Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian

Anastasia Kobzeva
Norwegian University of Science and Technology
anastasia.kobzeva@ntnu.no

Suhas Arehalli
Johns Hopkins University
suhas@jhu.edu

Tal Linzen
New York University
linzen@nyu.edu

Dave Kush
University of Toronto
dave.kush@utoronto.ca

Abstract

Recent research suggests that Recurrent Neural Networks (RNNs) can capture abstract generalizations about filler-gap dependencies (FGDs) in English and so-called *island* constraints on their distribution (Wilcox et al., 2018, 2021). These results have been interpreted as evidence that it is possible, in principle, to induce complex syntactic knowledge from the input without domain-specific learning biases. However, the English results alone do not establish that island constraints were induced from distributional properties of the training data instead of simply reflecting architectural limitations independent of the input to the models. We address this concern by investigating whether such models can learn the distribution of acceptable FGDs in Norwegian, a language that is sensitive to fewer islands than English (Christensen, 1982). Results from five experiments show that Long Short-Term Memory (LSTM) RNNs can (i) learn that Norwegian FGD formation is unbounded, (ii) recover the island status of temporal adjunct and subject islands, and (iii) learn that Norwegian, unlike English, permits FGDs into two types of embedded questions. The fact that LSTM RNNs can learn cross-linguistic differences in island facts therefore strengthens the claim that RNN language models can induce the constraints from patterns in the input.

1 Introduction

Human linguistic knowledge is complex and abstract, yet children master language relatively easily and quickly through exposure to their native language(s). A major debate centers around whether acquiring such knowledge requires complex domain-specific learning biases or whether it can be induced from the input using domain-general learning routines. We contribute to this debate by investigating whether Recurrent Neural Networks (RNNs), which are weakly biased language models, can induce complex knowledge of filler-gap dependencies and constraints on them from the input in Norwegian.

Filler-Gap Dependencies (FGDs) are contingencies between a displaced filler phrase and a later gap position where the filler is interpreted (denoted with __ throughout the paper). There are different types of FGDs. (1-a) is a *wh*-FGD where the filler *wh*-word is interpreted as the direct object of the verb *forged*. (1-b) is a Relative Clause (RC) FGD where the filler, the head of the RC, *painting*, is interpreted as the direct object of *forged* within the RC.

(1) a. They found out what the dealer forged __ using a new technique.
b. They found the painting that the dealer forged __ using a new technique.

FGDs have been the subject of extensive research because they require complex hierarchical generalizations about sentence structure to be interpreted. For example, establishing the RC FGD in (1-b) requires (i) identifying the head of the RC as a filler corresponding to a later empty NP position; (ii) knowing that *forged* requires a direct object; (iii) identifying the gap by recognizing the absence of an object next to *forged*, and (iv) associating the filler with the gap to form a dependency. There is a bidirectional relationship between the filler and the gap: fillers require gaps to be interpreted, and gaps require fillers to be properly licensed. This relationship can be established across a potentially unbounded structural distance as in (2).

(2) She knows what he thought they found out the dealer forged __ using a new technique.

FGDs are also constrained. Certain environments, called *islands* (Ross, 1967), block FGD formation. Various structures have been identified

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 175-185.
Amherst, Massachusetts, June 15-17, 2023

as islands. For example, embedded questions (3-a), sentential subjects (3-b), and adjuncts (3-c) are generally considered island domains in English.

(3) a. *What did he wonder [whether the dealer forged __]?
b. *What is [that the dealer forged __] extremely likely?
c. *What does the dealer worry [if they find out __]?

How do learners acquire island constraints? Nativist approaches hold that acquisition of islands would be impossible without innate domain-specific learning biases due to the induction problem known as the Poverty of the Stimulus (PoS; e.g., Chomsky 1986; Crain and Pietroski 2001). According to this argument, the input to the learner lacks direct evidence that islands exist. The input is therefore compatible with conflicting hypotheses about whether islands should be in the adult target state. The fact that learners nevertheless converge on the same set of island constraints has led the proponents of the nativist approach to suggest that innate domain-specific learning biases guide learners to the conclusion (for example, Subjacency Condition, Chomsky 1973).

Empiricist approaches, on the other hand, claim that the input is sufficiently rich to support learning island constraints when coupled with domain-general learning biases (Clark and Lappin, 2010). This position has recently gained support from neural network simulations. Wilcox and colleagues suggest that RNNs (and other autoregressive neural models) can capture the abstract generalizations governing *wh*-FGDs in English, as well as the associated island constraints (2018; 2019b; 2019a; 2021). They claim that this result militates against the PoS argument that islands cannot be induced from the input without domain-specific biases.

Wilcox and colleagues' results are suggestive, but they do not fully establish that the models 'learn' islands from the input. An alternate explanation is that the results are artifacts. Under this possibility, RNNs do not pursue FGDs into islands in English because the models are simply incapable of representing syntactic dependencies into island environments irrespective of the input they receive (either because the domains are too complex or because of some other unknown limitation inherent to the RNN architecture). One way of ruling out this explanation is to test the models' performance on a language that has a different set of island constraints. If the models can learn to pursue FGDs in another language into domains that are islands in English, that would constitute additional evidence that the models are inducing islands from the input.

