What Don't RNN Language Models Learn About Filler-Gap Dependencies?

Rui P. Chaves
Linguistics Department
University at Buffalo – SUNY
rchaves@buffalo.edu

Abstract

In a series of experiments Wilcox et al. (2018, 2019b) provide evidence suggesting that general-purpose state-of-the-art LSTM RNN language models have not only learned English filler-gap dependencies, but also some of their associated 'island' constraints (Ross, 1967)). In the present paper, I cast doubt on such claims, and argue that upon closer inspection filler-gap dependencies are learned only very imperfectly, including their associated island constraints. I conjecture that the LSTM RNN models in question have more likely learned some surface statistical regularities in the dataset rather than higher-level abstract generalizations about the linguistic mechanisms underlying filler-gap constructions.

1 Introduction

Recurrent Neural Networks (RNNs) are a class of abstract neural network where the connections between nodes consist of a directed graph along a temporal sequence. This architecture allows node outputs at current time step to depend on the current input as well as on the previous output state. Thus, the network can exhibit temporal dynamic behavior, since the internal state of the system is a kind of memory that can be used to process subsequent input. Such models are therefore well-suited for natural language tasks, among others. RNNs with a Long Short-Term Memory (LSTM) architecture have a far more elaborate and selective form of memory. A common LSTM node is composed of a cell, an input gate, an output gate and a forget gate. Such gates enable RNN nodes to remember values over arbitrary time intervals and the three gates regulate the flow of information into and out of the nodes.

LSTM RNNs are therefore better suited than plain RNNs to model long-distance dependencies of the kind found in natural languages (Linzen et al., 2016; Gulordava et al., 2018; Bernardy and Lappin, 2017). This includes filler-gap dependencies like (1), where the *wh*-phrase *what* is interpreted as the object of *do*, even though the two words are separated by four clausal boundaries as indicated by square brackets.

(1) What$_i$ do you think [the students will say [they believe [the TA claimed [he was trying to do _$_i$]]]]?

I refer to the 'extracted' phrase as the *filler* and to the canonical position where it would otherwise be realized as the *gap*, signaled via an underscore. The filler-gap dependency is the semantic and syntactic linkage that must be established between the filler and its *in situ* canonical location in order for such utterances to be interpretable.

1.1 Learning Filler-Gap dependencies

Recently, Chowdhury and Zamparelli (2018) provide some evidence that LSTM RNNs can store information about the filler phrase, and detect that the probability of the sentence-final NP in examples like (2) is low because of the presence of a filler-gap dependency.

(2) Who$_i$ should Mia discuss _$_i$ / *this candidate.

Wilcox et al. (2018) improve on this work, and propose a Surprisal-based (Hale, 2001; Levy, 2008) differences-within-differences design to measure the ability of the RNN to learn filler-gap dependencies, using a factorial design as in (3).

(3) a. I know that the lion devoured a gazelle **at** sunrise.
[NO WH-LICENSOR, NO GAP]

b.*I know what the lion devoured a gazelle **at** sunrise.
[WH-LICENSOR, NO GAP]

c.*I know that the lion devoured _ **at** sunrise.
[NO WH-LICENSOR, GAP]

d. I know what$_i$ the lion devoured _$_i$ **at** sunrise.
[WH-LICENSOR, GAP]

Wilcox et al. define $S(w)$ as the surprisal of a given word w, estimated in terms of the log inverse probability of w according to the RNN's hidden state softmax activation h before consuming w, given all previous words in the sentence:

(4) $S(w) = -log_2\ p(w|h)$

If the model has learned to represent filler-gap dependencies, then the surprisal of the proposition *at* in (3a) should be a small number, since the probability of *at* in this context is high, and the surprisal of 'at' in (3b) should be a large number, since the probability of 'at' in this context is low. Consequently, their difference $S(3b) - S(3a)$ should yield a large positive number. Similarly, $S(3d) - S(3c)$ should yield a large negative number, and the full **licensing interaction** $(S(3b) - S(3a)) - (S(3d) - S(3c))$ should be a large positive number. This licensing interaction represents how well the network learns both parts of the licensing relationship: a positive wh-licensing interaction means the model represents a filler-gap dependency between the wh-word and the gap site; a licensing interaction indistinguishable from zero indicates no such dependency. Wilcox et al. find that typical models show about 4 bits of licensing interaction in simple examples like (3).

Using this design, Wilcox et al. (2019b) found that LSTM RNNs can maintain filler-gap dependencies across up to four clausal boundaries, not unlike the ones in (1). Two models were used for these experiments: (i) the model in Gulordava et al. (2018) – henceforth the **Gulordava model** – which was trained on 90 million tokens of English Wikipedia, and has two hidden layers of 650 units each; and (ii) Jozefowicz et al. (2016) – henceforth the **Google model** – which was trained on the One Billion Word Benchmark (Chelba et al., 2013), has two hidden layers with 8196 units each, and employs a character-level convolutional neural network.

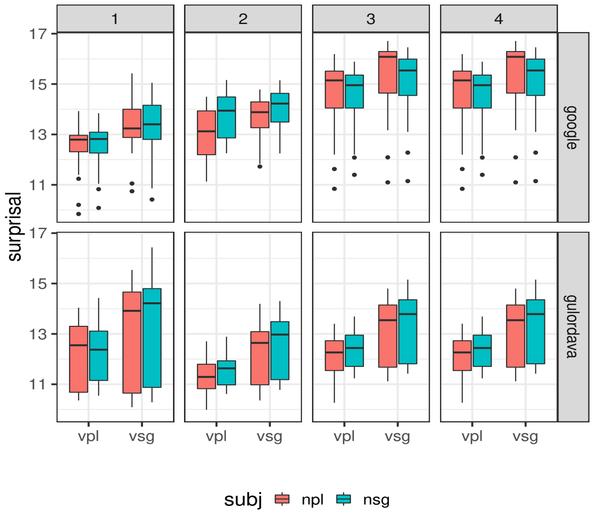

Figure 1: Surprisal at the gap-agreeing verb in 'which' interrogatives across embedding levels (LSTM RNNs)

But more recently Da Costa and Chaves (2020) shows that the Gulordava and Google LSTM models have learned filler-gap dependencies only very imperfectly. In particular, the models completely failed to learn that filler-gap constructions also impose agreement dependencies like those in (5). In such constructions, the singular/plural number information of the extracted phrase must match that of the verb from which the extraction takes place.

