Linguistic end-weight is really edge-weight Observing heaviness is a parsed corpus[⋆]

Ingunn Hreinberg Indriðadóttir[0000−0002−1863−0153]
and Anton Karl Ingason[0000−0002−2069−5204]

University of Iceland, Sæmundargata 2, 101 Reykjavík, Iceland

Abstract. This paper examines the relationship between heaviness and optional movement to the edge of a clause – demonstrating how a digitized and syntactically annotated corpus of historical texts can contribute to the study of phenomena associated with linguistic processing. We focus on so-called weight phenomena in word order variation and find that heaviness draws phrases to both edges of a clause – not just the right edge as sometimes assumed.

Keywords: historical corpora · heaviness · end weight · movement · processing.

1 Introduction

This paper examines the relationship between heaviness and optional movement to the edge of a clause – demonstrating how a digitized and syntactically annotated corpus of historical texts can contribute to the study of phenomena associated with linguistic processing. It is a well known observation that syntactic constituents sometimes appear at the end of a clause rather than in their canonical position when they are heavy/long; going back to Behagel [2] , see also [21, 22]. This tendency is manifested in Heavy NP shift, the type of alternation shown in where the direct object can shift to the right of the PP adjunct *on the street.*[1]

(1) a. I met [my rich uncle from Detroit] on the street.
    b. I met on the street [my rich uncle from Detroit].

Despite several studies on weight effects, it still remains a matter of investigation why such movement takes place. Proposed explanations appeal to some aspects of processing and include that such movement facilitates parsing [3, 5, 6, 7, 11] or utterance planning and production [22]. It even remains elusive which kind of measurement is most appropriate for deciding what counts as heavy [16, 18, 22, 23].

[⋆] Thanks to anonymous reviewers for helpful comments. This research was supported by The Icelandic Research Fund (Rannis) [grant number 185263-051].

[1] For a syntactic analysis of this example and other similar examples of Heavy NP shift, see [10, 15, 20].

We will not attempt to resolve these big questions. That task goes far beyond the scope of such a short paper. However, we do want to make an empirical point that in our opinion seems to escape attention in some of the most important studies on weight effects. 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 (2).

(2) a. I forgot about [my rich uncle from Detroit].
b. [My rich uncle from Detroit]$_1$, I forgot about him$_1$.

This is important because it suggests that weight-driven movement is, at least in part, about amending situations where one needs to backtrack from a deeply embedded structure in the middle of an utterance rather than moving to the right.

The paper is organized as follows: In section 2 we discuss the placement of old vs. new/given information in a sentence and the nature of rightward movement. In section 3 we introduce our study and the concept of Edge weight. In section 4 we discuss the results of our study. Section 5 concludes.

2 Optionally moving heavy elements to the edge

There is a well-known tendency for syntactic constituents that introduce new information to appear later in the sentence than constituents that present old/given information (see Prince 1981 for her account of definitions of old vs. new information).

Thrainsson (2005:506) argued that in languages such as Icelandic the basic word order of Subject-Verb-Object does not always agree with the tendency to present old information before new. As he demonstrates in example (3), both purposes can be fulfilled by choosing a passive sentence, rather than active.

(3) a. María lamdi strákinn.
Mary beat the.boy
'Mary beat the boy.'
b. Strákurinn var laminn af Maríu.
the.boy was beaten by Mary
'The boy was beaten by Mary.'

The sentences in (3) have more or less the same meaning but offer two ways of organizing old and new information without violating the rules of syntactic structure in Icelandic. The tendency for new information to appear at the right edge of a sentence is also manifested in various exceptions from the Definiteness Restriction [8, 13]. The restriction prohibits definite DPs from acting as late subjects in existential sentences with the dummy *það* (comparable to *there* in English) but this restriction can be violated under certain conditions [9].

(4) a. Það bilaði bíllinn.
there broke down the.car

‘The car broke down.’
b. Það er stíflaður vaskurinn.
there is clogged the.sink
‘The sink is clogged.’

The conditions for violating the Definiteness Restriction are, as Jónsson [9] demonstrates in (5) and (6), that the subject in the position to the right must present new information.

(5) A: Af hverju komstu ekki á bílnum?
‘Why didn’t you use the car to get here?’
B:
a. ??Nú, það bilaði bíllinn.
well there broke down the.car
‘Well, the car broke down.’
b. Nú, bíllinn bilaði.
well, the.car broke down
‘Well, the car broke down.’