To this end, we explore whether RNNs can learn the distribution of acceptable FGDs and island constraints in Norwegian – a language that differs from English in the set of domains that are islands. To preview our results, the models can learn that temporal adjuncts and subject phrases are islands in Norwegian, but that embedded questions are not (*wh*-islands). These results suggest that weakly-biased RNNs can capture patterns of island-insensitivity in Norwegian, thus providing empirical evidence that this pattern of cross-linguistic variation can be learned from the input.

2 Island constraints in Norwegian

Norwegian is similar to English in several respects when it comes to FGDs. Norwegian allows long-distance dependencies with gaps in various syntactic positions. Norwegian also exhibits sensitivity to some of the same islands that English does. FGDs into temporal adjuncts (4) or subject phrases (5) are unacceptable in Norwegian like English (Bondevik et al., 2021; Kush et al., 2019, 2018; Kobzeva et al., 2022b).

(4) *Hva spiste du kake [da han spiste __]?
What ate you cake when he ate __
*'What did you eat cake when he ate __?'

(5) *Hva har [brevet om __] skapt problemer?
What has letter.DEF about __ created problems
*'What has the letter about __ created problems?'

On the other hand, Norwegian allows FGDs into environments that are considered islands in English, such as Embedded Questions (EQs, Christensen 1982; Maling and Zaenen 1982). RC FGDs into embedded constituent questions like (6) are found in written corpora of Norwegian (Kush et al., 2021) and native speakers rate various types of FGD into EQs as acceptable in judgment studies (Kobzeva et al., 2022b).

(6) Vi var redde for noe vi ikke visste [hva __ var].
We were afraid of smth we NEG knew what __ was.
'We were afraid of something we did not know what __ was.'

This distribution of FGDs in Norwegian makes it a good testing ground for exploring whether RNNs can induce a set of islands that is different from what is observed in English. Recent research shows that RNNs can capture basic generalizations about *wh*- and RC FGDs in Norwegian: they learn that fillers can license gaps in different syntactic positions and across increased linear distance between the filler and the gap (Kobzeva et al., 2022a). Here we expand on this line of research by testing whether RNNs can learn that FGDs like (6) are acceptable in Norwegian, while simultaneously ruling out FGDs like (4) and (5). We do so by testing whether the models are less likely to expect FGDs in potential island environments relative to control sentences without island structures. We also test the robustness of the result by testing two more models with the same architecture but different initializations.

We ran five experiments. Experiment 1 tested whether the models learn that Norwegian FGDs are unbounded by seeing if they can successfully associate fillers and gaps across multiple embedded clauses. Establishing this basic result is a prerequisite for testing islands, which typically require cross-clausal dependencies. Experiments 2 and 3 tested if the models can learn that temporal adjunct clauses and complex subject phrases are islands in Norwegian, as in English. Finally, Experiments 4 and 5 tested if RNNs can learn that FGDs into embedded questions are possible in Norwegian. Experiments 1-4 evaluate the models performance on Norwegian only, while Experiment 5 directly compares *wh*-FGDs in Norwegian and English.

3 Method

3.1 Language models

We trained Long Short-Term Memory (LSTM) RNNs (Hochreiter and Schmidhuber, 1997) to take a sequence of words as input and compute a probability distribution of the next word over the model's vocabulary. We trained three such models with different random initializations following the procedure described in (Gulordava et al., 2018), using the code provided by the authors[1]. Each model was a 2-layer LSTM with 650 hidden units in each layer, trained for 40 epochs on 113 million tokens of Norwegian Wikipedia (in the Bokmål written standard) with a vocabulary size of 50000 most frequent words. The models achieved perplexities between 30.05 and 30.3 on the validation set.

3.2 Dependent measure

We test how the models would fare as incremental language processors by looking at *surprisal*, which measures how (un)predictable a word is given a specific prompt using the models' probability distribution. We measure the surprisal values by computing the negative log of the predicted conditional probability from the models' softmax layer.

3.3 Measuring FGDs

Wilcox et al. (2018) introduced a 2×2 factorial design for measuring FGDs inspired by psycholinguistic paradigms. The design independently manipulates the presence of a filler and the presence of a gap as in (7).

(7) They found out...
a. that the dealer forged the art -FILLER, -GAP
b. *what the dealer forged the art +FILLER, -GAP
c. *that the dealer forged __ -FILLER, +GAP
d. what the dealer forged __ +FILLER, +GAP
...using a new technique.

When both the filler and the gap are absent (7-a) or present (7-d), the sentences are grammatical. When either the filler or the gap is absent, (7-b) and (7-c), the sentences are ungrammatical. We measure *filler effects* – how the presence of a filler affects surprisal – in two different pairwise comparisons. *Filled gap effects* are measured by comparing surprisal associated with an NP in -GAP conditions. *Unlicensed gap effects* are measured by comparing surprisal associated with a gap in the +GAP conditions. We discuss each type of filler effect in more detail below.

3.3.1 Filled gap effects

In behavioral studies, filled gap effects are regarded as support for the *active gap-filling* strategy: after encountering a filler, the processor actively predicts a gap without waiting for the actual gap site. Stowe (1986) observed a slow-down in self-paced reading times at the direct object *us* in (8-b), which contains the filler *who*, compared to the same word in a corresponding sentence without a filler (8-a). The slow-down reflects a violated expectation: seeing a filler caused the processor to predict a gap in object position.

