What Syntactic Structures block Dependencies in RNN Language Models?

Ethan Wilcox[1], **Roger Levy**[2], and **Richard Futrell**[3]

[1]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[2]Department of Brain and Cognitive Sciences, MIT, rplevy@mit.edu
[3]Department of Language Science, UC Irvine, rfutrell@uci.edu

Abstract

Recurrent Neural Networks (RNNs) trained on a language modeling task have been shown to acquire a number of non-local grammatical dependencies with some success (Linzen, Dupoux, & Goldberg, 2016). Here, we provide new evidence that RNN language models are sensitive to hierarchical syntactic structure by investigating the **filler–gap dependency** and constraints on it, known as **syntactic islands**. Previous work is inconclusive about whether RNNs learn to attenuate their expectations for gaps in island constructions in particular or in *any* sufficiently complex syntactic environment. This paper gives new evidence for the former by providing control studies that have been lacking so far. We demonstrate that two state-of-the-art RNN models are are able to maintain the filler–gap dependency through unbounded sentential embeddings and are also sensitive to the hierarchical relationship between the filler and the gap. Next, we demonstrate that the models are able to maintain **possessive pronoun gender expectations** through island constructions—this control case rules out the possibility that island constructions block all information flow in these networks. We also evaluate three untested islands constraints: coordination islands, left branch islands, and sentential subject islands. Models are able to learn left branch islands and learn coordination islands gradiently, but fail to learn sentential subject islands. Through these controls and new tests, we provide evidence that model behavior is due to finer-grained expectations than gross syntactic complexity, but also that the models are conspicuously un-humanlike in some of their performance characteristics.

Keywords: Syntactic Islands, Recurrent Neural Networks, Blocking Effects, Acquisition of Syntax

Introduction

Recurrent Neural Networks (RNNs) with Long Short-Term Memory architecture (LSTMs) have achieved state-of-the-art scores at a number of natural language processing tasks, including language modeling and parsing (Hochreiter & Schmidhuber, 1997; Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016). In addition, they have begun to be used as a plausible sub-symbolic model for a variety of cognitive functions, including visual perception and language processing and comprehension (J. Elman, 1990). However, the distributed representations learned by RNNs and neural networks in general are notoriously opaque, posing a challenge for their interpretability as models of human sentence processing and for their controllability as NLP systems.

One recent line of work aims to uncover what these 'black boxes' learn about language by treating them like human psycholinguistic subjects. In this **psycholinguistic paradigm** RNNs trained on the language modeling task are fed handcrafted sentences, designed to expose their underlying syntactic knowledge (Linzen et al., 2016; McCoy, Frank, & Linzen, 2018). Much of this work has investigated what RNNs trained on a language modeling objective are capable of learning about natural syntactic dependencies. For the purposes of this investigation, we define **dependency** as any systematic co-variation between two words. For example, in one experiment networks were tested as to whether they had learned the number agreement dependency between a subject and a verb. They were fed with the prefix *The key to the cabinet...* and correctly gave a higher probability to the grammatical *is* over the ungrammatical *are*. Networks were shown to successfully complete this task for a number of languages, as well as for sentences whose content words were replaced with random alternatives of the same syntactic category rendering them syntactically licit but semantically implausible (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018).

But learning that covariance exists between certain words or word forms, without reference to their relative positions, is not enough to say that the RNN models have fully learned a dependency. Natural language dependencies consist of co-variation between two elements in *certain syntactic positions*. Agents must both attend to the structural relationship between the two elements bound by the dependency and filter out intervening material in syntactically irrelevant positions. The subject–verb number agreement task above provides compelling evidence that RNNs are capable of the latter: they were able to maintain correct predictions despite a number of *distractors* that mismatched the subject in number, such as *cabinet* in the example provided (Marvin & Linzen, 2018).

