Heaviness on the left edge
Observing linguistic processing in historical corpora

CLARIN 2019 – September 30 – October 2, 2019, Leipzig University

Ingunn Hreinberg Indriðadóttir
University of Iceland

Anton Karl Ingason
University of Iceland

UNIVERSITY OF ICELAND

Introduction

- It is a well known observation that heavy syntactic constituents sometimes appear at the end of a clause rather than in their canonical position.
- We examine the relationship between heaviness and optional movement to the edge of a clause.
- We demonstrate how a digitized and syntactically annotated corpus of historical texts can contribute to the study of phenomena associated with linguistic processing.
- We demonstrate that heaviness is not only positively correlated with movement to the right edge of a clause, but also to the left edge, e.g. by left dislocation

Optionally moving heavy elements to the edge

Placing syntactic elements at the right edge is often more natural if they are heavy (long).

(1) a. ?Stella read [to the children] [a book].
b. Stella read [to the children] [a book about lions and tigers].

This effect is found across languages, i.e., it is well-known in both English and Icelandic.

The current study uses the Icelandic treebank to make the point that heaviness can also increase the probability of leftward movement.

The Icelandic Treebank

- Our study is based on the Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al. 2011).
- IcePaHC contains about 1 million words of Icelandic prose, parsed syntactially for full phrase structure and hand-corrected.
- The treebank is available under a GPL-style license as a free and open source resource.
- While the raw data can be downloaded and used in research as well as commercial scenarios, search queries can also be submitted via an online search interface on `treebankstudio.org`

Edge weight rather than end weight

Left-Disolcation: Thráinsson (1979) described Left Dislocation in Icelandic as a construction with a similar discourse function as Topicalization: the targeted constituent has usually been introduced in the preceding discourse and its discourse function can be described as a reintroduction of a discourse topic or theme. For this reason, the targeted constituent is usually definite

(2) a. María sá prest í bænum í gær.
Mary saw priest downtown yesterday
'Mary saw a priest downtown yesterday.'
b. *[Prestur], María sá [hann] í bænum í gær.
priest Mary saw him downtown yesterday
Intended: 'A priest, Mary saw him downtown yesterday.'
c. [Presturinn], María sá [hann] í bænum í gær.
the.priest Mary saw him downtown yesterday
'The priest, Mary saw him downtown yesterday.'

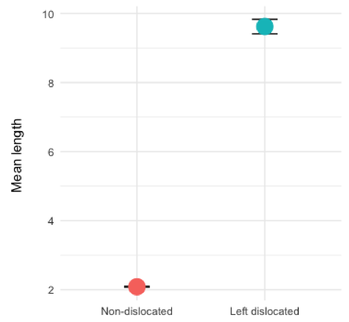
Beyond discourse: While discourse status matters for the probability of Left-Dislocation, so does the weight of the dislocated element.

Left Dislocated Subjects

Mean length of Subjects moved by Left Dislocation (μ: 2,1) vs Subjects in situ (μ: 9,6).

(Mann-Whitney U test: U = 77105, p<0.001).

We analyzed 34191 subjects, 193 of which were Left-Dislocated.

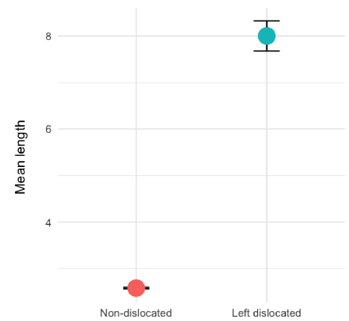


Left Dislocated Objects

Mean length of Objects moved by Left Dislocation (μ: 8) vs Objects in situ (μ: 2,57).

(Mann-Whitney U test: U = 614480, p<0.001).

We analyzed 25005 objects, 28 of which were Left-Dislocated.



Both subjects and direct objects that are moved by Left Dislocation tend to be long and, on average, considerably longer than the ones left in situ.

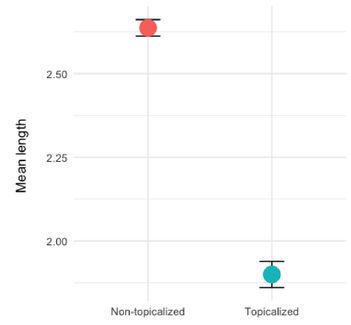
Move left only if not already on the right edge

Topicalized Direct Objects

Mean length of Non-Topicalized Direct Objects (μ: 2,6) vs Topicalized Direct Objects in situ (μ: 1,9).

(Mann-Whitney U test: U = 5442000, p=0.0128).

We analyzed 11688 Direct Objects, 1070 of which were Topicalized.

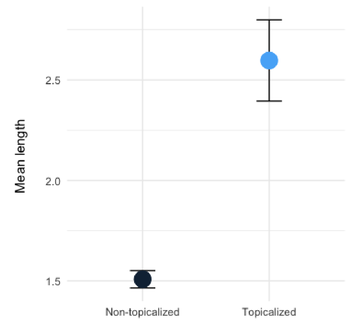


Topicalized Indirect Objects

Mean length of Non-Topicalized Indirect Objects (μ: 1,5) vs Topicalized Indirect Objects in situ (μ: 2,6).

(Mann-Whitney U test: U = 77105, p<0.001).

We analyzed 2012 indirect objects, 57 of which were topicalized.



Opposite pattern: Shifted phrases are shorter than the ones in situ.

Summary:

- Leftward movement, in particular Left Dislocation, moves heavy elements to the left edge, similarly to rightward movement.
- Heavy elements that are already on the right edge of the sentence do not need to undergo leftward movement, as they are already on an edge.
- Heavy elements that are placed in the middle of a sentence may be moved to either the left or right edge, whichever better suited in each case to facilitate grammatical parsing.

Conclusion

- Movement to both edges is associated with heaviness, not just to the right edge.
- Moving something to the edge can facilitate parsing in cases where speakers need to recover from a deeply embedded structure in the middle of a clause.
- Our study illustrates how digital parsed corpora of historical languages are useful for studying processing effects.

ingunn@hi.is • antoni@hi.is