Generalizations across filler-gap dependencies in neural language models

Katherine Howitt[1]   Sathvik Nair[1,3]   Allison Dods[1]   Robert Melvin Hopkins[1,2]

[1]Department of Linguistics, University of Maryland
[2]Department of Computer Science, University of Maryland
[3]University of Maryland Institute for Advanced Computer Studies
{kghowitt,sathvik}@umd.edu

Abstract

Humans develop their grammars by making structural generalizations from finite input. We ask how filler-gap dependencies, which share a structural generalization despite diverse surface forms, might arise from the input. We explicitly control the input to a neural language model (NLM) to uncover whether the model posits a shared representation for filler-gap dependencies. We show that while NLMs do have success differentiating grammatical from ungrammatical filler-gap dependencies, they rely on superficial properties of the input, rather than on a shared generalization. Our work highlights the need for specific linguistic inductive biases to model language acquisition.

1 Introduction

Human learners use their linguistic environment to acquire a grammar. At the same time, they come to generalizations that are not obviously signaled in the input. The central puzzle in language acquisition is to characterize the system that allows for human-like generalizations from finite input. Linguists posit that these generalizations are achieved through shared representations that allow learners to treat superficially distinct phenomena as a class (Chomsky, 1977; Kaplan and Bresnan, 1982; Gazdar, 1982; Gazdar et al., 1985; Pollard and Sag, 1987; Postal, 1999). The recent success of neural language models (NLMs) has caused many to question the necessity of linguistically-specific representational systems in language learning (Wilcox et al., 2018, 2023; Piantadosi, 2023).

We address this renewed controversy by conducting two experiments to uncover whether NLMs posit a shared representation for a particular syntactic dependency: filler-gap dependencies. We consider whether an NLM recognizes filler-gap dependencies in superficially distinct con-structions, as humans do (Crain and Fodor, 1985; Stowe, 1986; Bever and McElree, 1988; Traxler and Pickering, 1996; Sprouse et al., 2016). We further ask whether the NLM posits a shared representation for filler-gap dependencies, and thus systematically applies constraints across them.

Recent research shows NLMs can differentiate between grammatical and ungrammatical instances of filler-gap dependencies in individual constructions, but our study asks whether filler-gap dependencies are treated as a *class* by the NLM. If a shared structural relation is learnable by an NLM, which lacks language-specific biases, then, in principle, a learner does not need to have such biases to learn that relation. Although one could learn the correct pattern through piecemeal learning of each construction individually (given enough input), a shared representation across filler-gap dependencies would allow a learner to generalize from only a subset of constructions containing filler-gap dependencies. Whether an NLM posits this shared representation is the question.

We provide an NLM with direct evidence for a filler-gap dependency in one construction and test whether it generalizes to other constructions. In our first experiment, we augment an NLM's training data with specific instances of clefting, and in our second, topicalization. We compare performance on four constructions containing filler-gap dependencies: Wh-movement, clefting, tough-movement, and topicalization. The NLM treating filler-gap dependencies systematically would be evidence that this shared representation is learnable without language-specific inductive biases.

2 Filler-gap dependencies

Filler-gap dependencies share a set of properties across superficially distinct constructions, including sensitivity to islands(Chomsky, 1977). These

properties persist across constructions, despite variation in semantic contribution and discourse function (Schütze et al., 2015). In psycholinguistic experiments, humans have been shown to be sensitive to gaps across filler-gap dependencies, including wh-movement (Crain and Fodor, 1985; Stowe, 1986), tough-movement (Bever and McElree, 1988), and clefting (Traxler and Pickering, 1996), though see Sprouse et al. (2016) for variation in English relative clauses. These effects are sensitive to locality constraints, and appear to be mediated by the presence of islands (Phillips, 2006; Traxler and Pickering, 1996; McElree and Griffith, 1998; Omaki et al., 2015). Generalizing from surface forms on the basis of a shared representation could be critical to learning, especially if some constructions containing filler-gap dependencies do not occur frequently in the input.

Clefting (1) is one construction that contains a filler-gap dependency. In (1a), the filler *these snacks* forms a dependency with the gap site, marked with __ for readability, but silent in natural language. The filler is interpreted as the object of *bought*, despite not appearing linearly beside *bought* in the string. Strings lacking a filler but containing a gap (1b) are ungrammatical, and when the gap is filled (i.e., an object, such as *cheese*, immediately follows the verb), the acceptability pattern reverses (1c-d). In other words, neither a filler nor a gap can occur without the other. Importantly, clefts are superficially similar to sentences like (1d) which lack a filler-gap dependency, and thus are structurally quite distinct. A learner must distinguish between instances of clefting (1a) and other superficially similar sentences (1d).

