Heaviness and Constituent ordering: a Corpus-based study in Persian

Pegah Faghiri
Université Sorbonne Nouvelle / Mondes iranien et indien (CNRS) & Laboratoire de Linguistique Formelle

CECIL’S 2013
Peter Pazmany Catholic University, Hungary
22–23 August 2013
Effect of heaviness in the relative order between the verbal complements

• **Short-before-long** (end-weight) principle:
  - Processing & planning heavy constituents require more memory or resources
  - Costly constituents tends to be postponed.

(Wasow, 2002; Arnold et al, 2000; Stallingd et al, 1998; a.o.)

• Is this principle **universal**?
  - Hawkins *EIC* principle predicts an asymmetry in VO and OV languages
  - Long-before-short principle in OV languages (confirmed for Japanese by corpus and experimental studies)
Effect of heaviness in the relative order between the verbal complements

• Short-before-long principle:
  ➢ Processing & planning heavy constituents require more memory or resources
  ➢ Costly constituents tends to be postponed.

• Is this principle universal?
  ➢ Hawkins *Early Immediate Constituent* (EIC) principle
    Minimize domain → Maximize efficiency
    Predicting an asymmetry in VO and OV languages
    (Hawkins, 1994, 2008 a.o.)

  ➢ Long-before-short principle in OV languages
    Confirmed for Japanese by corpus and experimental data
    (Yamashita & Chang, 2001)
Object of study:

The preferential word order between the DO and the IO in preverbal domain in Persian

Methodology:

Corpus-based study using logistic regression modeling
Object of study:
The preferential word order between the DO and the IO in preverbal domain in Persian

Methodology:
Corpus-based study using logistic regression modeling
Essential properties of Persian

• A mixed head-direction language

• Head-final in verbal domain **but head-initial elsewhere**:
  - Nominal domain is head-initial: Det N Mod
  - Prepositions and no postpositions: Prep NP
  - Clausal phrase follow the complementizer: Comp P

• SOV is the canonical order but all variations are possible depending on register, information structure, prosody, etc.
  - E.g. goal arguments (locatives and datives) are post-verbal in informal language
  - Clausal arguments are strictly post-verbal
Essential properties of Persian

• A mixed head-direction language

• Head-final in verbal domain but head-initial elsewhere:
  ➢ Nominal domain is head-initial
  ➢ Prepositions and no postpositions
  ➢ Clausal phrase follow the complementizer

• SOV is the canonical order but **all variations are possible** depending on register, information structure, prosody, etc.
  ➢ e.g. goal arguments (locatives and datives) are post-verbal in informal language
  ➢ Clausal arguments are strictly post-verbal
Complex predicates (CPs)

• Only around 200 simplex verbs

• Verbal concepts are expressed by combination of a non-verbal element and a verb:
  - bāzi kardan: play do -> to play
  - harf zadan: speech hit -> to speak
  - be kār bordan: to work take -> to use
  - az dast dādan: of hand give -> to loose

→ From syntactic point of view the combination behaves like the combination of a verb with its complement

(Samvelian, 2012 a.o.)

• Prototypic pattern: N V and Prep N V
Complex predicates (CPs)

- Only around 200 simplex verbs
- Verbal concepts are expression by combination of a non-verbal element and a verb:
  - bāzi kardan: play do -> to play
  - harf zadan: speech hit -> to speak
  - be kār bordan: to work take -> to use
  - az dast dādan: of hand give -> to loose

→ From syntactic point of view the combination behaves like the combination of a verb with its complement

(Samvelian, 2012 a.o.)

- Prototypic pattern: N V and Prep N V
Does Hawkins’s *EIC* principle work for Persian?

Data for Japanese strict head-final language

Mary said that John got married yesterday.

```
[Mary-ga] [kinoo yesterday John-ga kekkonsi-ta to] it-ta
\[NP\] \[NP Prep\] V 1 2 3-5 6
\[NP Prep\] \[NP\] V 1 2 3
```

DO < IO (by 3 words)

```
[NP] [NP Prep] V
1 2 3-5 6
[NP Prep] [NP] V
1 2 3
```

The IO OD order should be preferred
Does Hawkins’s *EIC* principle work for Persian?

DO IO or IO DO?

