 2010, among others).

[7]Note that the combination of a numeral/quantifier and the plural form triggers a definite or a partitive reading, as in (i) and (ii), respectively.

(i) se=tā ketāb-hā gom šod-and
three=CLF book-PL lost became-3PL
'The three books were lost.'

(ii) čand=tā/se=tā az ketāb-hā gom šod-and
few=CLF/three=CLF of book-PL lost became-3PL
'A few/three of the books were lost.'

(13) (in) šiše emruz šekast
(this) glass today broke
'This/the glass broke today.'

(14) šiše-hā emruz šekast-and
glass-PL today broke-3PL
'The (*Some) glasses broke today.'

(15) yek ketāb-hā=i heyn=e asbābkeši gom šod-and
a book-PL=INDF during=EZ move lost became-3PL
'Some (of the) books get lost during the move.'

It should be noted that colloquial speech displays a definite suffix, realized as *-(h)e*, which marks a noun as being discourse-given or anaphoric, for example, *gorbe-he* 'the cat'. Since the data used in this study are limited to the written language, where this suffix is not likely to appear, we will not discuss it any further.

2.3 Differential Object Marking

Persian displays differential object marking (DOM),[8] realized by the enclitic *=rā*. Definite and/or specific direct objects are necessarily *rā*-marked.[9] Consequently, non-*rā*-marked direct objects receive an indefinite nonspecific reading, as in (16). DOM is not incompatible with the indefinite determination, as in (17). An indefinite NP like *ketāb=i* when *rā*-marked will receive a specific reading.

(16) Maryam ketāb=rā xarid *vs.* Maryam ketāb xarid
Maryam book=DOM bought Maryam book bought
'Maryam bought the book.' *vs.* 'Maryam bought a book/some books.'

(17) Maryam ketāb=i=rā xarid
Maryam book=INDEF=DOM bought
'Maryam bought a (specific) book.'

It should be noted that the use of the enclitic *=rā* is not limited to DOM. *Rā* is also used as a topicalizer for other non-subject functions, as illustrated by as in (18). Meanwhile, a more detailed discussion is beyond the scope of the present study (for further discussions see Lazard 1982, Meunier and Samvelian 1997, Dabir-Moghaddam 1992, among others).

(18) emruz=rā dars mi-xān-am
today=DOM lesson IPFV-read-1SG
'As for today, I (will) study.'

Note that *=rā* is a phrasal affix and is placed on the right edge of the NP, as in (19). Meanwhile, when the head noun is modified by a relative clause, *=rā* is either placed on the head noun, as in (20a), or on the right edge of the clause, as in (20b). The norm, however, states that it should be

[8]This designation coined by Bossong (1985) denotes the property of some languages with overt case-marking of direct objects to mark some objects, but not others, depending on semantic and pragmatic features of the object; see also Aissen (2003).

[9]In colloquial speech *=rā* is realized as *=(r)o*. We use the formal form throughout this paper for the ease of reading and also in coherence with our data, which are extracted from a written corpus.

placed as close to the head as possible. Due to the availability of two positions, double marking marginally happens, as in (20c).

(19) [ketāb=e dastur=e zabān=e fārsi=ye čāp=e jadid]=**rā** xarid-am
book=EZ grammar=EZ language=EZ Persian=EZ edition=EZ new=DOM bought-1sg
'I bought the last edition of (the book of) the Persian Grammar.'

(20) a. [ketāb=i[10]=**rā** ke ru=ye miz bud] xānd-am
book=RESTR=DOM that on=EZ table was read-1sg
b. [ketāb=i ke ru=ye miz bud]=**rā** xānd-am
book=RESTR that on=EZ table was=DOM read-1sg
c. [ketāb=i=**rā** ke ru=ye miz bud]=**rā** xānd-am
book=RESTR=DOM that on=EZ table was=DOM read-1sg
'I read the book that was on the table.'

2.4 Complex Predicates

Persian has a limited number of simplex verbs, around 250, half of which are currently used by the speech community. The verbal lexicon mainly consists of syntactic combinations, called "complex predicates", also known as Compound Verbs or Light Verb Constructions, including a verb and a non-verbal element, for example, a noun, as in *bāzi kardan* 'to play' (lit. 'play do'), an adjective, as in *derāz kešidan* 'to lay down' (lit. 'long pull'), a particle, as in *bar dāštan* 'to take' (lit. 'PARTICLE have'), or a prepositional phrase, as in *az dast dādan* 'to loose' (lit. 'of hand give'). New "verbal concepts" are regularly coined as complex predicates rather than simplex verbs (see Samvelian 2012, Samvelian and Faghiri 2013, Samvelian and Faghiri 2014, among many others).

Although, Persian complex predicates are multiword expressions and thus display some lexical properties such as lexicalization, they display all properties of syntactic combinations, including some degree of semantic compositionality. Hence, as Samvelian (2001, 2012) extensively argues, it is impossible to establish a clearcut distinction between (prep-)noun-verb complex predicates and "ordinary" object-verb combinations. In other words, the differentiation is better reflected by a continuum from highly lexicalized complex predicates to ordinary complement-verb combinations rather than a categorical distinction.