(1) a. It is *these snacks* that Mary bought __ today.
     b. * It is *apparent* that Mary bought __ today.
     c. * It is *these snacks* that Mary bought cheese today.
     d. It is *apparent* that Mary bought cheese today.

Filler-gap dependencies occur in many constructions, including Wh-movement (2), topicalization (3), and tough-movement (4), which differ in surface form but share the filler-gap dependency and its properties.

(2) I know *what* Mary bought __ today.

(3) *These snacks*, Mary bought __ today.

(4) *These snacks* are tough to buy __ here.

While it might initially appear that learning could occur from simply expecting a gap when presented with a filler, properties of this dependency also include specific constraints on when they can be formed. Some structural configurations, called *islands*, block the formation of a filler-gap dependency. For example, a filler-gap dependency cannot be formed inside a relative clause (e.g., *that carried* __) despite the fact that *carried* lacks an object (i.e., is followed by a gap). The relative clause blocks the dependency, and so a gap is unacceptable regardless of the presence of a filler. Examples (5)-(8) show that all filler-gap dependencies are subject to this same restriction.

(5) * It is *these snacks* that Mary bought [the bag that carried __] today.

(6) * I know what Mary bought [the bag that carried __] today.

(7) * *These snacks,* Mary bought [the bag that carried __] today.

(8) * *These snacks* are tough to buy [the bag that carried __] here.

One task for a learner is to recognize, on the basis of grammatical examples only (e.g., (1a) and (1d), but not (1b), (1c), or (5)), when each filler-gap dependency can and cannot occur. A further task is to recognize that the same properties apply to each construction containing a filler-gap dependency (2-4), and thus posit a shared representation underlying all filler-gap dependencies.

An alternative method to generalizing would be a piecemeal learning process: learning each construction separately. For the piecemeal process to work, sufficient examples of each construction must occur in the input, and similar constraints across these constructions would arise from distinct observations. NLMs here provide an opportunity to test whether a shared representation can *in principle* be extracted from the input without linguistic biases.

2.1 NLMs and Filler-gap dependencies

NLMs can learn at least some syntactic representations involving locality (see Linzen and Baroni (2021) for a review). NLMs have been shown to represent shared syntactic structure across different constructions in simulated priming (Prasad

et al., 2019) and simulated satiation (Lu et al., 2024) experiments, which compare measures from NLMs before and after exposing them to sentences with similar syntactic structures. Similarly, NLMs have been shown to generalize over syntactic structures that have been excluded from their training data (Jumelet et al., 2021; Warstadt, 2022; Misra and Mahowald, 2024; Patil et al., 2024). How human-like these generalizations are is still an open question.

With respect to filler-gap dependencies, NLMs capture language-specific island constraints in English (Wilcox et al., 2018; Ozaki et al., 2022; Wilcox et al., 2023) and Norwegian (Kobzeva et al., 2023) in sentences with embedded Wh-movement. Ozaki et al. (2022) analyze other constructions with filler-gap dependencies (clefting, topicalization, and tough-movement) and find that model performance varies by construction and is associated with the relative frequency of the constructions in texts resembling the training corpus. In other words, Ozaki et al. (2022) argue the model's ability to approximate human behavior is dependent on the availability of each construction type in the input. Whether this ability is modulated by a shared representation *across* different constructions is not known.

Finally, Lan et al. (2024) investigate the extent to which NLM performance with double gap phenomena (parasitic gaps and across the board movement) is in line with human judgments. In these constructions, a gap *can* occur inside an island, only if another gap is present. While they find that pretrained NLM performance is low for constructions with parasitic gaps or across the board movement, the authors show that adding examples of parasitic gaps and across the board movement to an NLM's training data adjusts its performance to be in line with human expectations, showing directly the relationship between NLM performance and surface forms in the training data. Thus, if the training data of an NLM does not contain sufficient instances of a particular construction, its ability to correctly capture the pattern of grammaticality suffers, strengthening Ozaki et al. (2022)'s claim that input frequency matters.

The methodology introduced by Lan et al. (2024) provides a path for exploring whether NLMs make generalizations that are not apparent from simply testing a pretrained model: if the model can improve on one construction from direct training on that construction, we can ask what other effects such training might have. Does training a model on one construction containing a filler-gap dependency affect its performance on *other* constructions containing filler-gap dependencies, the way one might expect given a shared representation?