Evidence suggesting that RNN language models are also sensitive to the structural relationship between the two bound elements has emerged from the study of **filler–gap dependencies** (Wilcox, Levy, Morita, & Futrell, 2018; Chowdhury & Zamparelli, 2018). The filler–gap dependency is the dependency between a *filler*—such as *who* or *what*—and and a gap, which is an empty syntactic position. Crucially, filler–gap dependencies are subject to a number of constraints, known as *island constraints*, which are a set of structural positions that prevent the filler and the gap from entering into a dependency with each other (Ross, 1967). (1-b) gives one example island, in which the dependency is blocked by a wh-complementizer.

(1) a. I know what the guide said that the lion devoured __ yesterday. NO VIOLATION
b. *I know what the guide said whether the lion devoured __ yesterday. WH-ISLAND ISLAND VIOLATION

While it has been shown that both simple Elman RNNs and more contemporary LSTMs are able to represent the basic covariance between fillers and gaps, as well as other non-structural aspects of dependency, it is still uncertain whether

the models are sensitive to island constraints (J. L. Elman, 1991). Previous work has demonstrated that two state-of-the-art models are sensitive to three of the most-studied island constraints (wh-islands, complex NP islands and adjunct islands) but insensitive to a fourth (subject islands) (Wilcox et al., 2018). Others have concluded that the models are merely sensitive to syntactic complexity plus order. Chowdhury and Zamparelli (2018) compared sentence-level perplexity scores obtained by RNN LMs for wh-questions that violate island constraints, and yes-no questions and statements that violate no grammatical rules but contain the same syntactic structures. While the models obtained better perplexity scores on the statements compared to the island-violation questions, they performed similarly on the island-violations and non-violating yes/no questions. These results may indicate that RNNs are not learning to attenuate their expectations for gaps in island constructions in particular, but in *any* sufficiently complex syntactic environment.

This paper adjudicates between these two accounts of model behavior by providing control studies that have been lacking so far. In the first section, we demonstrate that two state-of-the-art LSTM models are sensitive to some forms of syntactic complexity, but not to others. Models are able to maintain the filler–gap dependency through **unbounded sentential embeddings** and yet are sensitive to the **hierarchical relationship** between the filler and the gap, suggesting that only specific types of syntactic complexity block gap expectations. In the second section, we turn to **possessive pronoun gender dependencies**, demonstrating that the models are able to maintain general expectations through island constructions—it is not the case that island constructions block all information flow in these networks. In this section we also evaluate three untested islands constraints: **coordination islands**, **left branch islands**, and **sentential subject islands**. Models are able to learn left branch islands and coordination islands gradiently, but fail to learn sentential subject islands. Through these controls and new tests, we provide evidence that model behavior is due to finer-grained expectations than gross syntactic complexity, but also that the models are conspicuously un-humanlike in some of their performance characteristics.

Methods

Language Models

We assess two state-of-the-art pre-existing LSTM models trained on English text for a language modeling objective. The first model, which we refer to as the **Google Model**, was trained on the One Billion Word Benchmark and has two hidden layers with 8196 units each. It uses the output of a character-level convolutional neural network (CNN) as input to the LSTM (and was originally presented as the *BIG LSTM+CNN Inputs*) (Jozefowicz et al., 2016). The second model, which we refer to as the *Gulordava Model* was selected for its previous success at learning the subject-verb number agreement task. It was trained on 90 Million tokens of English Wikipedia, and has two hidden layers of 650 units each (Gulordava et al., 2018).

Dependent Measure: Surprisal

In this work we take a grammatical dependency to be the covariance between an upstream *licensor* and a downstream *licensee*. We assess the model's knowledge of the dependency by measuring the effect that the licensor has on the **surprisal** of the licensee, or on material immediately following the licensee when it is a gap. Surprisal, or negative log-conditional probability , $S(x_i)$ of a sentence's i^{th} word x_i, tells us how strongly x_i is expected under the language model's probability distribution. For sentences out of context, the surprisal is: $S(x_i) = -\log p(x_i|x_1 \dots x_{i-1})$. Surprisal is known to correlate directly with processing difficulty in humans (Smith & Levy, 2013; Hale, 2001; Levy, 2008). In this work, we expect that grammatical licensors set up expectations for licensee, reducing its surprisal compared to minimal pairs in which the licensor is absent. We derive the word surprisal from the LSTM langauge model by directly computing the negative log of the predicted conditional probability $p(x_i|x_1 \dots x_{i-1})$ from the softmax layer.