(8) a. My brother wanted to know if Ruth will bring *us* home to Mom at Christmas.
b. My brother wanted to know who Ruth will bring *us* home to __ at Christmas.

We test whether the models exhibit similar filled gap effects. We measure the surprisal difference between the ungrammatical +FILLER, -GAP condition as in (7-b) and the grammatical -FILLER, -GAP condition in (7-a) at the region of the filled NP (*the*

[1] https://github.com/facebookresearch/colorlessgreenRNNs

art in (7)). If seeing a filler sets up an expectation for a gap in object position, the NP should be more surprising in (7-b) than in (7-a), resulting in a *positive* surprisal difference.

Crucially, humans do not exhibit filled gap effects inside island environments (Stowe, 1986; Traxler and Pickering, 1996; Phillips, 2006), indicating that the active prediction of gaps is suspended where they are impossible. Following the same logic, if the models show sensitivity to island constraints, we expect to see no filled gap effects inside islands.

3.3.2 Unlicensed gap effects

Unlicensed gap effects provide a measure of how 'surprised' the model is to encounter a gap without a filler to license it. We measure these effects as a difference in surprisal between the grammatical +FILLER, +GAP (7-d) condition and ungrammatical -FILLER, +GAP (7-c) condition at the region following the gap (*using a new technique* in (7)). If a presence of a gap without a licensing filler is surprising to the models, the unlicensed gap effect should manifest as a negative difference between low surprisal in the post-gap region in (7-d) and high surprisal in (7-c).

Unlicensed gap effects show if the models recognize gaps as licit inside certain syntactic environments. Whereas filled gap effects measure the models' expectation for an upcoming gap, unlicensed gap effects arguably should reflect the models' understanding of grammaticality, as sentences with illicit gaps are ungrammatical (and, unlike filled gaps, cannot be 'rescued' by establishing another gap site later in a sentence). Analogous to filled gap effects, unlicensed gap effects should be close to zero in island environments if the models can derive their island status from their training data.

3.4 Statistical analysis

Following standard practice in psycholinguistics, statistical analysis was performed using mixed-effect linear regression models with sum-coded fixed effects of FILLER (0.5 for +FILLER, -0.5 for -FILLER) and CONDITION (0.5 for CONTROL and -0.5 for ISLAND except for Experiments 1 and 4, see details below). We fit the statistical models on differences in surprisal between +FILLER, -FILLER conditions with these fixed effects and a maximal random effect structure (Barr et al., 2013). We ran separate models for filled gap effects in the filled NP region and for unlicensed gap effects in the post-gap region. If a model failed to converge, we reduced the random effect structure until convergence was reached. Model formulas are presented in Appendix A.

4 Experiments

4.1 Experiment 1: Unboundedness

It is important to establish whether LSTMs can represent FGDs across hierarchical distance before testing island environments, as they involve cross-clausal dependencies. Therefore, in Experiment 1 we tested how increased hierarchical distance between the filler and the gap influences models' representations of FGDs. To do that, we manipulated the number of clausal embeddings between the filler and the gap (from 1 to 5 layers of clausal embedding, as illustrated in (9)). We created 30 items by crossing the factors FILLER and GAP in (7) with NUMBER OF LAYERS, resulting in a 2 × 2 × 5 design. Test sets were created for *wh*- and RC FGDs (600 test sentences per dependency type).

(9) a. 1 LAYER (+FILLER, +GAP)
Hun vet hva selgeren forfalsket __ ved hjelp
She knows what dealer.DEF forged __ with help
av moderne teknologi.
of modern technology.
'She knows what the dealer forged __ using modern technology'.

b. 5 LAYERS (+FILLER, +GAP)
Hun vet hva han trodde de fant ut
She knows what he thought they found out
avisen rapporterte politiet visste
newspaper.DEF reported police.DEF knew
selgeren forfalsket __ ved hjelp av moderne
dealer.DEF forged __ with help of modern
teknologi.
technology.
'She knows what he thought they found out the newspaper reported the police knew the dealer forged __ using modern technology'.

We tested all three models on all of the items, and we present the results averaged across the models for both dependency types together. Overall, filler effects decrease as layers of embedding increase (Figure 1). For *wh*-dependencies (blue bars), there was a significant reduction in both the filled gap effect and the unlicensed gap effect already at two layers of embedding, which was also true for every layer thereafter (p's <0.05 in all cases). For RC dependencies (orange bars), there was a significant reduction in filled gap effects at three layers (p <0.05), and in unlicensed gap effects at two layers (p's <0.001) of sentential embedding, as well as for every layer thereafter (p's <0.001 in all cases).

Tables with statistics summary can be found in Appendix A.

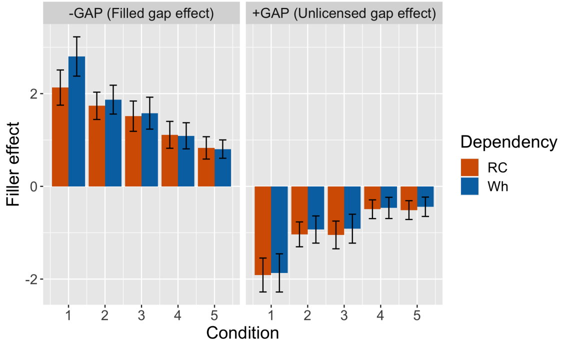

Figure 1: Unboundedness experiment: Filler effects by the number of embeddings for both dependency types. Bars represent an average over three models, error bars represent 95% confidence intervals.