3 Methods

3.1 Measuring Filler-gap dependencies and Island Effects

Psycholinguistic findings show structural constraints affect human expectations for gaps inside islands (Phillips, 2006; Traxler and Pickering, 1996; Stowe, 1986). One way to evaluate whether an NLM's predictions align with these effects is to measure its *surprisal*, the negative log probability of a word given context; less surprising words have higher probabilities. Surprisal quantifies the effect of processing difficulty (Levy, 2008). Investigating NLM surprisal at particular points in a sentence effectively treats the models like psycholinguistic subjects (Futrell et al., 2019).[1]

To determine whether the NLMs capture syntactically relevant knowledge, we evaluate surprisal at critical regions of grammatical and ungrammatical variants of superficially similar sentences, as in (1). We compute surprisal of the region following a verb, which can either consist of a direct object (a filled gap, **-gap**) or an adverb (a gap, i.e., no direct object, **+gap**). Each string also either contains a filler (**+filler**) or does not (**-filler**). This 2x2 design is illustrated in Table 1, with the critical region marked in bold. For example, the surprisal at *today* in (1a) should be lower than the surprisal at *cheese* in (1c) because in the latter case, given the filler *these snacks*, the reader expects a gap in the object position of *bought*. If the critical region consists of multiple words, we sum their surprisals.

If the NLM has learned the dependency, we expect to see high surprisal in the critical regions of ungrammatical sentences: both when it encounters a gap without having seen a prior filler (1b, +gap/-filler), as well as if it has seen a filler but then encounters a filled gap (1c, -gap/+filler). Likewise, we expect low surprisal in the critical regions of

---

[1]However, see Van Schijndel and Linzen (2021) and Huang et al. (2024) for arguments that surprisal is not always a good estimate of human behavior for some types of syntactically complex sentences.

	+filler	-filler	expected effect
+gap	It is these snacks that Mary bought _ **last week**.	*It is apparent that Mary bought _ **last week**.	negative
-gap	*It is these snacks that Mary bought **the cheese** last week.	It is apparent that Mary bought **the cheese** last week.	positive

Table 1: The expected effect is the difference in the LM's surprisal for versions of the same simple (non-island) construction with and without a filler.

	+filler	-filler	expected effect
+gap	*It is these snacks that Mary bought the bag that held _ **last week**.	*It is apparent that Mary bought the bag that held _ **last week**.	Closer to zero than simple effect
-gap	*It is these snacks that Mary bought the bag that held **the cheese** last week.	It is apparent that Mary bought the bag that held **the cheese** last week.	Closer to zero than simple effect

Table 2: For sentences containing islands, the expected effect is a reduction of the filler effect compared to the effect in simple sentences.

grammatical sentences: if it encounters a gap after having seen a filler (1a, +gap/+filler), as well as if it sees neither filler nor gap (1d, -gap/-filler).

To summarize these predictions, we calculate the **filler effect**: the *difference* in surprisal between two sentences that are identical except for the presence of a filler (Wilcox et al., 2018, 2023). We take the surprisal for a +filler sentence and subtract the surprisal of its -filler counterpart. Based on the predictions from the previous paragraph, our filler effect predictions for **simple** (non-island) sentences are as follows: a negative filler effect in the +gap condition and a positive filler effect in the -gap condition. These predictions are in Table 1.

The filler effect prediction for sentences with **islands** differs from the prediction for simple sentences. Filler-gap dependencies are not licensed into islands; sentences with islands are ungrammatical if they possess either a filler, a gap, or both. Only the sentences with no filler and no gap should be grammatical. Following Wilcox et al. (2023), we predict an NLM with human-like performance on island effects should show filler effects around zero in sentences with islands. If the NLM has learned that filler-gap dependencies are always unlicensed inside an island, the presence or absence of a filler should not affect the NLM's surprisal at a gap inside an island. Therefore, there should be no difference between the surprisal in the +filler and -filler conditions, i.e., a filler effect of zero. These predictions are summarized in Table 2. It is worth noting that Ozaki et al. (2022) have a different prediction for islands: they assume that grammaticality affects surprisal and that the NLM's surprisal

will be different at the filled gap in the grammatical -gap, -filler condition. We discuss islandhood and surprisal further in Section 5.