Experimental Design: Wh-Licensing Interaction

The filler–gap dependency is biconditional: Fillers set up expectations for gaps and gaps require fillers to be licensed. To measure this bi-directionality we employ the 2x2 interaction design proposed in Wilcox et al.. There, the authors measure the **wh-licensing interaction**, which they compute from four sentence variants, given in (2), that contain the four possible combinations of fillers and gaps for a specific syntactic position. Note that the underscores are for presentational purposes only, and were not included in test items. Subsequent examples will be given via the (2-d) example, but all four variants were created in order to compute the licensing interaction.

(2) a. I know that you insulted your aunt yesterday. [-FILLER -GAP]
b. *I know who you insulted your aunt yesterday. [+FILLER -GAP]
c. *I know that you insulted __ yesterday. [-FILLER +GAP]
d. I know who you insulted __ yesterday. [+FILLER +GAP]

If the filler sets up an expectation for a gap, then the filled syntactic position where a gap would typically occur should be more surprising in contexts that contain an upstream filler. That is $S(b) - S(a)$ should be a large positive number. If the gap requires a filler to be licensed, then the transition from the embedded verb to the S-modifying PP 'yesterday' that skips over the otherwise-required grammatical object should be more surprising in contexts without an upstream filler. That is, $S(d) - S(c)$ should also be a large negative number. We can assess how well the model has learned both expectations by measuring the difference of differences: $[S(b) - S(a)] - [S(d) - S(c)]$. This is the wh-licensing interaction. If the models are learning the filler–gap dependency, we expect this to be a large positive number, with typical models showing about 4 bits of licensing interaction in simple object extracted clauses such as (2). Although we might expect the

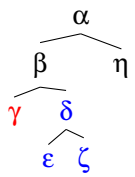

Figure 1: C-Command in a binary-branching tree structure. γ *c*-commands all the nodes in blue, but does not *c*-command the black nodes.

strongest difference in surprisal between (2-a) and (2-b) to be on the filled-gap position, *your aunt*, this material is elided in two of the conditions. Therefore, in order to keep the measurement site the same across all four conditions, we measure wh-licensing interaction in the post-gap prepositional phrase ('yesterday' in (2)).

In previous work using this methodology, RNN knowledge of island constraints was assessed by comparing the licensing interaction in island configurations to that in non-island minimal pairs. Strong evidence for an island constraint would be if the wh-licensing interaction dips to zero for a gap in island position, indicating that the model has decoupled expectations for fillers from gaps in this position. In practice we look for a significant decrease in wh-licensing interaction as indication that the models have learned to attenuate their expectations for gaps within islands. We derive the statistical significance of the interaction from a mixed-effects linear regression model, using some-coded conditions (Baayen, Davidson, & Bates, 2008). We include random intercepts by item but omit random slopes as we do not have repeated observations within items and conditions (Barr, Levy, Scheepers, & Tily, 2013). In our figures, error bars represent 95% confidence intervals of the contrasts between conditions, computed by subtracting out the by-item means before calculating the intervals as advocated in (Masson & Loftus, 2003). [1]

Syntactic Complexity

Unboundedness

The filler–gap dependency can span through a potentially unbounded number of sentential embeddings. To test whether models' expectations were attenuated with greater embedding depth, we created 23 items in five experimental conditions with between 0 and 4 layers of embedding and gaps in either object or indirect object (goal) position, following the examples in (3), and measured the licensing interaction in the post-gap material. (In this and subsequent examples, the material in which the interaction is measured will be highlighted in bold.)