3 The Position of the Direct Object

Several theoretical studies, mainly in the generative framework, have established a link between the position of the direct object and its specificity (e.g. Karimi 2003, Rasekhmahand 2004). Following Karimi's (2003) work in the minimalist framework, two different syntactic positions have generally been assumed for the DO depending on its specificity.[11]

[10]Persian grammars generally establish two distinct determiners *=i* in Persian. One is the indefinite determiner discussed in section 2.2.3. The other one, which occurs exclusively with restrictive relatives, is analyzed as a 'demonstrative' or 'definite' article (Lazard et al. 2006).

[11]The two positions assumed by Karimi (2003:105) are:

(i) a. $[_{VP}\ DP_{[+Specific]}\ [_{V'}\ PP\ V]]$
b. $[_{VP}\ [_{V'}\ PP\ [_{V'}\ DP_{[-Specific]}\ V]]]$

(21) a. Kimea aqlab barā mā še'r mi-xun-e (Karimi 2003:91–92)
Kimea often for us poem IPFV-read-3SG
'It is often the case that Kimea reads poetry for us.'
b. Kimea aqlab barā mā ye še'r az Hafez mi-xun-e
Kimea often for us a poem from Hafez IPFV-read-3SG
'It is often the case that Kimea reads a poem by Hafez for us.'
c. Kimea aqlab hame=ye še'r-ā=ye tāza=š=**ro** barā mā mi-xun-e
Kimea often all=EZ poem-PL=EZ new=3SG=DOM for us IPFV-read-3SG
'It is often the case that Kimea reads all her new poems for us.'
d. Kimea aqlab ye še'r az Hafez=**ro** barā mā mi-xun-e
Kimea often a poem from Hafez=DOM for us IPFV-read-3SG
'It is often the case that Kimea reads a (particular) poem by Hafez for us.'

In a neutral word order, nonspecific DOs are adjacent to the verb, as in (21a) and (21b), while specific DOs precede the indirect object, as in (21c) and (21d). Since specificity triggers *rā*-marking, this means that unmarked DOs occur adjacent to the verb while marked DOs do not. Hereafter, we refer to this hypothesis as the *DOM criterion*.

(22) The DOM Criterion
In a neutral word order, *rā*-marked DOs occur separated from the verb while unmarked DOs occur adjacent to the verb.

Furthermore, it is assumed that a nonspecific DO can be separated from the verb, that is, can undergo scrambling, only if it has a contrastive focus. The scrambling of specific objects, on the other hand, is less constrained, since they can additionally be topicalized.[12]

Grammarians have also formulated generalizations about the canonical position of the DO, which are mostly in accordance with the DOM criterion. However, some additionally establish a distinction between unmarked DOs, depending upon the presence of the indefinite determiner *-i*. Givi Ahmadi and Anvari (1995:305), for instance, state that *rā*-marked DOs should precede the IO, unmarked DOs should follow the IO, and *i*-marked (non *rā*-marked) DOs can either follow or precede the IO, as in (23).

(23) a. Yusef ketāb=rā be ketābxāne dād
Yusef book=DOM to library gave
'Yusef gave the book to the library.'
b. Yusef az ketābxāne ketāb gereft
Yusef from library book took
'Yusef took a book/some books from the library.'
c. Yusef ketāb=i az ketābxāne gereft *or* Yusef az ketābxāne ketāb=i gereft
Yusef book=INDEF from library took
'Yusef took a book from the library.'

It should be noted that these hypotheses remain theoretical and, to our knowledge, no systematic empirical verifications have ever been conducted. We have conducted a corpus-based study to investigate their validity and to study the factors that determine the preferential word

[12]Karimi (2003:106–111) assumes that discourse functions trigger movement in Persian and the landing site of a scrambled object is the specifier of a functional head, such as Topic or Focus.

order in ditransitive constructions in line with Arnold et al. 2000, Wasow 2002, Bresnan et al. 2007.

The focus of our study is the relative order between the DO and the IO in the preverbal domain. The data we use are extracted from a corpus compiled out of daily newspaper articles and thus are essentially of a formal register, where the word order variations are expected to be limited and the canonical SOV order to be dominant. [13]

4 Corpus Data

Our study is conducted on the Bijankhan corpus, a corpus collected from daily news and common texts, in particular, the newspaper *Hamshahri*, of about 2.6 million tokens, manually tagged for part-of-speech information. The corpus was created in 2005 by the DataBase Research Group at the University of Tehran and can be freely downloaded from their website.[14]

4.1 Constitution of the Dataset

The Bijankhan corpus does not contain any syntactic annotation, nor is it lemmatized or delimited for sentences. Our first step was to lemmatize the corpus[15] and to delimit finite clauses on the basis of the conjugated verbs.[16] In total, 185,015 finite verbs were lemmatized, representing 322 verb types, since we considered *Particle-Verb* complex predicates as *bar-dāštan* 'to take' (see section 2.4) as a distinct verb type from the simplex verb. The number of simplex lemmas is 228.

We selected the potentially ditransitive verbs in order to isolate the potentially ditransitive sentences: 122 verb types, corresponding to 42,550 tokens out of which we extracted a random sample of 2000 tokens. We then manually identified the relevant sentences, that is, sentences matching either of the following patterns: NP PP V or PP NP V. We did not take into consideration the preceding constituents of the sentence. This dataset, *Dataset1*, contains 541 sentences formed with 82 verb types. Following Samvelian's (2012) argumentation against a clearcut distinction between complex predicates and ordinary complement-verb combinations, we did not aim to exclude complex predicates from our dataset. Consequently, our dataset contains a number of lexicalized complex predicates, e.g. *qarār gereftan* 'to be installed' (lit. 'installation take').