3.2 Language Model

We estimate surprisal from a recurrent neural network (RNN) from Gulordava et al. (2018), which is a Long-Short-Term-Memory (LSTM) RNN (Hochreiter and Schmidhuber, 1997) with two hidden layers with 650 units in each layer, trained on data from an English Wikipedia corpus (90 million tokens, or around 3 million sentences). We chose to use this model because prior research evaluating it on filler-gap dependencies has shown success in capturing human-like knowledge of filler-gap dependencies, even relative to larger models (Wilcox et al., 2018; Ozaki et al., 2022; Lan et al., 2024; Wilcox et al., 2023; Kobzeva et al., 2023). Because it has transparent training data, we could carefully compare the pretrained RNN with models augmented with instances of different constructions, which we call Cleft-RNN and Topic-RNN. Details of the augmented training data for these models are explained in Section 4. [2]

---

[2]Transformers and LSTMs perform similarly on syntactic generalization tasks when trained on the same amounts of data, despite the transformers' lower perplexity (Patil et al., 2024). We did, however, replicate the results of our baseline for each construction with a pretrained GPT-2 model. See Appendix C1 for these results and more discussion on modeling choices.

3.3 Statistical Analysis

We test for two effects: the first is whether the models recognize that a filler must be associated with a gap in simple sentences, and the second is whether this expectation is modulated by the presence of an island. To determine whether the RNN learned the filler-gap dependency in simple sentences, we fit a linear mixed-effects regression model following Wilcox et al. (2023) using surprisal as the dependent variable, sum-coded features for the presence or absence of fillers and gaps which were fixed effects. If the RNNs learn the filler-gap dependency for a particular construction, we expect to see a negative interaction term between the presence of fillers and gaps, in line with Wilcox et al. (2023).

Additionally, Wilcox et al. (2023) fit mixed-effects models including islandhood as a fixed effect, claiming that a positive three-way interaction between the presence of fillers, gaps, and islands reflects the successful learning of island constraints. We apply this analysis, but also consider both directions of the dependency separately: unlicensed gap effects (UGE) in sentences containing a gap, and filled gap effects (FGE) in sentences without a gap (Kobzeva et al., 2023). We fit separate linear mixed-effects models for the surprisals of sentences with and without gaps, with fixed effects for fillers and islands.[3] This analysis allows us to tease apart the two-way nature of the dependency and analyze whether the RNNs' failures or successes are driven by only one direction of the dependency. Success requires the regression models' coefficients to all be negative for UGEs and all be positive for FGEs for both main effects and interactions. These analyses were repeated separately for the pretrained and augmented RNNs. All regression models were sum-coded, included random effects for each item (Barr et al., 2013), and fit using the Pymer library in Python (Jolly, 2018). All formulas and results are reported in Appendix B.

4 Experiments

We evaluate an RNN's behavior on four filler-gap dependency constructions in both simple sentences and sentences with an island. We augment the NLM's training data with instances of a single construction and then observe the effects of that augmentation on its performance on tests of *other* constructions. In other words, we ask whether "teaching" an NLM filler-gap dependencies in one construction helps it "learn" the dependency in other constructions. If and only if the NLM acts consistently with linguists' conclusion that these constructions share an underlying representation, an improvement in performance on one construction should generalize to others. Our implementation [4] and models [5] are both publicly available.

4.1 Materials

For embedded Wh-movement, we use the materials from Wilcox et al. (2023)'s experiment on complex NP islands. For each of the other 4 constructions in (1)-(4), we test a "simple" version (no island) - where a dependency can be formed - and an island version - which does not allow for the formation of a filler-gap dependency. For topicalization, we test two versions: one ("topicalization without intro"), to closely match the materials used by Ozaki et al. (2022), in which the topicalized element has no analog in the no filler sentences, and one ("topicalization with intro") in which the filler is replaced with an introductory string to control for the length of the sentence and the presence of a comma. See Tables 1 and 2 for a schema of the design using one construction, clefting. For each construction, we generate a set of items with a fixed syntactic template and a vocabulary that varies across the set. We ensure that the lexical items are all in the RNN's vocabulary. For each item, we modulate the presence of a filler, a gap, and an island structure, generating 8 sentence types per item. Our final testing set contains 486 clefting items, 486 topicalization with intro, 161 topicalization without intro, and 243 tough-movement items.