(3) a. I know who you insulted __ **at the party**. [OBJECT GAP, 0 LAYERS]

b. I know who the gardener reported the butler said the hostess believed her aunt suspected you insulted __ **at the party**. [OBJECT GAP, 4 LAYERS]

c. I know who you delivered a challenge to __ **at the party**. [GOAL GAP, 0 LAYERS]

d. I know who the gardener reported the butler said the hostess believed her aunt suspected you delivered a challenge to __ **at the party**. [GOAL GAP, 4 LAYERS]

The results for this experiment can be seen in figure 2, with the object gap results on the top and goal gap results on the bottom. First, we find a significant interaction between fillers and gaps resulting in supperaditive reduction of surprisal ($p < 0.001$ for all conditions) indicating that both models have learned the filler–gap dependency. Starting with the object gap conditions: For the google model, we find no effect of embedding depth on the wh-licensing interaction ($p > 0.85$ in all cases); for the gulordava model, we find a significant decrease in wh-licensing interaction only between the *no embedding* conditions and conditions with 3 or 4 additional layers of embedding ($p < 0.001$ in both). When the gap occurs in the goal position, for the google model, we find no significant effect of embedding depth of the wh-licensing interaction. For the gulordava model, we find a generally smaller wh-licensing interaciton, as well as a significant effect of embedding between the *no embedding* condition and conditions with two or more additional embedding layers ($p < 0.05, p < 0.05, p < 0.01$ for 2 ,3 and 4 layers). We take these results to indicate that the google model has learned the unboundedness of the filler–gap dependency whereas the gulordava model has learned only relative unboundedness and shows behavior that reflects human performance more than human competence. However, these results indicate that both models can, in principle, thread their expectations for gaps through complex syntactic structures, if we take the number of syntactic nodes as a proxy measure for syntactic complexity.

Syntactic Hierarchy

Although the filler–gap dependency is unbounded, it is subject to a number of hierarchical constraints, the most basic of which is that the filler must be "above" the gap, structurally. Here, we take this to mean that the filler must *c*-command the gap, although the precise relationship is more complex (Pollard & Sag, 1994). Structurally-speaking node γ *c*-commands node δ if neither node directly dominates the other and every node X that dominates γ also dominates δ. Figure 1 demonstrates this relationship, with the noes *c*-commanded by γ highlighted in blue.

To assess whether the models had learned this constraint on the structural relationship we created 24 variants following the examples in (4) and measured the wh-licensing interaction in the post-gap PP. If the model has learned the structural constraints on the filler–gap dependency, an undischarged filler in the matrix clause should not make a gap in subsequent parts of the sentence more or less likely, leading to near-zero licensing interaction in the *Matrix Clause* condition.

[1] Our studies were preregistered on `aspredicted.org`: To see the preregistrations go to `aspredicted.org/blind.php?=X` where $X \in \{$`sz8f5d,2r2eu7,zt73qt,es8rx7,f9pk9f,se6i2e`$\}$.

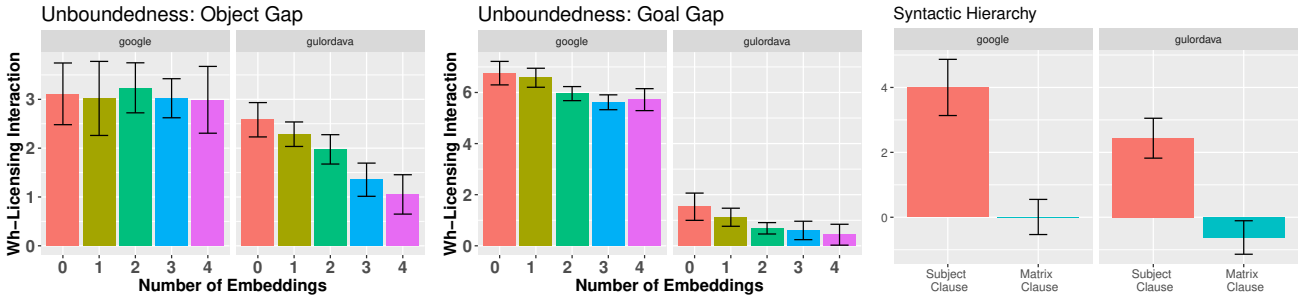

Figure 2: Effect of sentential embedding and syntactic hierarchy on wh-licensing interaction.