Despite the reduction in filler effects as a function of the number of sentential embeddings, the filler effects remain above zero even at the largest hierarchical distance. This suggests that the models have learned that FGD formation is unbounded and have the basic representational capacity required for testing FGDs inside islands.

4.2 Islands shared between Norwegian and English

Experiments 2 and 3 tested FGDs into constituents that are islands in Norwegian (just as in English) – subjects and temporal adjunct clauses – to see if the models' expectations for FGDs are attenuated within the two environments in Norwegian, as previously seen in English (Wilcox et al., 2018, 2021).

4.2.1 Experiment 2: Subject island

Fillers cannot be associated with gaps inside a subject phrase, like the gap inside the prepositional phrase attached to the subject in (10). Such sentences are rated as unacceptable by English speakers, and the same pattern is found in Norwegian (11-b). We compare the island condition in (11-b) to an NP-subject extraction as in (11-a).

(10) *The newspaper reported what [the agreement with __] will strengthen the political interaction after the elections.

(11) a. SUBJECT CONTROL (+FILLER, +GAP)
Avisen rapporterte hva som __ vil
Newspaper.DEF reported what REL __ will
forsterke det politiske samspillet etter
strengthen the political interaction.DEF after
valget.
election.DEF
'The newspaper reported what __ will strengthen the political interaction after the election.'

b. SUBJECT ISLAND (+FILLER, +GAP)
*Avisen rapporterte hva [avtalen
Newspaper.DEF reported what agreement.DEF
med __] vil forsterke det politiske
with __ will strengthen the political
samspillet etter valget.
interaction.DEF after election.DEF
'*The newspaper reported what the agreement with __ will strengthen the political interaction after the election.'

We created 30 items according to a 2 × 2 × 2 design that crossed the factors FILLER and GAP in (7) with a third factor: CONDITION (CONTROL, ISLAND). Again we created separate sets of sentences for *wh-* and RC FGDs (240 total test sentences per dependency type). The results of this experiment are presented in Figure 2.

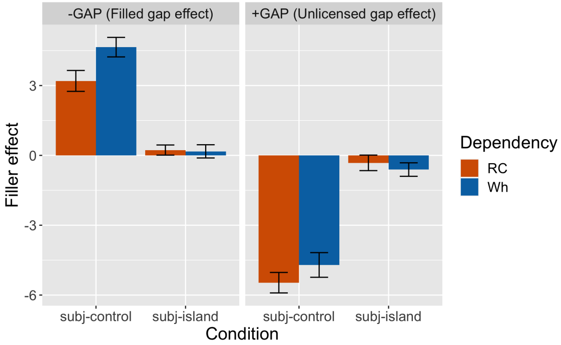

Figure 2: Subject island experiment: Filler effects by gap position for both dependency types.

Filled gap effects (Figure 2 left panel) were large in the control condition, but were significantly reduced in the island condition: statistical analysis revealed a main effect of CONDITION for both dependency types (both p's <0.001). The same pattern was found for unlicensed gap effects (Figure 2 right panel). For both dependency types, there was a significant effect of CONDITION (p's <0.001 in both cases). These results show that the models exhibit reduced filler effects within subject islands, which is in line with behavioral acceptability data from native Norwegian speakers.

4.2.2 Experiment 3: Adjunct island

Adjuncts are said to block FGD formation, which explains the unacceptability of (12): The filler *what* cannot be associated with the gap inside the adjunct *when*-clause. Norwegian, like English, does not al-low gaps inside temporal adjuncts (Bondevik et al., 2021; Bondevik and Lohndal, 2023).

(12) *What were the voters excited [when the politician visited
__ last week]?

We created 30 items according to a 2 × 2 × 3 design that crossed FILLER, GAP, and CONDITION for each dependency type (360 test sentences per dependency). CONDITION had three levels that determined the location of a direct object gap. In the LINEAR CONTROL (13-a) and STRUCTURAL CONTROL (13-b) the gap was not embedded in an island, whereas in ADJUNCT ISLAND (13-c), the gap was embedded inside a temporal adjunct (headed by *mens 'while', da 'when', etter at 'after'* and *før 'before'*). In the linear control condition (13-a), first used in (Wilcox et al., 2018), the filler and gap are in the same clause, but the linear distance between them is comparable to the distance in (13-c). In the structural control condition (13-b), our novel addition to the design, the filler and the gap are separated across two clauses, making the *structural* distance between the filler and the gap comparable to (13-c). We included these control conditions in order to estimate the independent effects of linear distance and structural distance on the model's performance, so as to better isolate island effects.