**DO < IO (by 3 words)**

\[
\begin{array}{cccc}
[\text{NP}] & [\text{Prep} & \text{NP}] & V \\
1 & 2 & 3-5 & 6 \\
\end{array}
\]

No preferential order based on relative length

**DO > IO (by 3 words)**

\[
\begin{array}{cccc}
[\text{NP}] & [\text{Prep} & \text{NP}] & V \\
1-5 & 6 & 7 & 8 \\
\end{array}
\]

No preferential order based on relative length
Most prominent hypothesis regarding complement ordering in Persian is the **Differential Object Marking criterion**
The DOM criterion

• DOM in Persian

- Definite and/or specific DOs are marked with the enclitic \( =rā \)

<table>
<thead>
<tr>
<th></th>
<th>in</th>
<th>ketāb=rā</th>
<th>be Nima</th>
<th>dād</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maryam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maryam</td>
<td>this</td>
<td>book=DOM</td>
<td>to Nima</td>
<td>gave</td>
</tr>
</tbody>
</table>

‘Maryam gave this book to Nima.’

- Indefinite non-specific DOs are unmarked

<table>
<thead>
<tr>
<th></th>
<th>be Nima</th>
<th>ketāb</th>
<th>dād</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maryam</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maryam</td>
<td>to Nima</td>
<td>book</td>
<td>gave</td>
</tr>
</tbody>
</table>

‘Maryam gave a book/books to Nima.’
The DOM criterion

• DOM in Persian
  - Definite and/or specific DOs are marked with the enclitic \( =r\dot{a} \)
  - Indefinite non-specific DOs are unmarked

• The hypothesis:
  - Marked DOs can be separated from the verb: \( DO \ IO \ V \)
  - Unmarked DOs should be adjacent to the verb: \( IO \ DO \ V \)

  Our Corpus study (at the preliminary stage) showed that part of this hypothesis fails usage data validation:
  - Marked DOs have a very strong (95%) preference for the NP PP order
  - But, unmarked DOs do not behave homogenously

(Karimi, 2005 a.o.)
The DOM criterion

• DOM in Persian
  ➢ Definite and/or specific DOs are marked with the enclitic =rā
  ➢ Indefinite non-specific DOs are unmarked

• The hypothesis:
  ➢ Marked DOs can be separated from the verb: DO IO V
  ➢ Unmarked DOs should be adjacent to the verb: IO DO V

❖ Our Corpus study (at the preliminary stage) showed that part of this hypothesis fails corpus data validation:
  ➢ Marked DOs have a very strong (95%) preference for the NP PP order
  ➢ But, unmarked DOs do not behave in a homogenous manner

(Faghiri & Samvelian, 2013)
Based on preliminary observations on corpus data, 4 DO types have been defined:

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare</td>
<td>(1) Maryam be Nima ketāb dād Maryam to Nima book gave</td>
<td>‘Maryam gave a book to Nima.’</td>
</tr>
<tr>
<td>Bare modified</td>
<td>(2) Maryam be Nima ketāb=e tārix dād Maryam to Nima book=EZ* history gave</td>
<td>‘Maryam gave a history book to Nima.’</td>
</tr>
<tr>
<td>Indefinite</td>
<td>(3) Maryam čand ketāb=e qadimi be Nima dād Maryam some book=EZ old to Nima gave</td>
<td>‘Maryam gave some old books to Nima.’</td>
</tr>
<tr>
<td>Marked</td>
<td>(4) Maryam in ketāb=rā be Nima dād Maryam this book=DOM to Nima gave</td>
<td>‘Maryam gave this book to Nima.’</td>
</tr>
</tbody>
</table>
Our corpus study
Corpus

- Bijankhan corpus (Bijankhan, 2004), freely available
- 2,6m tokens, extracted from newspaper
- Manually annotated for POS

Dataset

- Lemmatized verbs, extracted ditransitives (42k token, 122 lemmas)
- First dataset (541 tokens, 82 lemmas)
  - Random sample of 2000 tokens
  - Identified sentences corresponding to the NP PP V or PP NP V pattern
- Final dataset (908 tokens, 82 lemmas)
  - All instances of two low frequency typically ditransitive verbs ‘to send’ and ‘to pour’
  - Random samples of two high frequency typically ditransitive verbs ‘to give’ and ‘to take’
Corpus
• Bijankhan corpus (Bijankhan, 2004), freely available
• 2.6m tokens, extracted from newspaper
• Manually annotated for POS