(4) a. The fact that the mayor knows who the criminal shot __ **shocked the jury** during the trial. [SUBJECT]
b. *The fact that the mayor knows who the criminal shot the teller shocked __ **during the trial**. [MATRIX]

The results from this experiment can be seen in Figure 2, on the far right panel. We find strong licensing interaction for the grammatical *Subject Clause* conditions (in red), but a striking reduction in licensing interaction for the *Matrix Clause* conditions (in blue), which is significant for both models ($p < 0.001$). As the results in (2) and Wilcox et al. have shown that RNN models are insensitive to linear distance between the filler and the gap, we take these results suggest that it is the relevant structural properties which block the models' expectations for gaps inside the matrix clause.

Island Effects: Gender Expectation vs. Filler–Gap Dependency

Island constraints are specific syntactic configurations that block the filler–gap dependency. One way to show that the RNN models are learning island conditions as constraints on the filler–gap dependency is to demonstrate that they are capable of threading other expectations into island configurations. To do this, we used **pronoun gender expectation** between a gendered noun, such as 'actress' or 'husband', and a possessive pronoun such as 'his' or 'her.'. Nouns that carry overt gender marking or culturally-imbued gender bias set up expectations that subsequent pronominals match them in gender. Previous work has shown that humans thread expectations set up by *cataphoric pronouns* into syntactic islands (Yoshida, Kazanina, Pablos, & Sturt, 2014). Cataphoric pronouns are pronouns that precede the nominal element to which they refer, as in (5).

(5) **Her** manager revealed that the studio notified **Judy Dench** about the new film.

Because cataphoric pronouns are relatively less frequent than anaphoric pronouns, which follow the nominal to which they refer, we use sentences such as those in (6) to assess whether RNN LMs can thread expectations into island environments. We measure the strength of the gender expectation by calculating the difference in surprisal between the matching condition and the mismatching condition, or S((6-b))-S((6-a)). If the models attenuate their expectation for gender agreement in island positions, then we expect an interaction between MISMATCH and ISLAND resulting in supperaditivally lower surprisal.

(6) a. The actress said that they insulted **her** friends. [MATCH, CONTROL]
b. #The actress said that they insulted **his** friends. [MISMATCH, CONTROL]
c. The actress said whether they insulted **her** friends. [MATCH, ISLAND]
d. #The actress said whether they insulted **his** friends. [MISMATCH, ISLAND]

In order to test whether the models maintained their gender expectations through island constructions, we created six suites of experiments following the pattern of (6) for six of the most frequently studied islands constructions. For each of the gender expectation experiments, we created 30 variants, 15 with masculine subjects and 15 with feminine subjects and measured the surprisal at the possessive pronoun. The results are presented on the bottom row in Figure 3 alongside model performance on the filler–gap dependency for the same syntactic constructions (top row). For the filler–gap dependency, results for four islands had already been tested in Wilcox et al. (2018), which we present alongside novel results for *Coordination Islands*, *Sentential Subject Islands* and *Left-Branch Islands*, the latter separately without a gender expectation control. For these experiments, we created between 20-24 experimental items and measured the wh-licensing interaction in the post-gap material. We take a reduction in wh-licensing interaction in island constructions and no such reduction in the gender expectation as evidence that the model has both learned the island constraint, and has applied that constraint uniquely to the filler–gap dependency.

Wh-Islands The wh-constraint states that the filler–gap dependency is blocked by S-nodes introduced by a wh-complimentizer, as demonstrated in the unacceptability of (7-b) compared to (7-a). We created experimental items following the examples in (7) and measured their gender expectation and filler–gap dependency (filler–gap dependency materials were taken from Wilcox et al.).