(13) a. LINEAR CONTROL (+FILLER, +GAP)
Jeg husker hva politikeren med godt
I remember what politician.DEF with good
omdømme besøkte __ forrige uke.
reputation visited __ last week.
'I remember what the politician with a good reputation visited __ last week.'
b. STRUCTURAL CONTROL (+FILLER, +GAP)
Jeg husker hva avisen rapporterte at
I remember what newspaper.DEF reported that
politikeren besøkte __ forrige uke.
politician.DEF visited __ last week.
'I remember what the newspaper reported that the politician visited __ last week.'
c. ADJUNCT ISLAND (+FILLER, +GAP)
*Jeg husker hva velgerne var begeistret
I remember what voters.DEF were excited
da politikeren besøkte __ forrige uke.
when politician.DEF visited __ last week.
'*I remember what the voters were excited when the politician visited __ last week.'

We defined two contrasts for analysis: CONTROL contrast compared effect size between the two control conditions (linear vs. structural). ISLAND contrast compared effects between the structural control and the adjunct island condition.

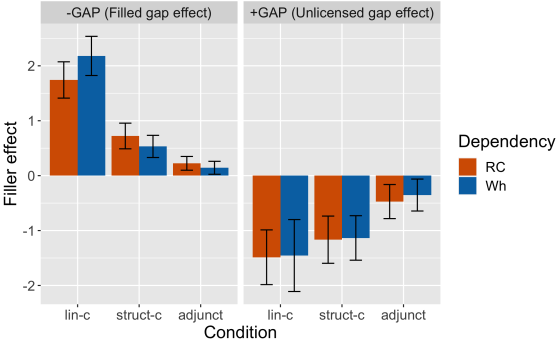

Figure 3: Adjunct island experiment: Filler effects by condition for both dependency types. Control conditions are `lin-c` and `struct-c`.

The results of the experiment are presented in Figure 3. Filled gap effects for both dependency types (left panel) were largest in the linear control condition, significantly larger than in the structural control condition (CONTROL contrast p's <0.001). Filled gap effects were in turn significantly larger in the structural control condition than in the adjunct island condition (ISLAND contrast p's <0.001), where filled gap effects were close to zero.

The same qualitative pattern was observed with unlicensed gap effects for both dependency types (right panel). Unlicensed gap effects were larger in the linear control condition compared to the structural control, and in the structural control condition compared to the island condition (p's <0.001 in all cases). Therefore, the models show reduced filler effects inside temporal adjuncts in Norwegian. However, the average filler effects are not 0 in the adjunct island condition, suggesting that the models might not treat them as full islands.[2] Norwegian shows some variation in adjunct island effects, with extraction from conditional adjuncts rated higher than from temporal and reason-adjuncts (Bondevik et al., 2021; Bondevik and Lohndal, 2023). The result obtained here could be explained by the models' sensitivity to this variation (and potential overgeneralization).

4.3 Islands contrasting English and Norwegian

The results of Experiments 2 and 3 suggest that the models learn that subjects and temporal adjuncts are islands in Norwegian, similar to the conclusions

[2] On around 65% of the trials, the models show filled-gap effects greater than zero, while unlicensed gap effects are less than zero on around 70% of the trials. However, the effects are mostly small, under 1 bit of surprisal 90% of the time.

made for English by Wilcox et al.. Experiments 4 and 5 test whether the models can learn that embedded questions (EQs) are not islands in Norwegian. We test two types of EQs in Norwegian: 1) interrogative EQs, and 2) *whether*-EQs.

4.3.1 Experiment 4: Interrogative EQ

According to Kush et al. (2021), the most common type of extraction from EQs (in a children's fiction corpus) includes a subject gap inside an interrogative EQ as in (14).

(14) Vi var redde for noe vi ikke visste [hva __ var].
We were afraid of smth we NEG knew what __ was.
'We were afraid of something we did not know what __ was.'

We chose to first test such EQs because we reasoned that they were likely the most frequent in the model's training data. We created 30 items that crossed FILLER, GAP, and CONDITION for each dependency type (240 test sentences per dependency). CONDITION controlled whether the embedded clause was an EQ (15-b) or a declarative complement (15-a) control.[3]

(15) a. DECLARATIVE CONTROL (+FILLER, +GAP)
Han sa hvem som sjåføren glemte at __
He said who REL driver.DEF forgot that __
skulle hentes i sentrum den dagen.
should be.picked.up in center.DEF that day.DEF.
'He said who$_i$ the driver forgot (that) __$_i$ should be picked up in the center that day.'

b. WH-ISLAND (+FILLER, +GAP)
Han sa hvem som sjåføren glemte hvor __
He said who REL driver.DEF forgot where __
skulle hentes __ den dagen.
should be.picked.up that day.DEF.
'He said who$_i$ the driver forgot where$_k$ __$_i$ should be picked up __$_k$ that day.'

We expected clear filled gap effects and unlicensed gap effects in the declarative clauses. If the models recognize that interrogative EQs are not islands in Norwegian, the filled gap effects and unlicensed gap effects in the EQ sentences should be comparable to their declarative counterparts, or at least greater than zero.

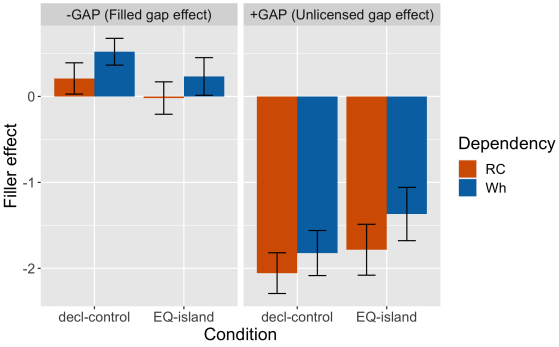

Figure 4: Interrogative EQ island experiment: Filler effects by condition for both dependency types.

