

Modeling the acquisition of island constraints: a corpus study of Norwegian

Anastasia Kobzeva¹ and Dave Kush^{1,2}

¹Norwegian University of Science and Technology (NTNU), ²University of Toronto

anastasia.kobzeva@ntnu.no

Background & Method. Languages allow filler-gap dependencies (FGDs), where a filler (such as *what* in 1) is interpreted as an argument of a distant verb (e.g. *bought*). There are, however, environments, known as *islands* [1], where FGDs are judged unacceptable. For instance, English embedded questions (EQs, 2) are typically considered islands. Importantly, there appears to be cross-linguistic and cross-dependency variation [2] in what domains are islands. For example, Norwegian reportedly allows FGDs into EQs, relative clauses (RCs), and some adjuncts [3-5]; (3) is an example of relativization from an EQ.

- (1) What did Mary hear that John bought ____?
(2) *What did Mary wonder [who bought ____]?
(3) Det var det (som_i) jeg ikke skjønte [hva_k ____ var ____].
That was the RP I NEG understood what was
“That was the thing that I didn’t understand what ____ was.”

A long-standing question is how learners induce knowledge of what environments are islands in their languages from exposure to everyday language. Pearl and Sprouse [6] proposed a computational learning model for acquiring the set of acceptable FGDs from distributional regularities in the input to children. Their method included computing the frequency of trigrams of container nodes – select phrase structure (PS) categories along the path from filler to gap, such as *IP-VP-CP-IP-VP* in (4) – in child-directed utterances with *wh*-dependencies. Pearl and Sprouse trained their model on five English corpora from CHILDES [7]. They operationalized the acceptability of a FGD as the probability of its container node sequence and evaluated the predicted probabilities of *wh*-FGDs against English native speaker judgments (Table 1). The model predicted that island-violating FGDs (Table 1, d) were significantly less probable than non-island violating FGDs (Table 1, a-c), closely aligning with human judgments. The authors thus argued that a trigram-based algorithm could induce island constraints from input distributions.

- (4) [_{CP} What did [_{IP} Mary [_{VP} hear [_{CP} that [_{IP} [John [_{VP} bought ____]]]]]]?

Norwegian learners must master a different set of islands compared to English, and this set might differ across dependency type [2-5]. Following Pearl and Sprouse, we tested whether a similar computational learner could successfully learn the correct set of Norwegian *wh*-dependencies first. We trained a model on the Norwegian children’s fiction subcorpus of NorGramBank, a treebank annotated using Lexical-Functional Grammar [8] and computed a frequency distribution over labelled nodes along the path of *wh*-FGDs (n=20497). Instead of PS categories, we calculated the trigram probability of nodes in the LFG functional structure to estimate the probability of four island types (Complex NP, Subject, Adjunct, and *Whether*-islands).

Results & Discussion. Somewhat unexpectedly, the algorithm induced strong island effects in Norwegian, comparable to the results of our replication study for English (Figure 1 and Table 2). These results suggest that the input distributions of *wh*-FGDs are similar in Norwegian and English: in neither language is there significant evidence of ‘island violations’ with *wh*-FGDs. If Norwegians do learn that some constituents are not islands from the input, learning must be driven by observing island violations with different (*non-wh*) FGDs, such as relativization (as in 4) or topicalization. Therefore, our next step will be to apply the model to relativization FGDs in the corpus to find out whether there is a cross-dependency variation, and whether the reported pattern of Norwegian island judgements can be induced based on relativization dependencies.

Table 1. Example stimuli set using 2x2 factorial design (manipulating distance between the filler and the gap and presence of an island structure).

	Sentence	Distance	Structure
a.	Who ___ heard that John bought a house?	Short	No island
b.	What did Mary hear that John bought ___?	Long	No island
c.	Who ___ heard [the rumor that John bought a house]?	Short	Island
d.	*What did Mary hear [the rumor that John bought ___]?	Long	Island

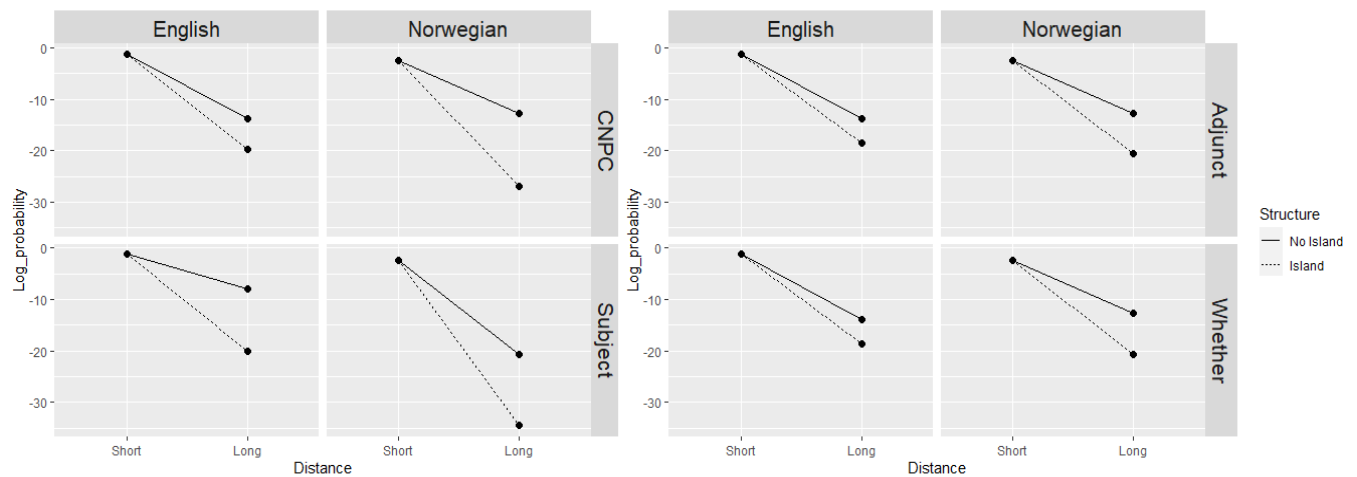


Figure 1. Modeling results for English and Norwegian.

Table 2. Differences-in-differences (DD) scores for island types (based on induced log probabilities from Table 1).

Language/Island type	CNP	Subject	Adjunct	Whether
English	5.97	12.28	4.7	4.7
Norwegian	14.26	13.82	7.95	7.95

References

- [1] Ross, (1967). Doctoral dissertation, MIT; [2] Abeillé, Hemforth, Winckel, & Gibson (2020). *Cognition*; [3] Kush, Lohndal, & Sprouse (2018). *NLLT*; [4] Kush, Lohndal, & Sprouse (2019). *Language*; [5] Bondevik, Kush, & Lohndal (2020). *Nordic Journal of Linguistics*; [6] Pearl & Sprouse (2013). *Language Acquisition*; [7] MacWhinney, B. (2000). Psychology Press; [8] Dyvik, Meurer, Rosén, De Smedt, Haugereid, Losnegaard, ... Thunes (2016). In *LREC*.