(7) a. I know who Alex said your friend insulted __ **yesterday**. [CONTROL, FILLER–GAP]
b. *I know who Alex said whether your friend insulted __ **yesterday**. [ISLAND, FILLER–GAP]
c. The actress said they insulted {**his/her**} friends. [CONTROL, GENDER EXP.]
d. The actress said whether they insulted {**his/her**}

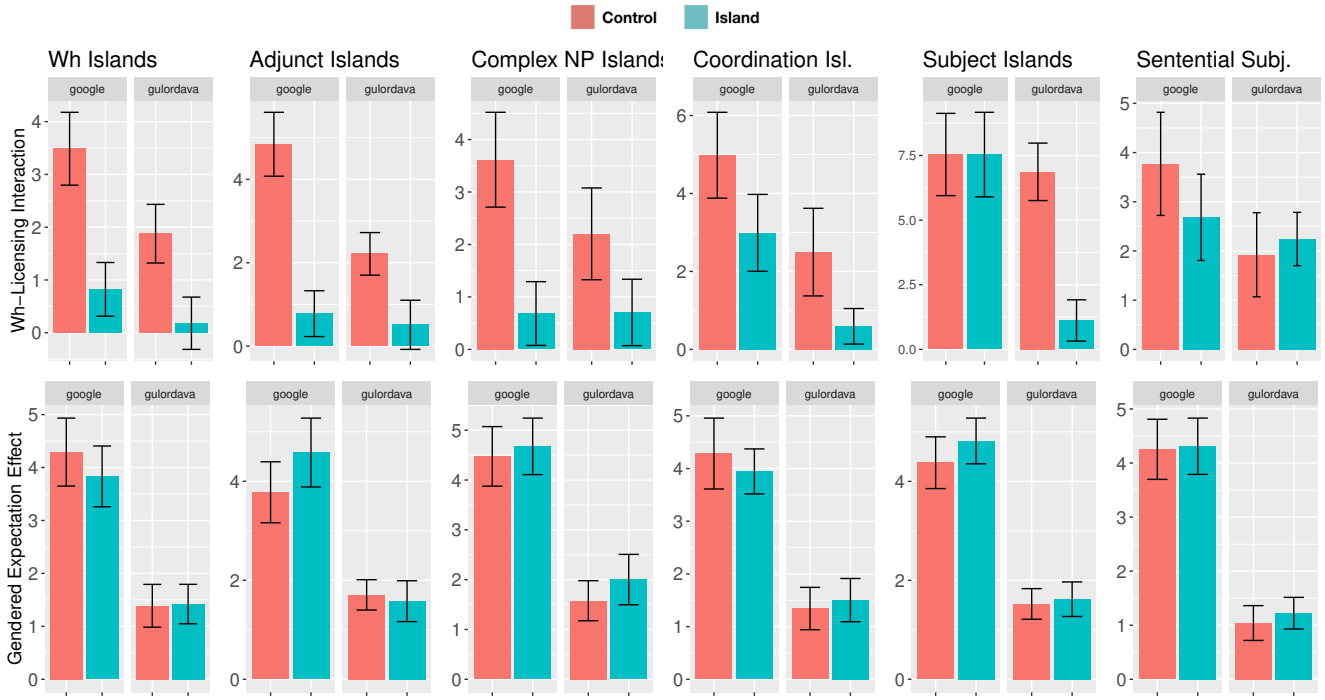

Figure 3: Effect of island construction on gender dependency.

friends. [ISLAND, GENDER EXP.]

The results for this experiment can be seen in the far left panel of Figure 3, with island structures graphed in blue and non-island controls in red. We find a significant difference in licensing interaction between the island and non-island conditions for both the google and gulordava models ($p < 0.001$ for both models), but no such difference in gender expectation.

Adjunct Islands Gaps cannot be licensed inside an adjunct clause, as demonstrated by the relative unacceptability of (8-a) over (8-b).

(8) a. I know what the librarian placed __ **on the wrong shelf**. [CONTROL, FILLER–GAP]
b. *what the patrong got mad after the librarian placed __ **on the wrong shelf**. [ISLAND, FILLER–GAP]
c. The actress thinks they insulted {**his/her**} performance [CONTROL, GENDER EXP.]
d. The actress got mad after they insulted {**his/her**} performance. [ISLAND, GENDER EXP.]