Figure 4 shows that filled gap effects were small or close to 0 across all conditions and dependency types, while unlicensed gap effects were large. Statistical analysis revealed a main effect of CONDITION for both filled gap effects and unlicensed gap effects with *wh*-dependencies (p's <0.01). With RC dependencies, the same was true for the filled gap effect (p <0.05, orange bars on the left panel). For the unlicensed gap effect with RC dependencies, the effect of CONDITION was not significant (p <0.1). Importantly, despite the significant effect of CONDITION in three out of four cases tested, both filled gap effects and unlicensed gap effects in the island condition were comparable to the declarative control, suggesting that the models treat EQs and embedded declarative clauses similarly with respect to FGD formation in Norwegian.

4.3.2 Experiment 5: *Whether*-EQ

In Experiment 4, we tested FGDs into interrogative EQs with gaps in subject position. However, previous research in English has not tested interrogative EQs and has instead focused on FGDs into polar EQs, *whether*-islands. For example, Wilcox et al. tested *whether*-islands with gaps in object position in English. An example of +FILLER, +GAP, ISLAND condition from their *whether*-island experiment is presented in (16).

(16) *I know what my brother said whether our aunt devoured __ at the party.

In order to facilitate more direct cross-linguistic comparison, and to test the robustness of the result of Experiment 4, we decided to run an experiment comparing FGDs into *whether*-EQs in English and Norwegian side by side. To do so, we slightly modified the 24 English items from (Wilcox et al., 2018) and created 24 novel items following the same tem-

[3] The direct translation of (15-b) would be ungrammatical in English due to *that*-trace effects. Norwegian exhibits some variation in *that*-trace effects; theoretical and experimental work shows that it mostly allows subject gaps after *that* (Lohndal, 2009; Kush and Dahl, 2020). We return to this issue in the Discussion.

plate, resulting in 48 items total. We then translated them into Norwegian. As the original (Wilcox et al., 2018) items did not include RC dependencies, we restricted dependency types to *wh*-FGDs in this experiment. We compared the performance of the Gulordava model (used by Wilcox et al., 2018) on English stimuli and the performance of one of the Norwegian models (used by Kobzeva et al., 2022a). The results are presented in Figure 5.

Overall, filler effects are smaller in English (light blue bars) than in Norwegian (dark blue bars; main effect of LANGUAGE, p <0.001). The pattern of island sensitivity also differs. In Norwegian, robust filled gap effects were observed in both declarative control and *whether*-island environments, while in English, no filled gap effect was observed inside a *whether*-island (left panel). Statistical analysis confirmed a significant CONDITION × LANGUAGE interaction for filled gap effects (p <0.01). Similar differences were observed for unlicensed gap effects (right panel): In Norwegian, unlicensed gap effects are equally large in declarative complements and *whether*-islands, whereas there is no unlicensed gap effect inside a *whether*-island in English compared to the declarative control (CONDITION × LANGUAGE p <0.05).

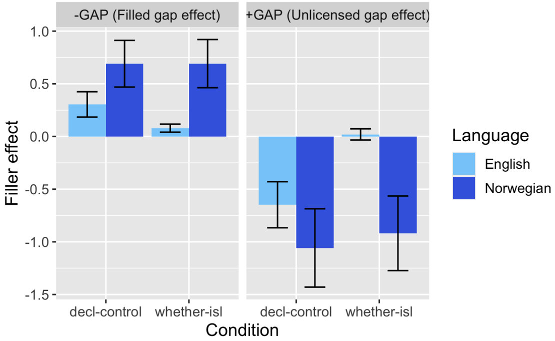

Figure 5: *Whether*-island experiment (with *wh*-dependencies): Comparison of filler effects in English and Norwegian.

Taken together with the fact that the architecture of the English and the Norwegian model was the same, and that they were trained using the same hyper-parameter combination for the same number of epochs on input data that were comparable in size and genre, these results suggest that RNNs can come to different conclusions about the status of *whether*-islands based on different language input. This provides further evidence for the claim, made in Wilcox et al., that autoregressive language models can learn the distribution of FGDs in a language from their input.

5 Discussion

In this paper, we tested LSTMs' ability to establish FGDs in Norwegian by looking at filled gap effects and unlicensed gap effects. Experiment 1 found non-zero filled gap effects and unlicensed gap effects across multiple layers of embedding suggesting that the models learn that FGDs are unbounded. Experiments 2 and 3 showed that filled gap effects and unlicensed gap effects are significantly reduced inside subject phrases and temporal adjuncts, suggesting that the models learned that these domains are islands in Norwegian, mirroring previous findings for English (Wilcox et al., 2018, 2019a,b, 2021).