(6) A: Hvað gerðist eiginlega?
‘What happened?’
B:
a. Það bilaði bíllinn.
there broke down the.car
‘The car broke down.’
b. Bíllinn bilaði.
the.car broke down
‘The car broke down.’

Various suggestions have been made to explain why heavy elements can be moved to the right edge of a sentence, whereas lighter elements are not as easily shifted, as demonstrated in example (7).

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

One of these suggestions is that it serves the purpose of placing old information closer to the sentence initial position and new and less predictable information further to the right [12]. Other accounts give more value to the sheer length and/or complexity of the shifted element [4, 11, 15], rather than information structure, although it has been demonstrated that both factors are weight predictors for word order in English, independantly and simultaneously [1]. Some accounts have suggested that the relative length of the word-string the shifted constituent moves over is also important [17, 22].

The question of how syntactic heaviness is best defined will not be addressed in this paper, but whether it be information structure or length and/or complexity, most of these accounts agree that rightward movement of heavy elements is a means of facilitating processing and parsing. The question that remains to be answered is whether heavy elements can only be moved to the right edge of a sentence. The results from our study suggest that leftward-movement may serve the same purpose.

3 Edge weight rather than end weight

Thráinsson described Left Dislocation in Icelandic [19] 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.

(8) 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.'

The Left-Dislocated constituent is always in the nominative case but the pronominal copy in situ carries the appropriate case.

Prince [14] argued that, in some cases, Left Dislocation is in fact Topicalization where Topicalization isn't possible (e.g. if the extraction site is in a relative clause). She described Left Dislocation, at least in those instances, as a means to amend a situation where grammatical processing is difficult or impossible.

For our study, we searched Icelandic Parsed Historical Corpus (IcePaHC)for examples of Left-Dislocated Subjects and Direct Objects and Topicalized Direct and Indirect Objects. We compared the average length of the moved constituents vs. the average length of constituents left in situ in each case.

The aim of this study is to demonstrate that heavy syntactic constituents are not only moved to the right edge, as discussed in section 2, but may also be moved to the left edge, e.g. by Left Dislocation or Topicalization.

Or first search was for Subjects moved by Left Dislocation vs Subjects in situ. Or search gave 34191 examples, 193 of which had Left-Dislocated Subjects (such as example (9)) and we found a significant difference in the length between the two.

(9) en fiskarnir sem þar inni lifa, þeir eru þó ekki saltir
but the fish that there inside life, they are though not salty
'But the fish that live in there are not salty.'
(ID 1720.VIDALIN.REL-SER,.53)[2]

[2] Examples from the IcePaHC corpus are shown along with their unique tree ID in parentheses.

Our search showed that subjects in situ had the average length of μ: 2,1, whereas Left-Dislocated subjects had the average length of μ: 9,6, as shown in Fig. 1 (Mann-Whitney U test: U = 77105, p<0.001).

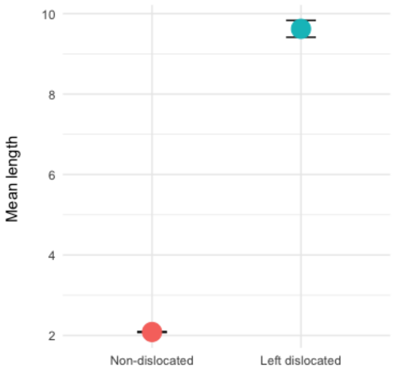

Fig. 1. Subject length by Left Dislocation

As we see in Fig. 1, Left-Dislocated subjects are not only considerably longer by number of words on average than subjects in situ, they also tend to be very long in general. These results suggest that very long subjects are more likely to be moved out of the main clause by Left Dislocation. The moved subjects would have been on the left edge of the sentence anyway if they hadn't been moved. We wanted to know what happens with long constituents that are further away from the clause initial position.

4 Move left only if not already on the right edge

Our second search was for Direct Objects that have been moved by Left Dislocation, such as example (10). We found 25005 examples, 28 of which had Left Dislocated Objects.