First, we annotated the DO for two properties, markedness and bareness: a) Markedness, to test the DOM criterion; b) Bareness, since bare objects correspond to the opposite extremity on the scale of specificity and/or definiteness compared to marked objects. Furthermore, they tend to form a semantic predicate with the verb. The distribution of the relative order with respect to these two variables is given in Table 1.

We observe that the data are globally consistent with the DOM criterion, as seen in Table 2. Marked DOs tend to be separated from the verb: 248 over 258 tokens are in DO-IO order. Unmarked DOs, that is, BARE and OTHER, tend to be adjacent to the verb: 74 over 283 tokens are in IO-DO order. However, marked DOs have a very consistent behavior compared to unmarked

[13]The postverbal realization of the IO, an ordering possibility prevailing in colloquial speech but expected to be limited in the written language (see section 2.1), is thus excluded by this methodological choice. To give an estimation, among all occurrences of the verbs *rixtan* 'to pour' and *ferestādan* 'to send' in the corpus, 254 and 219 respectively, there are only 8 cases where the IO is realized postverbally.

[14]http://ece.ut.ac.ir/dbrg/bijankhan/

[15]Given the limited number of Persian simplex verbs, we developed a dictionary-based lemmatizer. It should be noted that some finite verbs of the corpus remained unrecognized due mainly to tagging errors and orthographic anomalies. We ignored these verbs.

[16]Periphrastic verbal forms, that is, conjugations involving auxiliaries, were considered as single finite verbs.

Table 1
Distribution of word order by *markedness* and *bareness* in *Dataset1*

	DO			
	Marked	Bare	Other	Total
DO-IO-V	248	27	47	322
IO-DO-V	10	183	26	219
Total	258	210	73	541

Table 2
Contingency table for DOM and word order

	DO			
	Marked		Unmarked	
DO-IO-V	248	(96.12%)	74	(26.15%)
IO-DO-V	10	(3.88%)	209	(73.85%)

DOs, which show more versatility. 96% of marked DOs precede the IO, while 74% of unmarked DOs follow the IO.

A closer look at unmarked DOs reveals an inconsistency between bare nouns and unmarked non-bare DOs (labeled OTHER in Table 1). 87% of the former follow the IO while 64% of the latter precede the IO. To summarize, on the one hand, marked and bare objects not only verify the DOM criterion but also show only a slight variation. On the other hand, unmarked non-bare objects present a more significant amount of variation and more importantly, their preferred position goes against the DOM criterion.

With this observation, we felt the necessity for a more fine-tuned classification of unmarked non-bare DOs. We defined two classes on the basis of the degree of determination of the NP (see section 2.2). We separated determined NPs, that is, quantified or indefinite NPs, from non-determined NPs, that is, bare-modified NPs. Recall that the latter only differ from bare nouns by the presence of a modifier. Consequently, we end up with four DO types: BARE, BARE-MODIFIED, INDEFINITE (unmarked indefinite to be more precise), and MARKED.

The distribution of the relative order with regards to DO type is given in Table 3. The new classification provides some insights into the unbalanced variation observed with DOM. Indeed, the three types of unmarked DOs do not behave similarly. Interestingly, indefinite DOs seem to group with marked DOs, contrary to what is expected from the DOM criterion. Meanwhile, the preferred position of bare-modified DOs remains unclear and our dataset appears to be inconclusive. Nevertheless, it is clear that the DO type and relative order are strongly related (χ^2=348.7374, df = 3, p-value < 2.2e-16). Hence, the DO type is a relevant variable and probably a better predictor than the DOM criterion, since it captures more variation.

To remedy to this insufficiency, we enlarged our dataset. Given our first experience of token

Table 3
Distribution of word order by DO-type in *Dataset1*

	DO-type			
	Bare	Bare-modified	Indefinite	Marked
DO-IO-V	27	11	36	248
IO-DO-V	183	11	15	10
Total	210	22	51	258

Table 4
Distribution of word order by DO-type in *Dataset2*

	DO-type								
	Bare		Bare-modified		Indefinite		Marked		Total
DO-IO-V	43	(0.158 ***)	22	(0.333 **)	111	(0.770 ***)	403	(0.950 ***)	579
IO-DO-V	228		44		33		21		326
Total	271		66		144		424		905

Significance codes for p-values obtained by the χ^2 test: 0 '***' 0.001 '**'

identification (rate of 541/2000), we decided to modify our sampling method. We considered all occurrences of two typically ditransitive low frequency verbs of the corpus, *rixtan* 'to pour' and *ferestādan* 'to send' (219 and 254 tokens, respectively), and a random sample out of all occurrences of two high frequency typically ditransitive verbs, *gereftan* 'to give' and *dādan* 'to take' (10494 and 6849 tokens, respectively). This dataset (*Dataset2* hereafter) contains 905 tokens. The distribution of the relative order and the DO type is given in Table 4.