Broadly speaking, results from Experiments 4 and 5 suggest that the models can learn that embedded questions are not island environments in Norwegian. In both Experiment 4 and 5, we found large unlicensed gap effects in Norwegian interrogative EQs and in Experiment 5 we observed filled gap effects inside Norwegian *whether*-EQs. Taken together, the results are consistent with the conclusion that LSTM RNNs can learn cross-linguistic differences in island facts from different language input. We do not know whether the model's generalization was derived from actual examples of FGDs into embedded questions in the training data, or whether the model learned the distribution indirectly. We cannot verify that in this case that the models learned from direct evidence, but it is plausible that such evidence would be available in the Wikipedia corpus given that FGDs into embedded questions are found (in relatively small numbers) in other corpora (such as the child fiction corpus investigated by Kush et al., 2021).

One potentially surprising finding was the asymmetry in filled and unlicensed gap effects between Experiments 4 and 5. In Experiment 4, filled gap effects were not robust in subject position, but unlicensed gap effects were. In Experiment 5, both filled gap effects and unlicensed gap effects were observed in object position. We take this effect to mean that the model was not actively pursuing embedded subject gaps in our stimuli. There are various possible interpretations for this effect. One possibility is that the model avoids gaps after overt material in left edge of a clause (a kind of *that-trace* effect, see Lohndal, 2009). Another possibility is that embedded subject gaps were not frequent enough in the training data to establish strong expectations for them.

We do not take the fact that filled gap effects are absent in some EQs as evidence against the models being able to establish FGDs into EQs. Even in the absence of filled gap effects, unlicensed gap effects show that the models can still recognize gaps in EQs as licit in Norwegian. We think that unlicensed gap effects provide a better indication of what the models have learned is possible. In other words, the two effects measure different aspects related to an FGD: While filled gap effects measure active expectation/prediction for a gap inside a particular structural configuration (i.e. whether the models think that a gap is *likely* in a given position), unlicensed gap effects reflect whether the models 'understand' that FGDs are in principle possible in that configuration. We suggest that future work using this paradigm should keep this dissociation in mind when interpreting results: Learning what a possible FGD is, does not necessarily entail active expectation in RNN language models.

One outstanding question is how well the model's active gap-filling behavior mirrors how actual humans would process these sentences. Native English speakers do not actively pursue gaps inside islands (Stowe, 1986; Traxler and Pickering, 1996; Phillips, 2006). In this regard, the English models of Wilcox et al. mimic human behavior. It is unknown whether native Norwegian speakers suspend active gap-filling inside islands, but pursue active gap-filling inside structures like EQs, that are not islands in their language. Future work should test the alignment between the model's performance and human behavior.

6 Conclusion

In this study, we tested whether LSTMs, an RNN architecture without language-specific bias, can learn two types of filler-gap dependencies in Norwegian in several (potential) island environments. We found evidence that the models can pick up patterns of island-insensitivity when it comes to embedded questions in Norwegian, while still inducing island effects in subject and adjunct islands. Our results also show that RNNs are sensitive to differences in the distribution of FGDs in English and Norwegian, suggesting that the input to the models must provide enough evidence for the diverging patterns. Our results lead us to reassess the importance of domain-specific learning biases in acquiring island constraints from the input.

Acknowledgements

All experimental materials and analysis scripts are available in the project's OSF repository. Our models were trained with the resources from NTNU's IDUN computing cluster (Själander et al., 2019). We thank the members of NTNU's EyeLands Lab (Øyelab) for their comments on an earlier version of this paper. We also thank Sigurd Farstad Iversen for his help in creating the stimuli in Norwegian.

References

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Ingrid Bondevik, Dave Kush, and Terje Lohndal. 2021. Variation in adjunct islands: The case of Norwegian. *Nordic Journal of Linguistics*, 44(3):223–254.

Ingrid Bondevik and Terje Lohndal. 2023. Extraction from finite adjunct clauses: an investigation of relative clause dependencies in norwegian. *Glossa: a journal of general linguistics*, 8(1).

Noam Chomsky. 1973. Conditions on transformations. In Morris Halle, Stephen R. Anderson, and Paul Kiparsky, editors, *A Festschrift for Morris Halle*, pages 232–286. Holt, Rinehart and Winston, New York.

Noam Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Kirsti Koch Christensen. 1982. On multiple filler-gap constructions in Norwegian. In Elisabet Engdahl and Eva Ejerhed, editors, *Readings on unbounded dependencies in Scandinavian languages*, pages 77–98. Almquist & Wiksell, Stockholm.

Alexander Clark and Shalom Lappin. 2010. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Stephen Crain and Paul Pietroski. 2001. Nature, nurture and universal grammar. *Linguistics and philosophy*, 24(2):139–186.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL 2018*, pages 1195–1205.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2022a. LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Anastasia Kobzeva, Charlotte Sant, Parker T. Robbins, Myrte Vos, Terje Lohndal, and Dave Kush. 2022b. Comparing island effects for different dependency types in Norwegian. *Languages*, 7(3):195–220.

Dave Kush and Anne Dahl. 2020. L2 transfer of L1 island-insensitivity: The case of Norwegian. *Second Language Research*, pages 1–32.

Dave Kush, Terje Lohndal, and Jon Sprouse. 2018. Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory*, 36(3):743–779.

Dave Kush, Terje Lohndal, and Jon Sprouse. 2019. On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language*, 95(3):393–420.

Dave Kush, Charlotte Sant, and Sunniva Briså Strætkvern. 2021. Learning island-insensitivity from the input: A corpus analysis of child- and youth-directed text in Norwegian. *Glossa: a journal of general linguistics*, 6(1):1–50.