4.2 Predictions

Each graph shows the average filler effect with 95% confidence intervals for simple and island sentences for each construction, before and after augmentation. For **simple** constructions, we expect to see a negative filler effect when the gap is present (blue bars) and a positive effect when

---

[3]Since Kobzeva et al. (2023) were only testing for island effects, their regression models were fit on filler effects rather than raw surprisals based on the presence of an island. We consider the joint presence of filler and island in our analysis.

the gap is absent (orange bars). This is because a human-like learner should find a gap less surprising when a filler is present than when there is no filler, and vice versa when there is no gap. (See Section 3.1 for more detail.) For **island** constructions, we expect that the confidence interval for the filler effect should overlap with zero because a human-like knowledge of islands suggests that a gap should be equally surprising regardless of the upstream presence of a filler; the same is true in the no-gap condition, using one of Wilcox et al. (2023)'s criteria. A less stringent relative metric for learning islands is a reduced effect relative to the filler effect in simple sentences (Wilcox et al., 2023); in other words, the difference in surprisal at the critical region decreases rather than disappears entirely.[6]

4.3 Experiment 1: Training on clefting

Our initial test of the pretrained RNN yielded variation across constructions consistent with Ozaki et al. (2022). Based on these results, we chose to augment the pretrained RNN's training data with instances of clefting because it fails to demonstrate knowledge of islands in clefting; also, clefting is reported as less frequent compared to Wh-movement in Ozaki et al. (2022). We hypothesize that Wilcox et al. (2023)'s robust effects with embedded Wh-movement may be due to the relative frequency of the construction in the training corpus.

We create a training set for clefting, using the same syntactic template for test sentences but with different lexical items. We then retrain the RNN following the same configurations in Gulordava et al. (2018), with training data that include all original training material and 864 additional examples of grammatical simple clefting. Half the examples contain a gap, as in (1a), half do not, as in (1d). We refer to this model as Cleft-RNN.

4.3.1 Results

Simple constructions. We first present the filler effects for the simple sentences of each construction of the pretrained RNN (before augmentation) and Cleft-RNN, plotted in Figure 1. Testing the pretrained RNN on simple constructions, we replicate Wilcox et al. (2023)'s findings for embedded Wh-movement. The pretrained RNN also shows the desired filler effect pattern in simple sentences

for clefting and tough-movement, but not for either form of topicalization. For each construction type, we looked at the two-way interaction of filler and gap, confirming a positive result for Wh-movement, clefting, and tough-movement (negative interaction terms with $p < 0.001$). The interaction effects for topicalization were positive, indicating that the pretrained RNN did not learn the dependency in either type of topicalization constructions.

Training on clefting had no significant effect on knowledge of the dependency in simple sentences of any construction, confirmed both by the qualitative appearance of the graphs and by the mixed-effects models for each construction type (negative interaction terms with $p < 0.001$). [7] Since Cleft-RNN did not learn the dependency in either form of topicalization, we do not report island effects.

Island constructions. The results for constructions containing islands before and after augmentation, presented in Figure 2, are less straightforward. For no construction did the pretrained RNN meet the most stringent criteria for recognizing island constraints: that is, both filler effects in the island condition equaling zero. We do see varying degrees of *reduction* in the filler effects for each construction in the island vis-a-vis the simple condition. We consider each direction of the dependency separately: filled gap effects (FGE, orange) and unlicensed gap effects (UGE, blue).

The mixed-effects model using Wilcox et al. (2023)'s methods shows negative filler-gap interaction terms and positive three-way interaction terms for Wh-movement, clefting, and tough-movement, suggesting that the pretrained RNN correctly captures island constraints for all constructions where it knows the simple dependency. However, this result is at odds with our qualitative findings, which suggest the model did not learn the relevant generalization in clefting. We observe that in the -gap condition (orange bars), the filler effect is equal in magnitude in both the simple and island conditions. In fact, the filler effect is negative, suggesting high surprisal in the grammatical (-filler, -gap) sentence (see Table 2). This qualitative result is confirmed by our separate mixed-effects model for FGEs, which shows sta-
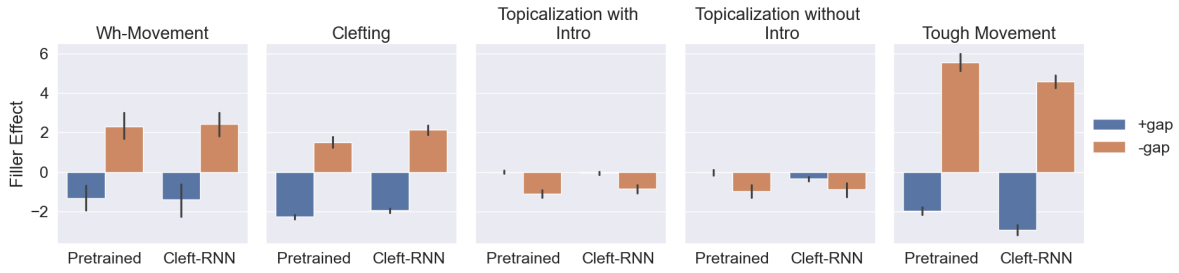