The new dataset confirms our observations concerning marked, bare, and indefinite DOs. Moreover, we can track down a preferential position for bare-modified DOs, which group with bare DOs, in conformity with the DOM criterion. Our data are particularly interesting for indefinite DOs, since their preferential position goes against the received hypothesis, the DOM criterion, according to which these DOs should group with bare nouns and bare-modified DOs, rather than *rā*-marked DOs. In *Dataset2* the DO type provides an accuracy of 86.8%, as against 78% for the DOM criterion.

4.2 Multifactorial Analysis

Our data reveal two different preferential orders for the IO and the DO in the preverbal domain, depending on the degree of determination of the DO. The DO type is indeed a very efficient predictor for the relative order between the DO and the IO; however, it leaves some variation unexplained. Given that studies on word order preferences on other languages have singled out factors such as heaviness, collocationality and lexical bias, we annotated *Dataset2* for these variables and performed mixed-effect logistic regression modeling (Agresti 2007) in order to study the effect of these variables independently and in interaction with each other.[17] Moreover, likelihood ratio tests were used to assess main effects and interactions and their contribution to the fit. In the remainder of this section, we will focus on the effect of the above-mentioned factors, heaviness in particular, without discussing the technical details of the modeling more than necessary.

[17]Logistic regression allows for the modeling of a categorical variable – in our case the binomial variable ORDER$_{\{DO-IO, IO-DO\}}$ - with a combination of categorical and continuous variables without any assumption about the distribution of the data. The *logit* transformation returns a value in the range of 0 and 1, which models the probability of the success scenario, in our case ORDER=DO-IO. It predicts ORDER=DO-IO, if the return value is bigger than 0.5, and ORDER=IO-DO otherwise. When the model returns 0, the return value of the *logit* transformation, that is, the probability of the success scenario, would be 0.5, which means no prediction is possible; likewise, negative return values correspond to failure and positive ones to success. In other words, positive coefficients vote for ORDER=DO-IO and negative ones for the inverse. The bigger the absolute value, the stronger the probability for either one. Wald tests are used to obtain p-values for individual coefficients.

4.2.1 Lexical Bias It has been shown that in preferential constituent ordering, the verb may exhibit a bias towards one order rather than the other (Wasow 1997, Stallings et al. 1998). Thus, verbal lemmas can be a source of variation in the preferential order and this is the case in our data as well. This variation is commonly dealt with using mixed models (e.g. Bresnan et al. 2007), which have the advantage of capturing the variation due to non-predicting variables, that is, random effects, in order to allow better estimates for the predictors, that is, fixed effects. Accordingly, we have included verbal lemmas as a random intercept. [18]

4.2.2 Collocationality Studies on word-order variations have pointed out that semantic connectedness can influence the ordering of constituents (e.g. Wasow 1997, Hawkins 2001). Constituents semantically connected to the verb, that is, constituents whose interpretation depends on the verb, tend to occur adjacent to it. In particular, Wasow (2002, 1997) provides corpus evidence on heavy-NP shift in English, showing that constituent ordering and semantic connectedness are correlated. The more the V-PP combination is semantically connected the more it is likely to appear adjacent and trigger the NP shift.[19] For Persian, semantic connectedness seems even more relevant, given the productivity of complex predicates, that is, syntactic combinations displaying a high degree of collocationality.

Both the IO and the DO can have a collocational relation with the verb and while this collocational relation does not necessarily imply adjacency, the prototypical pattern for a lexicalized complex predicate is either N-V, as in *qarār gereftan* 'to be installed' (lit. 'establishment take'), or P-N-V, as in *be kār bordan* 'to use' (lit. 'to work take'). As mentioned earlier, there are no formal criteria to systematically differentiate complex predicates from ordinary complement-verb combinations. Furthermore, there is no exhaustive list of (lexicalized) complex predicates available (Samvelian and Faghiri 2013, 2014). Hence, annotating the data for collocationality is not straightforward. A manual annotation based on native speakers' intuition would not only be subjective but also hardly independent of the word order. Consequently, we opted for an automatically annotated measure based on the frequency of the sequence N-V or P-N-V in the whole corpus (185k verbs). We operationalized this measure by a categorical variable, COLL-MES, with three levels depending on the frequency, NP-COLL, PP-COLL and NONE.[20] This variable has the advantage of being independent of annotators' judgments, but it has the disadvantage of being "blind", hence approximate and corpus-dependent.

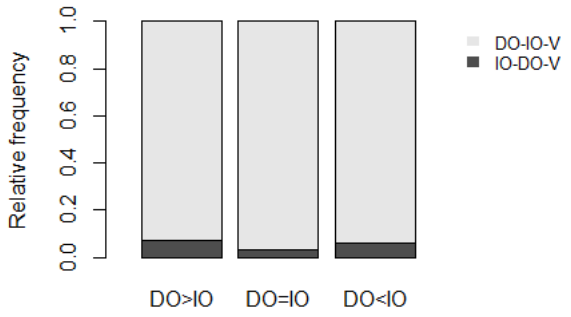
COLL-MES turned out to be significant (p-value < 0.001 for COLL-MES=NP-COLL) with the expected effect, that is, favoring the IO-DO order when the sequence N-V is coded as collocational. However, COLL-MES and DO-TYPE are highly related (χ^2 = 397.8262, df = 6, p-value < 2.2e-16) in

[18] An anonymous reviewer suggested that we group these verbs semantically and examine whether these classes correlate with the word order. Even though we did not classify verbal lemmas, we annotated the data for the preposition lemma, which reflects a semantic classification to some extent, and did not find a significant correlation. Note that this is indeed an important clue for the study of ordering preferences in the postverbal domain, which we will undertake in future studies.