Terje Lohndal. 2009. Comp-t effects: Variation in the position and features of C. *Studia Linguistica*, 63(2):204–232.

Joan Maling and Annie Zaenen. 1982. A phrase structure account of Scandinavian extraction phenomena. In Pauline Jacobson and Geoffrey K. Pullum, editors, *The Nature of Syntactic Representation*, pages 229–282. Springer Netherlands, Dordrecht.

Colin Phillips. 2006. The real-time status of island phenomena. *Language*, pages 795–823.

John Robert Ross. 1967. *Constraints on variables in syntax*. PhD dissertation, MIT.

Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. 2019. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure.

Laurie A Stowe. 1986. Parsing wh-constructions: Evidence for on-line gap location. *Language and cognitive processes*, 1(3):227–245.

Matthew J Traxler and Martin J Pickering. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, 35(3):454–475.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pages 181–190.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019b. What syntactic structures block dependencies in RNN language models? *arXiv preprint arXiv:1905.10431*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler-Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pages 211–221.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2021. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–88.

A Results of Statistical Tests

The levels of significance used in the tables below: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The statistics are presented separately for filled gap effects (FGE) and unlicensed gap effects (UGE) by each dependency type and experiment. The response variable *s* in `lmer` formulas is the difference in surprisal between +FILLER, -FILLER conditions.

1. Unboundedness			
$s \sim lyrs + (1+lyrs \mid model) + (1+lyrs \mid item)$			
	FGE, *wh*-dependencies		
	Est.	S.E.	t
(Intercept)	2.801	0.304	9.221***
layers2	−0.931	0.220	−4.240*
layers3	−1.223	0.204	−5.980***
layers4	−1.711	0.246	−6.959***
layers5	−1.997	0.219	−9.104***
	UGE, *wh*-dependencies		
(Intercept)	−1.867	0.147	−12.681***
layers2	0.936	0.099	9.488***
layers3	0.954	0.099	9.671***
layers4	1.402	0.099	14.212***
layers5	1.427	0.099	14.465***
	FGE, RC dependencies		
(Intercept)	2.131	0.194	10.971***
layers2	−0.394	0.281	−1.402
layers3	−0.617	0.237	−2.598*
layers4	−1.019	0.203	−5.024***
layers5	−1.301	0.233	−5.593**
	UGE, RC dependencies		
(Intercept)	−1.912	0.192	−9.954***
layers2	0.877	0.161	5.447***
layers3	0.864	0.156	5.557***
layers4	1.419	0.166	8.564***
layers5	1.400	0.158	8.885***

2. Subject island

s ~cond + (1+cond | model) + (1+cond | item)

	Est.	S.E.	t
	FGE, *wh*-dependencies		
(Intercept)	2.411	0.255	9.459***
condition	4.476	0.335	13.368***
	UGE, *wh*-dependencies		
(Intercept)	−2.658	0.255	−10.437***
condition	−4.098	0.488	−8.390***
	FGE, RC dependencies		
(Intercept)	1.713	0.132	12.944***
condition	2.970	0.254	11.697***
	UGE, RC dependencies		
(Intercept)	−2.895	0.223	−13.008***
condition	−5.147	0.383	−13.455***

3. Adjunct island

s ~cntrs + (1+cntrs | model) + (1+cntrs | item)

	Est.	S.E.	t
	FGE, *wh*-dependencies		
(Intercept)	0.952	0.127	7.476***
controlCntrs	2.457	0.232	10.609***
islandCntrs	1.618	0.221	7.323***
	UGE, *wh*-dependencies		
(Intercept)	−0.981	0.208	−4.714***
controlCntrst	−0.948	0.298	−3.182**
islandCntrst	−1.255	0.273	−4.602***
	FGE, RC dependencies		
(Intercept)	0.896	0.136	6.593***
controlCntrst	1.692	0.200	8.454***
islandCntrst	1.344	0.234	5.755***
	UGE, RC dependencies		
(Intercept)	−1.042	0.187	−5.569***
controlCntrst	−0.889	0.201	−4.423***
islandCntrst	−1.139	0.178	−6.382***

4. Interrogative EQ

s ~cond + (1+cond | model) + (1+cond | item)

	Est.	S.E.	t
	FGE, *wh*-dependencies		
(Intercept)	0.376	0.081	4.647***
condition	0.288	0.107	2.690**
	UGE, *wh*-dependencies		
(Intercept)	−1.595	0.260	−6.142***
condition	−0.454	0.153	−2.961**
	FGE, RC dependencies		
(Intercept)	0.095	0.080	1.189
condition	0.228	0.100	2.271*
	UGE, RC dependencies		
(Intercept)	−1.920	0.220	−8.707***
condition	−0.272	0.152	−1.795+

5. *Whether*-EQ

*s ~condition*language + (1+condition | item)*

	Est.	S.E.	t
	FGE		
(Intercept)	0.617	0.074	8.388***
condition	0.109	0.102	1.074
language	0.700	0.102	6.880***
condition:language	−0.625	0.204	−3.073**
	UGE		
(Intercept)	−0.652	0.099	−6.570***
condition	−0.354	0.132	−2.690**
language	−0.676	0.127	−5.346***
condition:language	0.627	0.253	2.477*