Figure 1: Filler effects for simple constructions for the pretrained model and Cleft-RNN.
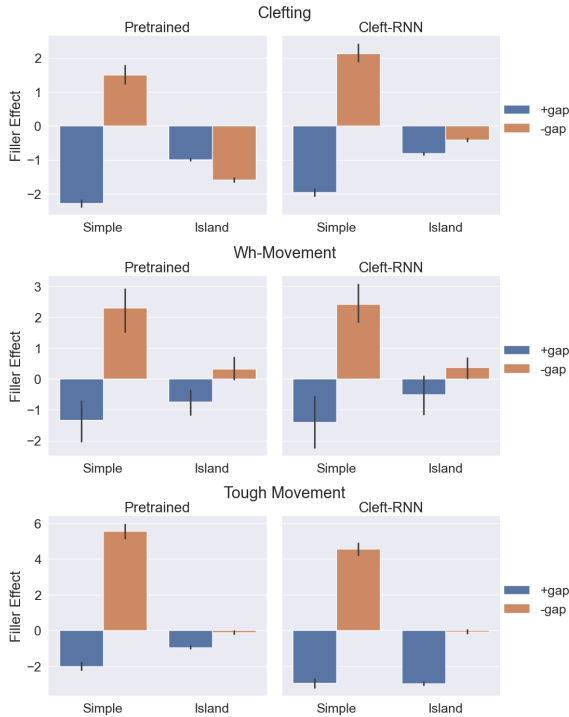

Figure 2: Filler effects for simple and island constructions for the pretrained model and Cleft-RNN. Since this dependency was not learned for topicalization, we do not display these results.

tistically significant positive coefficients for Wh-movement and tough-movement ($p < 0.001$), but not for clefting. In the regression models for UGE, we observe statistically significant negative coefficients for all three constructions ($p < 0.001$). In other words, the pretrained RNN is not sensitive to FGEs for clefting constructions, consistent with the qualitative pattern.

We now review Cleft-RNN's behavior with island constructions. The direction and magnitude of the effects in the three-way interaction model are consistent with the conclusion that Cleft-RNN captures island constraints in clefting and Wh-movement. For the clefting construction (in other words, when presented with examples structurally identical to those it was trained on), our regres-

sion models for FGEs and UGEs support this result. All coefficients for the UGEs are negative, and all coefficients for the FGEs are positive. Cleft-RNN is less sensitive to the presence of an upstream filler at a filled gap in an island construction than the pretrained RNN, though qualitatively the grammatical form is still more surprising than the ungrammatical form. For Wh-movement, Cleft-RNN's confidence interval of the filler effect for islands in the +gap condition overlaps with zero, achieving our most stringent criterion for displaying knowledge of island constraints. Results were statistically significant, both for clefting ($p < 0.001$) and for Wh-movement ($p < 0.01$), which was tested on a smaller stimulus set.

In tough-movement, however, augmentation has a detrimental effect. The magnitude of the filler effect in the +gap condition for islands is equivalent to that in the simple cases. The positive, non-significant interaction term for the regression model for UGEs in tough-movement supports this observation; Cleft-RNN lacks sensitivity to islands in tough-movement constructions.

For topicalization, our regression models do not have the correct signs for either construction type, confirming, as a qualitative inspection of Figure 1 suggests, that Cleft-RNN does not learn the dependency in these constructions.

4.4 Experiment 2: Training on topicalization

Since neither the pretrained RNN nor Cleft-RNN are able to arrive at the correct generalization for cases of topicalization, we now determine if providing the pretrained RNN with positive direct evidence of topicalization is sufficient for learning this dependency. We follow a similar procedure to the previous experiment, generating sentences from the same syntactic template for topicalization with intro and ensuring the lexical items do not appear in the testing sentences. We augment the RNN's Wikipedia corpus with 864 grammati-
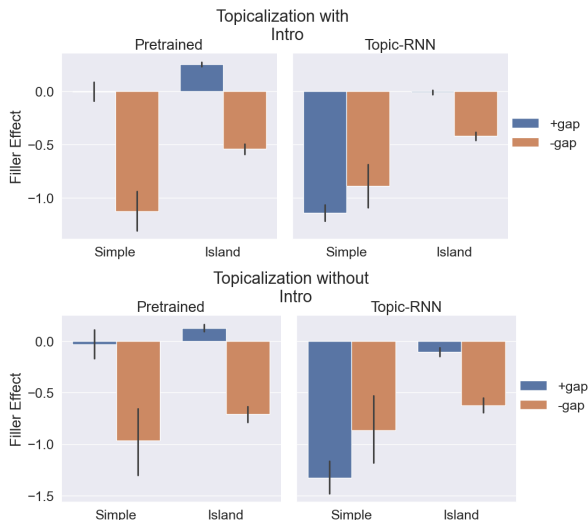
Figure 3: Filler effects for simple and island constructions for the pretrained model and Topic-RNN.