[19] Wasow classifies V-PP combinations on the basis of their degree of collocationality and idiomaticity into the three following classes: non-collocations, semantically transparent collocations and semantically opaque collocations, that is, idioms, and observes that the rate of the NP shift, 26 %, 47%, and 60% respectively, increases with the degree of semantic connectedness.

[20] It should be noted that we tried different ways to operationalize this measure. The frequency as a continuous variable, a categorical variable with six levels (NPH, NPL, PPH, PPL and NONE), a categorical variable with three levels (NPH, NPL and NONE) and another one with (PPH, PPL and NONE). We opted for COLL-MES because it had a better performance on the data compared to the others.

Figure 1
Distribution of word order and relative length for marked DOs



our data and when we consider their interaction in the model, the significant effect of COLL-MES disappears. Moreover, this variable does not help to capture the variation in the data beyond the DO type. In other words, non-canonical orders, that is, where the order does not conform to the preferred order predicted by the DO type, cannot be explained by COLL-MES. More precisely, in the case of BARE and BARE-MODIFIED types, where 65 (out of 337) tokens do not follow the predicted IO-DO order, only 6, that is, less than 10%, are coded as PP-COLL. Likewise, in the case of MARKED and INDEFINITE types, where 54 (out of 514) tokens do not follow the predicted DO-IO order, only 3, that is, 5.5%, are coded as NP-COLL. Consequently, the significant effect of this variable in our data seems to be an illustration of the fact that bare objects tend strongly to participate in the formation of complex predicates rather than that of providing an explanation for the relative order.

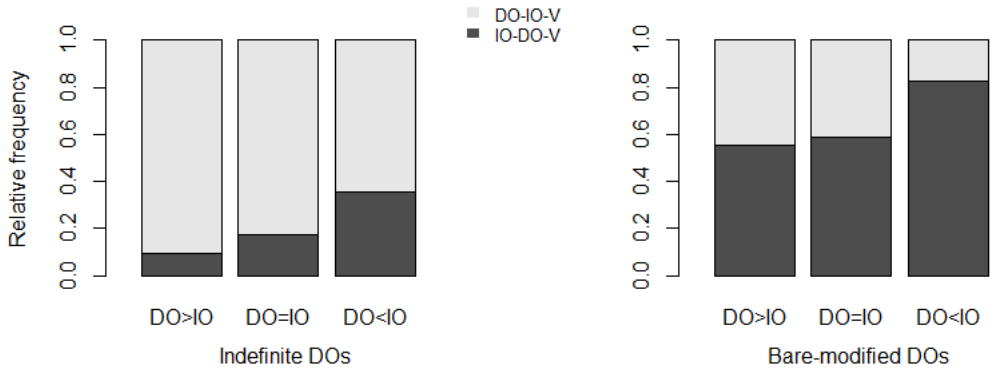
4.2.3 Heaviness Heaviness is one of the most frequently evoked factors in studies on constituent-ordering preferences in other languages. Yet, to our knowledge, it has not been investigated for Persian. As mentioned earlier, in head-initial languages, e.g. English (Wasow 2002) and French (Thuilier 2012), heaviness is shown to have an effect corresponding to the "short-before-long" tendency. In head-final languages, e.g. Japanese (Hawkins 1994, Yamashita and Chang 2001) and Korean (Choi 2007), the mirror-image effect is observed. Like Japanese and Korean, Persian is an SOV language, hence the "long-before-short" tendency would be expected.

In line with Wasow (1997, 2002), we operationalized the weight factor in terms of the relative length between the DO and the IO in number of words. First of all, we observe that the relative length is not relevant for all DO types and its influence on word order varies from one type to another. Relative length is irrelevant for bare DOs, given that it is by definition negative in this case.[21] As for the marked DOs, more than 95% of them are in the DO-IO order and, as illustrated by Figure 1, the data show no significant bias with respect to the relative length.

Focusing on indefinite and bare-modified DOs, however, it appears that the order is influenced by relative length. As illustrated by Figure 2, longer IOs are more likely to precede the

[21] Given that the NP in the IO can have an enclitic realization, the IO can consist of only one (phonological) word. Hence, 0 is also a possible value for this variable. We only had two such cases in the whole dataset; and they followed the IO-DO order.

Figure 2
Distribution of word order and relative length for indefinite and bare-modified DOs



DO. More precisely, in the case of indefinite DOs the shift from the (preferred) DO-IO order is reinforced when the IO is longer than the DO. In the case of bare-modified DOs, the general preference for the IO-DO order is reinforced when the IO is longer than the DO.

Given these observations, we built a model with only a subset of the data, that is, excluding bare nouns and marked DOs, with DO-TYPE and REL-LEN[22] as main effects[23] and VERB as a random intercept. The model is summarized in Table 5, where success corresponds to ORDER=DO-IO.

As expected, DO-TYPE has a significant effect: BARE-MODIFIED favors the IO-DO order and INDEFINITE the inverse. Interestingly, REL-LEN turned out also to have a significant effect with a positive coefficient, favoring the DO-IO order, when the DO is longer than the IO and the inverse, when the IO is longer than the DO. Thus, the effect of the relative length corresponds to the "long-before-short" tendency.