Dataset

1. We lemmatized verbs and extracted ditransitives \((42k \text{ token, } 122 \text{ lemmas})\)
2. First dataset \((541 \text{ tokens, } 82 \text{ lemmas})\)
   Random sample of 2000 tokens
   – Identified sentences corresponding to the NP PP V or PP NP V pattern
3. Final dataset
   – All instances of two low frequency typically ditransitive verbs ‘to send’ and ‘to pour’ \((219, 254 \text{ tokens respectively})\)
   – Random samples of two very high frequency typically ditransitives ‘to give’ and ‘to take’ \((10494, 6849 \text{ tokens respectively})\)
Methodology

• Mixed-effect regression model*
  – Dependent variable: Order (NP PP V = 1)
  – Random effect: verbal lemma
  – Predicting variables:
    • DO type
    • Relative length (nb of words): log(NP) – log(PP)
    • Collocational relation with the verb: Frequency of the sequences N-V or Prep-N-V in the whole corpus

*Executed with R
Data description:

- Average preference of 59% for NP PP V order
- All variables came out to have a significant effect
- DO type and order are strongly correlated
Data description:

- Average preference of 59% for NP PP V order
- All variables came out to have a significant effect
- DO type and order are strongly correlated

<table>
<thead>
<tr>
<th></th>
<th>Bare</th>
<th>Bare-Modified</th>
<th>Indefinite</th>
<th>Marked</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP V</td>
<td>43</td>
<td>(16%)</td>
<td>23 (34%)</td>
<td>112 (77%)</td>
</tr>
<tr>
<td>PP NP V</td>
<td>228</td>
<td>(84%)</td>
<td>44 (66%)</td>
<td>33 (23%)</td>
</tr>
<tr>
<td>Total</td>
<td>271</td>
<td></td>
<td>145</td>
<td>425</td>
</tr>
</tbody>
</table>
Data description:

- Average preference of 59% for NP PP V order
- All variables came out to have a significant effect
- DO type and order are strongly correlated

<table>
<thead>
<tr>
<th></th>
<th>Bare</th>
<th>Bare-Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP V</td>
<td>43</td>
<td>23</td>
</tr>
<tr>
<td>PP NP V</td>
<td>228</td>
<td>44</td>
</tr>
<tr>
<td>Total</td>
<td>271</td>
<td>67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Indefinite</th>
<th>Marked</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP V</td>
<td>112</td>
<td>404</td>
</tr>
<tr>
<td>PP NP V</td>
<td>33</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>145</td>
<td>425</td>
</tr>
</tbody>
</table>
Data description:

- Average preference of 59% for NP PP V order
- All variables came out to have a significant effect
- DO type and order are strongly correlated

DO type predict order with 87% of accuracy in our data

N.b. the DOM provide 78% of accuracy

<table>
<thead>
<tr>
<th></th>
<th>Bare</th>
<th>Bare-Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP V</td>
<td>43 (16%)</td>
<td>23 (34%)</td>
</tr>
<tr>
<td>PP NP V</td>
<td>228 (84%)</td>
<td>44 (66%)</td>
</tr>
<tr>
<td>Total</td>
<td>271</td>
<td>67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Indefinite</th>
<th>Marked</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP V</td>
<td>112 (77%)</td>
<td>404 (95%)</td>
</tr>
<tr>
<td>PP NP V</td>
<td>33 (23%)</td>
<td>21 (5%)</td>
</tr>
<tr>
<td>Total</td>
<td>145</td>
<td>425</td>
</tr>
</tbody>
</table>
Data description:

- Average preference of 59% for NP PP V order
- All variables came out to have a significant effect
- DO type and order are strongly correlated
- The previous hypothesis with regards to the DOM criterion is only partially valid:

<table>
<thead>
<tr>
<th></th>
<th>Bare</th>
<th>Bare-Modified</th>
<th>Indefinite</th>
<th>Marked</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP V</td>
<td>43 (16%)</td>
<td>23 (34%)</td>
<td>112 (77%)</td>
<td>404 (95%)</td>
</tr>
<tr>
<td>PP NP V</td>
<td>228 (84%) ✓</td>
<td>44 (66%) ✓</td>
<td>33 (23%)</td>
<td>21 (5%)</td>
</tr>
<tr>
<td>Total</td>
<td>271</td>
<td>67</td>
<td>145</td>
<td>425</td>
</tr>
</tbody>
</table>
The relative length effect:

- Average preference of 59% for NP PP V order
- All variables came out to have a significant effect
- DO type and order are strongly correlated
- The previous hypothesis with regards to the DOM criterion is only partially valid:

<table>
<thead>
<tr>
<th></th>
<th>Bare</th>
<th>Bare-Modified</th>
<th>Indefinite</th>
<th>Marked</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP V</td>
<td>43 (16%)</td>
<td>23 (34%)</td>
<td>112 (77%)</td>
<td>404 (95%)</td>
</tr>
<tr>
<td>PP NP V</td>
<td>228 (84%)</td>
<td>44 (66%)</td>
<td>33 (23%)</td>
<td>21 (5%)</td>
</tr>
<tr>
<td>Total</td>
<td>271</td>
<td>67</td>
<td>145</td>
<td>425</td>
</tr>
</tbody>
</table>

The relative length has an effect only in these cases
The relative length effect:

Beyond the strong effect of DO type

**Relative length** shows a significant effect \((p\text{-value} < 0.001)\)

corresponding to the **long-before-short** tendency

**Improving accuracy by 2%**
Long-before-short tendency

Relative length have an effect in the case of **Indefinite** and **Bare-Modified DO**

As for Bare DOs, Relative length is meaningless

NP is always shorter (or equal) to PP
Long-before-short tendency

Shorter DOs prefer the **PP NP V** order significantly more often

NP PP = 1
PP NP = 0

Pegah Faghiri - Heaviness and Constituent ordering in Persian
CECIL’S3 – 22 August 2013
Discussions

**Short-before-long is not universal**
Not only Japanese (strictly head-final) but also Persian (mixed head-direction) presents the long-before-short tendency

→ The verbal position has to be taken into account in the effect of relative length on preferential order between verbal complements
→ Theories solely based on general principles ignoring linguistic parameters would eventually fail cross-linguistic validity
→ Theories proposing accounts in terms of dependency seems to be more appropriate

❖ However Hawkins’s *EIC* principles fails to account for Persian data
Discussions

Short-before-long is not universal
Not only Japanese (strictly head-final) but also Persian (mixed head-direction) presents the long-before-short tendency

→ The **position of the verb** has to be taken into account in the effect of relative length on preferential order between verbal complements
→ Theories solely based on general principles ignoring linguistic parameters would eventually fail cross-linguistic validity
→ Theories proposing accounts in terms of dependency seems to be more appropriate
  ❖ However Hawkins’s *EIC* principles fails to account for Persian data
Discussions

**Short-before-long is not universal**
Not only Japanese (strictly head-final) but also Persian (mixed head-direction) presents the long-before-short tendency

→ The verbal position has to be taken into account in the effect of relative length on preferential order between verbal complements

→ Theories solely based on general principles ignoring linguistic parameters would eventually fail cross-linguistic validity

→ Theories proposing accounts in terms of dependency seems to be more appropriate

  ❖ However Hawkins’s *EIC* principles fails to account for Persian data

**Furthermore:**

In Persian the relative length plays only a secondary role while the **DO type, which depends on the information status of the NP, plays the essential role.**
To go further: Experimental methods

We are currently running a couple of experiments to explore the effect of information structure and relative length independently

- For Indefinite and Bare-Modified DOs (2 experiments):
  Semi-guided production task (online questionnaire on Ibex)

  2 conditions (2x2):
  - Givenness: IO given vs IO new (DO always new)
  - Length: DO > IO vs DO < IO (at least 6 syllables)
  - With control for Animacy (DO –animate, IO +animate)

  20 items (7 verbs)
To go further: Experimental methods

We are currently running a couple of experiments to explore the effect of information structure and relative length independently.

- For Indefinite and Bare-Modified DOs (2 experiments):
  Semi-guided production task (online questionnaire on Ibex-farm)

  2 conditions (2x2):
  - Givenness: **IO given** vs **IO new** (DO is always new)
  - Length: **DO > IO** vs **DO < IO** (at least 6 syllables)

- With control for Animacy: **DO –animate, IO +animate**

  Schema: ‘**someone (something) (to someone) give**’

  20 items (7 verbs) / 40 fillers
References


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


Thanks to my advisors:

Pollet Samvelian (Université Sorbonne Nouvelle / MII)
Barbara Hemforth (Université Paris-Diderot / LLF)

Thank you for your attention!

pegah.faghiri@univ-paris3.fr

This study is part of a project on word order effects across languages in the Labex Empirical Foundations of Linguistics (ANR/CGI).