cal examples of topicalization, half with a gap and half without a gap, and train it using the hyperparameters in Gulordava et al. (2018). The augmented model is referred to here as Topic-RNN.

4.4.1 Results

Figure 3 shows filler effects in the pretrained RNN and Topic-RNN. Training explicitly on simple instances of topicalization does not lead the model to posit the dependency in both directions. Here, positive evidence is not enough to learn even the simple dependency. The regression model does not show significant effects for basic filler-gap licensing in Topic-RNN.

Topic-RNN *does* learn that the presence or absence of an upstream filler should modulate surprisal at a gap (UGEs, blue bars), but it fails to learn the correct relationship between a filled gap and the absence of an upstream filler (FGEs, orange bars). In fact, Topic-RNN's surprisal is consistent with the non-human-like hypothesis that *the cheese* in the sentence *The snacks, Mary bought the cheese last week* is less surprising than in the sentence *In fact, Mary bought the cheese last week*.

5 Discussion

In this paper, we test whether NLMs can generalize knowledge of filler-gap dependencies across different constructions when their input is augmented with only one construction containing a filler-gap dependency. The pattern of knowledge of the pretrained RNN potentially reflects piece-

meal learning, where the frequency of particular constructions modulates the model's recognition of the filler-gap dependency for each construction type individually (Ozaki et al., 2022). However, based on the pretrained results alone we cannot determine whether the NLM's inferences for one construction are based on others, hence the need for an augmentation-based procedure. Experiment 1 found that while Cleft-RNN behaves differently than the pretrained RNN on clefting, Wh-movement, and tough-movement, it fails to generalize systematically across all the types of filler-gap dependencies we test. Cleft-RNN's failure to learn the relevant dependency for simple topicalization sentences further confirms that these models do not arrive at their knowledge of filler-gap dependencies through a shared representation.

Cleft-RNN does improve its representation of island constraints in clefting, the construction it was augmented with. However, this improvement still preserves the incorrect prediction for grammaticality. In this case, positive evidence of grammatical forms is still insufficient for human-like learning. This finding supports a conclusion drawn by Lan et al. (2024): that given sufficient evidence of a construction, NLMs can arrive at a correct representation of the constraints. However, the ability to generalize from one construction appears weak at best.

Cleft-RNN's filler effect for Wh-movement is the only instance among our findings where an NLM achieves the most stringent measure of islands: a confidence interval overlapping with zero. However, we are cautious to over-interpret this finding: our test set for Wh-movement was far smaller than that of the other construction types.

Further, the failure of Cleft-RNN to capture islands in tough-movement relative to the pretrained RNN highlights a more pressing issue with NLMs that rely only on surface distributions: exposure to one construction type can cause a degradation in an NLM's knowledge of a different type, when the two share superficially similar characteristics at odds with the dependency. The learner would then need even more positive direct evidence of the other type to offset such erroneous conclusions.

We found that the NLM we tested fares worse at recognizing island constraints than past studies would suggest: both the pretrained RNN and our augmented models failed to arrive at the most stringent measure of islandhood in all cases but

Wh-movement in Cleft-RNN. We suspect this is greatly influenced by the frequency of surface forms of particular constructions, as hypothesized by Ozaki et al. (2022).

Why, then, did Wilcox et al. (2023)'s method of using the presence of fillers, gaps, *and* islands as predictors of surprisal yield a significant interaction for islandhood in instances where the filler effect suggested otherwise? We believe that the interaction collapses effects across different combinations of features. Our NLMs succeed with UGEs, which likely obscures their corresponding failure to recognize FGEs. However, the failure in FGEs suggests that the NLM is recognizing neither grammaticality nor the presence of an island. For FGEs, the filler effect is in the wrong direction; grammatical continuations (i.e., those without a filler or a gap) are more surprising than ungrammatical ones (a filler and no gap). This contradicts both the measures for islands proposed by Wilcox et al. (2023) and Ozaki et al. (2022), who suggest that rather than no difference at the gap site, surprisal should align with grammaticality. Here we find that surprisal does not align with either measure and is in fact showing the reverse pattern for grammaticality.