5 Long-before-short Tendency in OV Languages

Availability-based production accounts of word-order preferences suggest the universality of the "short-before-long" principle. According to these accounts, which are almost exclusively underpinned by studies on Germanic languages, short simple constituents can be processed and formulated faster than long ones and thus become available for production sooner. Hence, the "long-before-short" tendency observed in OV languages challenges this widely accepted view of sentence production.[24]

Building on extensive corpus studies from typologically different languages, Hawkins (1994, 2004) proposes a theory of word-order preferences based on the human parsing mechanism, which predicts opposite tendencies for VO and OV languages. Specifically, he postulates a

[22]We used the logarithmic transformation to minimize the effect of outliers. The exact value of REL-LEN is $\log(\text{DO}_{\text{Nb-of-words}}) - \log(\text{IO}_{\text{Nb-of-words}})$.

[23]The maximal model also included COLL-MES which was eliminated because it did not have a significant effect (p-values > 0.99).

[24]See Jaeger and Norcliffe (2009) for a discussion.

Table 5
Summary of results of mixed-effect model for ORDER

Random effects:					
	Groups	Name	Variance	Std. Dev.	
	VERB	(Intercept)	0.2245	0.4738	
	Number of obs: 210, groups: VERB, 31				
Fixed effects:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.5933	0.2947	5.406	6.45e-08	***
DO=BARE-MOD	-2.0397	0.3485	-5.852	4.85e-09	***
REL-LEN	0.8435	0.2609	3.233	0.00122	**

distance-minimizing dependency-based principle, the *Early Immediate Constituent* (EIC), according to which, other things being equal, the parser prefers a word order that allows the listener to recognize the phrase and its immediate constituents in the quickest possible manner. This principle is sensitive to the direction of the head. In a head-initial language like English, shifting a heavy NP to follow the PP allows the two constituents of the VP to be recognized more quickly, as illustrated by (24). All the words in the NP need to be processed before the PP is identified. Hence, in the case of a heavy NP, that is, when the NP is longer than the PP, reversing the order allows the identification of the two constituents by processing a smaller number of words. Likewise, in a head-final language like Japanese, the mirror-image shift minimizes the distance between the heads of the two constituents of the VP and allows them to be recognized more quickly than in the reverse ordering. However, in the case of a mixed head-direction language like Persian, EIC does not provide an adequate prediction. For instance, EIC does not provide any predictions for the preferred ordering of the IO and the DO when the DO is an indefinite NP, since in both orderings, as illustrated by (25b) and (25c), the same number of words must be processed in order to recognize the VP.

(24) a. I [$_{VP}$ introduced [$_{NP}$ some friends that John had brought] [$_{PP}$ to Mary]]
1 2 3 4 5 6 7 8

b. I [$_{VP}$ introduced [$_{PP}$ to Mary] [$_{NP}$ some friends that John had brought]]
1 2 3 4

(25) a. Yusef yek ketāb=e āmuzeš=e akkāsi az ketābxāne gereft
Yusef a book=EZ teaching=EZ photography from library took
'Yusef borrowed a photography tutorial book from the library.'

b. Yusef [$_{VP}$ [$_{NP}$ yek ketāb=e āmuzeš=e akkāsi] [$_{PP}$ az ketābxāne] gereft]
1 2 3 4 5 6 7

c. Yusef [$_{VP}$ [$_{PP}$ az ketābxāne] [$_{NP}$ yek ketāb=e āmuzeš=e akkāsi] gereft]
1 2 3 4 5 6 7

Despite the fact that the EIC principle correctly predicts the "long-before-short" preference in Japanese, Yamashita and Chang (2001, 2006) feel the need for a production-oriented account in the framework of the theory of grammatical coding (Bock and Levelt 1994, Garrett 1980) that could explain these seemingly contradictory tendencies. For these authors, acknowledging language-specific differences in sentence production is the key to a uniform account of word-order preferences. Since word-order preferences can be influenced by both conceptual and form-

related factors (Bock 1982), the sensitivity of a production system to these factors can be viewed as language-specific.

According to Yamashita and Chang (2001, 2006) the production system of Japanese, contrary to English, is more sensitive to conceptual factors than to form-related ones. This is because Japanese (and Persian for that matter) is a far less "rigid" language than English.[25] Moreover, in English Heavy-NP shift happens in the postverbal domain, where it is shown that the verb exerts strong influence, contrary to the preverbal domain (Stallings et al. 1998). These syntactic constraints presumably increase the effect of form-related factors over more conceptual ones. Longer constituents have competing properties. On the one hand, from a formal point of view, they are slower to process, therefore less accessible. On the other hand, they contain more lexical items, which makes them richer in meaning and more salient and hence more accessible from a conceptual point of view. Consequently, in Japanese, more sensitive to conceptual factors, placing long constituents before shorter ones is favored, while in English, more sensitive to form-related factors, placing short constituents before longer ones is favored.