The results for this experiment can be seen in Figure 3, second panel from the left. We find a significant reduction of wh-licensing interaction between the control and island conditions in the case of the filler–gap dependency for both models ($p < 0.001$ google; $p < 0.01$ gulordava; materials taken from]Wilcox et al.). However, we find no effect of syntactic structure on the gender effect.

Complex NP Islands Gaps are not licensed inside S-nodes that are dominated by a lexical head noun, as demonstrated by the relative badness of (9-b) compareid to (9-a).

(9) a. I know what the actress bought __ **yesterday**. [CONTROL, FILLER–GAP]
b. *I know what the actress bought the painting that depicted __ **yesterday**. [ISLAND, FILLER–GAP]
c. The actress said they saw her {**his/her**} performance. [CONTROL, GENDER EXP.]
d. The actress said they saw the exhibit that featured {**his/her**} performance. [ISLAND, GENDER EXP.]

We created items follwing the examples in (9), with filler–gap items adopted from (Wilcox et al., 2018). The results from this experiment can be found in the middle-left panel of Figure 3. We found an effect of syntactic location on wh-licensing interaction for both models ($p < 0.001$ google; $p < 0.01$ gulordava) but no such interaction for gender expectations.

Coordination Islands The coordination constraint states that a gap cannot occur in one half of a coordinate structure as demonstrated by the difference between (10-b) and (10-a), in which a whole conjunct has been gapped.

(10)a. I know what the man bought __ **at the antique shop**. [CONTROL, FILLER–GAP]
b. *I know what the man bought the painting and __ **at the antique shop**. [ISLAND, FILLER–GAP]
c. The fireman knows they talked about {**his/her**} performance. [CONTROL, GENDER EXP.]
d. The fireman knows they talked about the football game and {**his/her**} performance. [ISLAND, GENDER EXP.]

We created experimental items following the examples in (10). Results can be seen in 3 center-right panel. For the filler–gap dependency, in both models there is a significant difference between the *control* condition and *island* conditions ($p < 0.05$ for both models). These results indicate that the models have somewhat attenuated expectations for gaps when they occur in the second half of a coordinate struc-

ture. However, note that, at least for the google model, the wh-licensing interaction is significantly greater than zero, indicating that this model still maintains *some* expectation for gaps in this syntactic location. For both models there is no difference in gender expectation between the *control* and *island* conditions).

Subject Islands Gaps are generally licensed in prepositional phrases, except when they occur attached to sentential subjects. We created experimental items following the examples in (11), with filler–gap materials adapted from Wilcox et al..

(11)a. I know what _ **fetched** a high price. [CONTROL, FILLER-GAP]

b. *I know who the painting that depicted _ **fetched** a high price. [ISLAND, FILLER–GAP]

c. The actress said they sold the painting by {**his/her**} friend. [CONTROL, GENDER EXP.]

d. The actress said the painting by {**his/her**} friend sold for a lot of money. [ISLAND, GENDER EXP.]

The results from this experiment can be seen in Figure 3, second panel from the right. For the filler–gap dependency, we found a significant difference between the *control* and *island* condition in the case of the gulordava model ($p < 0.01$), but no such reduction in the case of the google model. For gender expectation, we found no significant difference between the two conditions.

Sentential Subject Islands The sentential subject constraint states that gaps are not licensed within an S-node that plays the role of a sentential subject. To assess whether the RNN models had learned this constraint we created items following the variants in (12).

(12)a. I know who the seniors defeated _ **last week**. [CONTROL, FILLER–GAP]

b. I know who for the seniors to defeat _ **will be trivial**. [ISLAND, FILLER–GAP]

c. The fireman knows they will save {**his/her**} friend. [CONTROL, GENDER EXP.]

d. The fireman knows for them to save {**his/her**} friend will be difficult. [ISLAND, GENDER EXP.]

The results for this experiment can be seen in Figure 3, in the far right panel. We found no decrease in gender expectation between the *control* and *island* conditions for either model. Likewise, for the filler–gap dependency we found no significant decrease in wh-licesning interaction between the island and non island conditions in either model. These results indicate that neither model suspends its expectations for gaps within sentential subjects.