The results of Experiment 1 strongly suggest that the NLM arrives at its knowledge of filler-gap dependencies through piecemeal learning and that positive direct evidence of a filler-gap dependency in each construction is required to learn the dependency for that construction. We conducted Experiment 2 to explore whether positive evidence of simple topicalization sentences is sufficient for Topic-RNN to make predictions consistent with a human-like understanding of both the simple dependency and islands. Topic-RNN learns to expect a gap given a filler, but fails to learn the other direction of the dependency: that in the absence of a filler, there should be no gap. The model's failure to learn the simple topicalization dependency even in the face of direct evidence is an additional challenge to claims that language-specific biases are not necessary to learn such dependencies.

Taken together, the results from Experiments 1 and 2 show that NLMs do not generalize from a shared representation to learn filler-gap dependencies. Instead, they rely heavily on input that closely aligns with individual constructions. Further, in cases such as topicalization, NLMs appear to struggle with learning the dependency. Our findings are particularly important as researchers consider in what ways NLMs might and might not serve as good proxies for language learners. Our work reiterates the importance of specific linguistic inductive biases to model language acquisition.

Acknowledgements

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.

Thomas G. Bever and Brian McElree. 1988. Empty categories access their antecedents during comprehension. *Linguistic Inquiry*, 19(1):35–43. Publisher: The MIT Press.

Noam Chomsky. 1977. On Wh-Movement. *Formal Syntax*, pages 71–132. Publisher: Academic Press.

Stephen Crain and Janet Dean Fodor. 1985. Rules and constraints in sentence processing. *North East Linguistics Society*, 15(1).

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Gerald Gazdar. 1982. Phrase Structure Grammar. In Pauline Jacobson and Geoffrey K. Pullum, editors, *The Nature of Syntactic Representation*, Synthese Language Library, pages 131–186. Springer Netherlands, Dordrecht.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard Univ. Press, Cambridge, Mass.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically.

In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

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

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Eshin Jolly. 2018. Pymer4: Connecting r and python for linear mixed modeling. *Journal of Open Source Software*, 3(31):862.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969. Association for Computational Linguistics.

Ronald Kaplan and Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2023. Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 175–185, Amherst, MA. Association for Computational Linguistics.

Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–28.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Jiayi Lu, Jonathan Merchan, Lian Wang, and Judith Degen. 2024. Can syntactic log-odds ratio predict acceptability and satiation? In *Proceedings of the Society for Computation in Linguistics 2024*, pages 10–19, Irvine, CA. Association for Computational Linguistics.

Brian McElree and Teresa Griffith. 1998. Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(2):432–460.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *arXiv preprint arXiv:2403.19827*.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Grokking of hierarchical structure in vanilla transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.

Akira Omaki, Ellen F. Lau, Imogen Davidson White, Myles L. Dakan, Aaron Apple, and Colin Phillips. 2015. Hyper-active gap filling. *Frontiers in Psychology*, 6.

Satoru Ozaki, Dan Yurovsky, and Lori Levin. 2022. How Well Do LSTM Language Models Learn Filler-gap Dependencies? *Proceedings of the Society for Computation in Linguistics*, 5(1):76–88.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (fict) shows that language models can generalize from indirect evidence. *arXiv preprint arXiv:2405.15750*.

Colin Phillips. 2006. The Real-Time Status of Island Phenomena. *Language*, 82(4):795–823.

Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *LingBuzz Preprint*.

Carl Jesse Pollard and Ivan A. Sag. 1987. *Information-Based Syntax and Semantics: Volume 1, Fundamentals*. Number no. 13 in CSLI lecture notes. Cambridge University Press, Stanford, CA.

Paul M. Postal. 1999. *Three Investigations of Extraction*. The MIT Press.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Carson T. Schütze, Jon Sprouse, and Ivano Caponigro. 2015. Challenges for a theory of islands: A broader perspective on ambridge, pine, and lieven. *Language*, 91(2):31–39.

Jon Sprouse, Ivano Caponigro, Ciro Greco, and Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in english and italian. *Natural Language & Linguistic Theory*, 34(1):307–344.

Laurie A. Stowe. 1986. Parsing wh-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, 1:227–245.

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

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5831–5837.

Alex Warstadt. 2022. *Artificial Neural Networks as Models of Human Language Acquisition*. New York University.

Ethan Wilcox, Richard Futrell, and Roger Levy. 2023. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

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: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.