6 Discussion

6.1 The DOM Criterion Revisited

According to our data, the preferential position of the DO is adjacent to the verb for bare nouns and bare-modified DOs and separated from the verb for marked and indefinite DOs. The degree of variation that each DO-type presents varies. Marked and bare nouns DOs behave in a very consistent manner and present a small (arbitrary or stylistic) variation, while indefinite and bare-modified DOs present a considerable amount of variation. In the light of these observations, it seems appropriate to revisit the DOM criterion. Indeed, it appears that subordinating the position of the DO to its degree of determination provides an account closer to reality than an account based on markedness only. Note that variation in the strength of these preferences can also be explained.

The more a DO is determined, that is, the more (discourse) accessible a DO, the more it is likely to be placed leftward in the sentence and separated from the verb. And the less a DO is determined, that is, the less (discourse) accessible a DO, the more likely it is to be placed adjacent to the verb. Put this way, it is plausible for DOs located in the middle of the hierarchy to show more variability than the ones located in the two extremities.

6.2 Relative Length

The data examined in this study show that despite its significant effect in the relative order of the DO and IO, relative length is of secondary importance in Persian, since relative order mainly depends on the type of the DO:

1. The position of *rā*-marked and bare DOs is totally independent of relative length;
2. Relative length has a significant effect on the ordering of indefinite and bare-modified DOs, conforming to the "long-before-short" tendency observed in OV languages.

Persian is very similar to Japanese with respect to the properties singled out by Yamashita and Chang (2001, 2006). Like Japanese and contrary to English, it displays a relatively free word

[25] Japanese has a fairly free word order and allows null pronouns. English, in contrast, has a fairly strict word order that requires all arguments to be overtly present (Yamashita and Chang 2001:54).

order and does not require all arguments to be overtly realized. Moreover, the ordering preferences under study take place in the preverbal domain. Following Yamashita and Chang (2001, 2006), we attribute the "long-before-short" tendency to the sensitivity of the preverbal domain in Persian to conceptual factors rather than to form-related ones. We assume that longer constituents are lexically richer and hence more salient.

We note that the "long-before-short" tendency can be integrated in the continuum established previously on the basis of the degree of determination of the DO, given that relative length plays a significant role for the DOs located in the middle of the hierarchy. In the case of these DOs, lexical richness contributes to the accessibility of the DO and hence a relatively more salient DO would be located higher in the continuum and therefore is more likely to be separated from the verb, whereas at the two extremities of the continuum, that is, marked and bare DOs, the nature of the DO determines its preferred position regardless of relative length.

6.3 Information Structure

Another highly discussed factor, influencing ordering preferences, alongside heaviness, is givenness (or newness) in discourse, that is, the information status (see Gundel 1988, Arnold et al. 2000, Bresnan et al. 2007). Although the study of the information structure suffers from some inconsistencies in terminology and analysis (see Gundel 1988, Lambrecht 1996, Ward and Prince 1991), the effect of givenness corresponding to the "given-before-new" principle seems uncontroversial, especially since it is consistent with accessibility-based production models.

At this stage of the study, we have not annotated the data for the information status of the DO or the IO and consequently have not been able to study the effect of the relative givenness on the word order. Nevertheless, we can discuss this factor to some extent on the basis of the referential givenness[26] of the DO. We observe that the continuum established based on the degree of determination of the DO conforms to the *Givenness Hierarchy* (Gundel et al. 1993).[27] Indeed, for NPs in the DO position in Persian, we can assume that *ra*-markedness corresponds to the highest degree of (referential) givenness, and bareness to the lowest degree of givenness. Consequently, given the continuum from the very strong preference of marked DOs to be separated from the verb to the very strong preference of bare DOs for adjacency, we observe that the preferred position of the DO is consistent with the "given-before-new" principle.

7 Conclusion

In this paper, we have presented corpus data on the relative order between the DO and the IO in Persian, which support the "long-before-short" tendency observed in other OV languages like Japanese and Korean. Yet, given that Persian, contrary to the latter, has a mixed head-direction behavior, Hawkins's (1994) EIC principle does not provide the expected prediction. On the contrary, Yamashita and Chang's (2001) production-oriented account is grounded in properties shared by Japanese and Persian. Consequently, in line with Yamashita and Chang (2001), we attribute this to the fact that the extra lexical material in longer constituents makes

[26] Gundel (1988) proposes two distinct and logically independent senses of givenness-newness: referential givenness and relational givenness. Relational givenness is about the partition of the semantic/pragmatic representation of the sentence into topic and focus. Referential givenness describes the relationship between a linguistic expression and a corresponding non-linguistic entity in the speaker's/hearer's mind.

[27] Gundel et al. (1993) define the (referential) *Givenness Hierarchy* with six cognitive statuses in the following increasing order: Type identifiable, Referential, Uniquely identifiable, Familiar, Activated and In focus.

them conceptually more accessible and that ordering preferences in Persian, like in Japanese, are more sensitive to conceptual factors than to form-related ones.

Furthermore, in Persian, relative length is only of secondary importance, since the position of the DO mainly depends on its degree of determination. The more a DO is determined the more it is likely to be separated from the verb. We can trace a continuum from the *rā*-marked DOs to bare DOs which conforms to the Givenness Hierarchy and supports the "given-before-new" principle.

We are currently undertaking a series of controlled experiments to verify the results of our corpus study with respect to relative length and to further investigate the role of the information structure.

References

Agresti, Alan. 2007. *An introduction to categorical data analysis.* Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley-Interscience.

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21:435–483.