(10) [Þau orð] eg tala til yðar þau tala eg ei af sjálfum
[those words.ACC] I speak to you they speak I not from self
mér
me
'The words I speak to you, I speak not from myself'
(ID 1593.EINTAL.REL-OTH,.1039)

Similarly to the left dislocated subjects, we found a significant difference in the average length of left dislocated direct objects (μ: 8) and direct objects in situ (μ: 2,57) (Mann-Whitney U test: U = 614480, p<0.001).

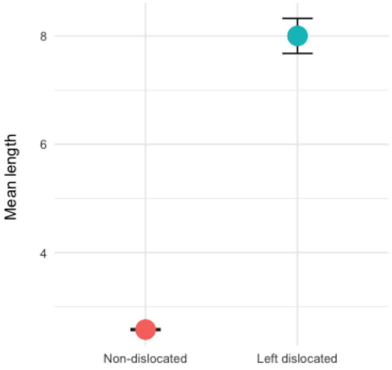

Fig. 2. Direct Object length by Left Dislocation

These search results confirm that both subjects and direct objects that are moved by Left Dislocation tend to be very long and, on average, considerably longer than the ones left in situ.

Our next search was for examples with Topicalized vs. Non-Topicalized constituents. First we looked for Topicalized Direct Objects, like we see in excample (11) vs. Direct Objects in situ. We found 11688 examples, 1070 of which had Topicalized Objects. Our search revealed that Non-Topicalized direct objects tend to be significantly longer (μ: 2,6) than Topicalized ones (μ: 1,9) (Mann-Whitney U test: U = 5442000, p=0.0128).

(11) [Öllum þessum móðgunum] tóku landsmenn með þögn og
[All these insults.DAT] took countrymen with silence and
þolinmæði
patience
'All these insults countrymen took with silence and patience.'
(ID 1907.LEYSING.NAR-FIC,.521)

Although the length difference is nowhere near as great as in the Left-Dislocated examples, it is still significant and, interestingly, it is the opposite to what we've previously seen, as the shifted constituents are, in this case, shorter than the constituents in situ.

We hypothesize that this could be explained by the fact that Direct Objects in Icelandic tend to already be located on the right edge (11a), whereas Indirect Objects are usually found in the middle, between the verb and Direct Object (12b).

(12) a. Pétur borðaði [hafragraut].
Peter ate [porridge]
'Peter ate porridge.'

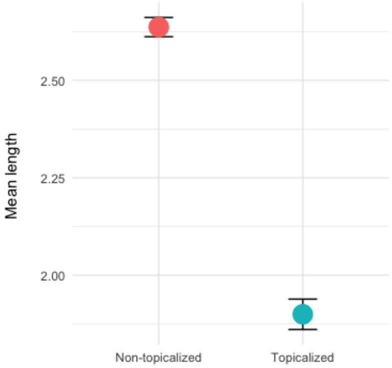

Fig. 3. Direct Object length by Topicalization

b. Pétur gaf [Maríu] hafragraut.
Peter gave [Mary] porridge
'Peter gave Mary porridge.'

We decided to also search for Topicalized Indirect Objects vs Indirect Objects in situ. Our search gave 2012 examples, out of which 57 had Topicalized Indirect Objects, and it revealed the opposite results to the Direct Objects: that Topicalized indirect objects include a larger number of words (μ: 2,6) than those left in situ (μ: 1,5a) (Mann-Whitney U test: U = 77105, p<0.001).

(13) [Þeim sem við hallardyr sat] gaf hann digran gullhring ...
[They.DAT that at palace door sat] gave he thick golden ring
'Those that sat by the palace door he gave a thick golden ring.'
(ID 1480.JARLMANN.NAR-SAG,.790)

To briefly summarize, our main findings were that constituents moved to the left edge by Left Dislocation, Subjects and Direct Objects, tend to be very long (on average they include more than 8 words) and significantly much longer than constituents left in their original place. Topicalized Indirect Objects follow the same pattern, although the average length difference is much smaller, whereas Topicalized Direct Objects are significantly shorter by average number of words than Direct Objects in situ.

From these results we have drawn the following conclusions:

(14) a. Leftward movement, in particular Left Dislocation, is used to move heavy elements to the left edge of the sentence, similarly to rightward movement.
b. 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.