Left Branch Islands The left-branch constraint states that modifiers which appear on the left branch under an NP cannot be gapped, which accounts for the relative ungrammaticality of (13-b) compared to (13-a). Because possessive pronouns cannot grammatically occur in left-branches under an NP, this experiment examines only the filler–gap dependency. We created 20 items following the examples in (13) and measured the wh-licensing interaction in the post-gap material.

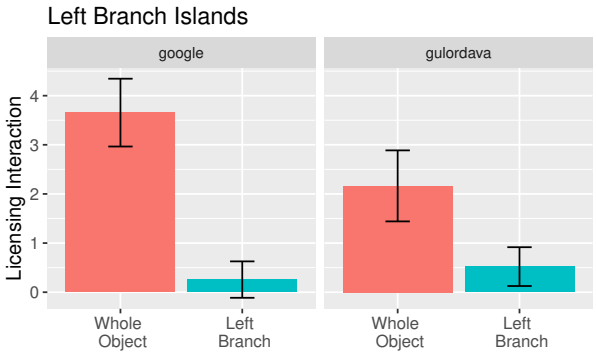

Figure 4: Left Branch Islands.

(13)a. I know what color car you bought _ **last week**. [WHOLE OBJECT]

b. I know what color you bought _ **car last week**. [LEFT BRANCH]

The results from this experiment can be seen in Figure 4 with experimental conditions on the x-axis and wh-licensing interaction on the y-axis. We see strong wh-licensing interaction in the two *whole object* conditions, but a significant reduction in licensing interaction when the gap consists of the Adjective Phrase modifier ($p < 0.001$ for the google model; $p < 0.05$ for the gulordava model). This results indicate that the models have learned the left branch islands, insofar as they do not expect left-branching modifiers to be extracted without the NP to which they are attached.

For every condition tested we found that the expectation set up by gendered subjects for possessive pronouns is not affected by the pronoun's location inside island constructions. For the three novel structures, we found that the two models tested are sensitive to left branch islands and gradiently to coordination islands, but not to sentential subject islands.

Discussion

The filler–gap dependency has been the focus of intense research for over fifty years because it is both far reaching and tightly constrained. It can be threaded through a potentially unbounded number of sentential embeddings; yet the filler must syntactically dominate the gap and the dependency is subject to a number of highly-specific blocking 'island' conditions. In this work we have shown that RNNs trained on a language modeling objective have learned both the power and the constraints imposed on this dependency. First, we provided evidence that they are able to thread the dependency through an unbounded number of sentential embeddings, and have also learned the constraints that govern the syntactic hierarchy of the filler relative to the gap.

Second, using gender expectation effects, we have demonstrated that the models are able to thread some contextually-dependent expectations into island constructions, providing evidence that previously-observed island effects have been learned for the filler–gap dependency *in particular*, and are not due to the model's inability to thread *any* information into syntactic islands. In addition, we have increased the experimental coverage of island effects, demonstrating that the models were able to learn left-branch islands and gradiently

learn coordination islands, but failed to learn sentential subject islands. This brings the total number of islands learned to 5/7 for the google model and 6/7 for the gulordava model. Although some of the model behavior remains strikingly unlike human acceptability judgements (in e.g. coordination islands), these experiments demonstrate that sequence models trained on a language modeling objective are able to separate natural language dependencies from each other and learn different fine-grained syntactic rules for each.

References

Baayen, R. H., Davidson, D., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Chowdhury, S. A., & Zamparelli, R. (2018). Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics* (pp. 133–144).

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*(2-3).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of naacl.*

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics and language technologies* (pp. 1–8).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv*, *1602.02410*.

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

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *57*(3), 203.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Stanford, CA: Center for the Study of Language and Information.

Ross, J. R. (1967). Constraints on variables in syntax.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Yoshida, M., Kazanina, N., Pablos, L., & Sturt, P. (2014). On the origin of islands. *Language, Cognition and Neuroscience*, *29*(7), 761–770.