Arnold, Jennifer E., Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language* 76:28–55.

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* 89:1–47.

Bock, J. Kathryn, and Willem Levelt. 1994. Language production: Grammatical encoding. In *Handbook of psycholinguistics*, ed. Morton A. Gernsbacher, 945–984. New York: Academic Press.

Bossong, Georg. 1985. *Empirische Universalienforschung: differentielle Objektmarkierung in den neuiranischen Sprachen.* Tübingen : Gunter Narr Verlag.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, ed. Boume Gerlof, Irene Kraemer, and Joost Zwarts, 69–94. Royal Netherlands Academy of Science.

Choi, Hye-Won. 2007. Length and order: A corpus study of Korean dative-accusative construction. *Discourse and Cognition* 14:207–227.

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. In *Cooperating with written texts: The pragmatics and comprehension of written texts*, 549–573. Mouton de Gruyter.

Faghiri, Pegah. 2010. La morphologie du pluriel nominal du persan d'après la théorie whole word morphology. Master's thesis, Université de Montréal.

Garrett, Merrill F. 1980. Levels of processing in sentence production. *Language production* 1:177–220.

Ghomeshi, Jila. 2003. Plural marking, indefiniteness, and the noun phrase. *Studia linguistica* 57:47–74.

Givi Ahmadi, Hassan, and Hassan Anvari. 1995. *Dastur zabāne fārsi [Persian grammar].* Mo'assese farhangi Fātemi.

Gundel, Jeanette K. 1988. Universals of topic-comment structure. In *Studies in syntactic typology*, ed. M. Hammond, E.A. Moravcsik, and J.R. Wirth, Typological studies in language, 209–239. John Benjamins.

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

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge studies in linguistics. Cambridge University Press.

Hawkins, John A. 2001. Why are categories adjacent? *Journal of Linguistics* 37:1–34.

Hawkins, John A. 2004. *Efficiency and complexity in grammars.* Oxford Linguistics. Oxford University Press.

Jaeger, T. Florian, and Elisabeth J. Norcliffe. 2009. The cross-linguistic study of sentence production. *Language and Linguistics Compass* 3:866–887.

Karimi, Simin. 2003. On object positions, specificity, and scrambling in Persian. In *Word order and scrambling*, ed. Simin Karimi, 91–124. Blackwell Publishing Ltd.

Karttunen, Lauri. 1976. Discourse referents. In *Syntax and semantics 7: Notes from the linguistic underground*, ed. James D. McCawley, 363–85. Academic Press, New York.

Kempen, Gerard, and Karin Harbusch. 2003. Word order scrambling as a consequence of incremental sentence production. In *Mediating between concepts and grammar*, ed. Holden Härtl and Heike Tappe, 141–64. Berlin, Germany: Mouton De Gruyter.

Lambrecht, Knud. 1996. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents.* Cambridge Studies in Linguistics. Cambridge University Press.

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

Lazard, Gilbert, Yann Richard, Rokhsareh Hechmati, and Pollet Samvelian. 2006. *Grammaire du persan contemporain.* Bibliothèque iranienne. Institut français de recherche en Iran.

Meunier, Annie, and Pollet Samvelian. 1997. La postposition *rā* en persan : son rôle dans la détermination et sa fonction discursive. *Cahiers de Grammaire* 25:187–232.

Rasekhmahand, Mohammad. 2004. Jāygāh maf'ul mostaqim dar fārsi [The position of the direct object in Persian]. *nāme farhangestān* 6:56–66.

Samvelian, Pollet. 2001. Le statut syntaxique des objets nus en persan. *Bulletin de la Société de Linguistique de Paris* 96:349–388.

Samvelian, Pollet. 2007. A (phrasal) affix analysis of the Persian ezafe. *Journal of Linguistics* 43:605–645.

Samvelian, Pollet. 2012. *Grammaire des prédicats complexes : les constructions nom-verbe.* Herms–Lavoisier.

Samvelian, Pollet, and Pegah Faghiri. 2013. Introducing PersPred, a syntactic and semantic database for Persian complex predicates. In *Proceedings of the Workshop on Multiword Expressions*, 11–20. Atlanta, Georgia, USA.

Samvelian, Pollet, and Pegah Faghiri. 2014. Persian complex predicates: How compositional are they? *Semantics-Syntax Interface* 1:43–74.

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. *Journtal of Memory and Language* 39:392 – 417.

Thuilier, Juliette. 2012. Contraintes préférentielles et ordre des mots en français. Doctoral Dissertation, Université Paris Diderot.

Ward, Gregory L., and Ellen F. Prince. 1991. On the topicalization of indefinite nps. *Journal of Pragmatics* 16:167–177.

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

Wasow, Thomas. 2002. *Postverbal behavior.* CSLI lecture notes. CSLI.

Yamashita, Hiroko, and Franklin Chang. 2001. "Long before short" preference in the production of a head-final language. *Cognition* 81:B45–B55.

Yamashita, Hiroko, and Franklin Chang. 2006. Sentence production in Japanese. *Handbook of East Asian psycholinguistics* 2:291–297.

(Faghiri)
Université Sorbonne Nouvelle - Paris 3
pegah.faghiri@univ-paris3.fr

(Samvelian)
Université Sorbonne Nouvelle - Paris 3
pollet.samvelian@univ-paris3.fr