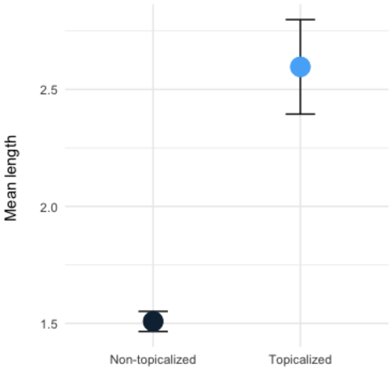

Fig. 4. Indirect Object length by Topicalization

c. 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.

5 Conclusion

We found that 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. Some more general implications involve the fact that our study illustrates how digital parsed corpora of historical languages are useful for studying processing effects. Of course, while our results are already interesting, the effects needs to be studied in more detail in experiments, taking into account other variables and painting a clearer picture of how similar heaviness-driven leftward movement is to heaviness-driven rightward movement. A further avenue of future inquiry is to explore in more detail the relationship between Left Dislocation and Topicalization in the light of analyses such as the one presented by Prince. [14].

The main point here to show that movement to both edges is associated with heaviness, not just to the right edge. Understanding what exactly the parsing problem is is a bigger problem for a bigger research program. However, it intuitively seems on the surface that both types of movement to the edge, i.e., to the left and right, amend some kind of a processing difficulty – and that moving heavy elements to the edge can sometimes ameliorate the situation.

References

[1] Arnold, J.E., Losongco, A., Wasow, T., Ginstrom, R.: Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. Language **76**(1), 28–55 (2000)

[2] Behaghel, O.: Beziehungen zwischen umfang und reihenfolge von satzgliedern. Indogermanische Forschungen **25**, 110 (1909)

[3] Bever, T.G.: The cognitive basis for linguistic structures. Cognition and the development of language **279**(362), 1–61 (1970)

[4] Chomsky, N.: The logical structure of linguistic theory. Plenum press New York (1975)

[5] Frazier, L., Fodor, J.D.: The sausage machine: A new two-stage parsing model. Cognition **6**(4), 291–325 (1978)

[6] Hawkins, J.A.: A parsing theory of word order universals. Linguistic inquiry **21**(2), 223–261 (1990)

[7] Hawkins, J.A.: A performance theory of order and constituency. Cambridge University Press (1994)

[8] Jónsson, J.G.: Definites in Icelandic existentials. The Nordic Languages and Modern Linguistics X pp. 125–134 (2000)

[9] Jónsson, J.G.: Merkingarhlutverk, rökliðir og fallmörkun. In: Setningar III. Almenna bókafélagið (2005)

[10] Kayne, R.S.: The antisymmetry of syntax. MIT Press, Cambridge (1994)

[11] Kimball, J.: Seven principles of surface structure parsing in natural language. Cognition **2**(1), 15–47 (1973)

[12] Kuno, S., Takami, K.i.: Grammar and discourse principles: Functional syntax and GB theory. University of Chicago Press (1993)

[13] Milsark, G.: Toward an explanation of certain peculiarities of the existential construction in English. Linguistic Analysis **3**, 1–29 (1977)

[14] Prince, E.F.: On the limits of syntax, with reference to left-dislocation and topicalization (1998)

[15] Ross, J.R.: Constraints on variables in syntax. Ph.D. thesis, Massachusetts Institute of Technology (1967)

[16] Shih, S., Grafmiller, J.: Weighing in on end weight. In: annual meeting of the Linguistic Society of America (2011)

[17] Stallings, L.M., MacDonald, M.C.: It's not just the "heavy np": relative phrase length modulates the production of heavy-np shift. Journal of psycholinguistic research **40**(3), 177–187 (2011)

[18] Szmrecsanyi, B.: On operationalizing syntactic complexity. In: Le poids des mots. Proceedings of the 7th international conference on textual data statistical analysis. Louvain-la-Neuve. vol. 2, pp. 1032–1039 (2004)

[19] Thráinsson, H.: On complementation in Icelandic. Garland, New York (1979)

[20] Wallenberg, J.: Antisymmetry and the Conservation of C-Command: scrambling and phrase structure in synchronic and diachronic perspective. Ph.D. thesis, University of Pennsylvania (2009)

[21] Wasow, T.: Remarks on grammatical weight. Language variation and change **9**(01), 81–105 (1997)

[22] Wasow, T.: End-weight from the speaker's perspective. Journal of Psycholinguistic research **26**(3), 347–361 (1997)

[23] Wasow, T., Arnold, J.: Intuitions in linguistic argumentation. Lingua **115**(11), 1481–1496 (2005)