(5) a. They wondered which lawyer I think you said _ was/*were upset.

b. They wondered which lawyers I think you said _ *was/were upset.

Following the same factorial approach and code of Wilcox et al. (2018), Da Costa and Chaves (2020) extracted the softmax activation of the verbs *were*/*was* in 20 items like those illustrated in (6), up to four levels of clausal embedding.

(6) a. Someone wondered which lawyer(s) I think was/were ...
[N$_{sg/pl}$, LEVEL1, V$_{sg/pl}$]

b. Someone wondered which lawyer(s) I think you said was/were ...
[N$_{sg/pl}$, LEVEL2, V$_{sg/pl}$]

c. Someone wondered which lawyer(s) I think you said you thought was/were ...
[N$_{sg/pl}$, LEVEL3, V$_{sg/pl}$]

d. Someone wondered which lawyer(s) who people believe I think you said you thought was/were ...
[N$_{sg/pl}$, LEVEL4, V$_{sg/pl}$]

The results in Figure 1 show that both the Gulordava and the Google models failed. Had the

LSTM RNNs succeeded at this task, the conditions where the noun and verb agree (i.e. N_{pl}-V_{pl} and N_{sg}-V_{sg}) would be lower in surprisal than the conditions where the agreement is mismatched (i.e. N_{pl}-V_{sg} and N_{sg}-V_{pl}). Note also that in the Google model surprisal increased with the level of embedding, so that the correct verb form is more unexpected in level 4 than the incorrect verb forms in levels 1 and 2. Da Costa and Chaves (2020) tested other types of construction and the results are equally bad, suggesting that the Gulordava and Google models have not learned the morphosyntax of filler-gap dependencies, even though they were trained on datasets larger than what a child learner is exposed to; according to Atkinson et al. (2018), children begin to exhibit adult-like active formation of filler-gap dependencies by age 6.

1.2 Learning Island Constraints

Wilcox et al. (2018, 2019b) in addition claim that the Gulordava and Google models have learned certain constraints on filler-gap dependencies known as **Islands** (Ross, 1967). In particular, Wilcox et al. claim that the models learn that the subordinate clauses introduced by *whether* have reduced acceptability as in (7a), that relative clauses and adverbial adjuncts are difficult to extract from as in (7b,c), and that conjuncts and the left branches of NP are not possible to extract, as in (7d,e). All reported examples below are from Wilcox et. al experiments. Square brackets indicate the island-establishing environments.

(7) a.* I know what Alex said [whether your friend devoured _ at the party].
(**Wh-Island**)

b.*I know (that/what/who) the family bought the painting [that depicted _ last year].
(**Complex NP Constraint Island**)

c.*I know what the patron got mad [after the librarian placed _ on the wrong shelf].
(**Adjunct Constraint Island**)

d.*I know what the man bought [the painting and _] at the antique shop.
(**Conjunct Constraint island**)

e.*I know what color you bought [_ car] last week.
(**Left Branch Constraint island**)

However, Wilcox et.'s claims are too strong. First, most of these island constraints are more complex than Wilcox et. al's discussion suggest, and before it cannot be claimed that a model learns island constraints before all the associated conditions are shown to have been learned as well. For example, the Conjunct Constraint is but a piece of a larger set of constraints that are specific to coordination, known as the Coordinate Structure Constraint (CSC). The CSC consists of the Conjunct Constraint, the Element Constraint, the ATB Exception, and the Asymmetric Exception; see Kehler (2002, Ch.5) for a detailed overview and an account of most of these constraints that is based on pragmatic discourse relations.

The Complex NP Constraint (CNPC) is similarly complex. First, it is not restricted to relative clauses: nouns that semantically introduce propositional complements like in *the claim that Robin stole a book* also induce such extraction limitations (e.g. *$What_i$ did you reject the claim [that Robin stole $\_{i}$]?'). Second, it is also known that the CNPC vanishes in presentational relatives (i.e. in relatives that express assertions rather than presupposed content), as we discuss below.

Moreover, some of the island constraints that Wilcox *et al.* probed are know to be weakened when the island phrase is untensed, and vanish altogether if there is a secondary (i.e. 'parasitic') gap outside the adjunct (Engdahl, 1983); see Phillips (2006) for experimental evidence. In sum, there is a complex array of facts that still need to be tested.

Finally, the Left Branch Constraint (LBC) items that Wilcox *et al.* used, like (7e), have a critical confound. The sentences are not licit even without the extraction (i.e. **what color car*). And since the sentences are ill-formed, with or without extraction, it remains unclear whether the RNNs have or not learned the LBC.

But even conceding that the results are overall on the right track, there is one final problem. Both the Gulordava and Google models failed to learn that extraction from subject phrases (phrasal or clausal) is hampered, as illustrated in (8).

(8) a.*I know who [the painting by _] fetched a high price at auction.
(**Subject Constraint Island**)

b.*I know who [for the seniors to defeat _] will be trivial.
(**Sentential Subject Constraint Island**)

The difficulty in learning clausal Subject Island effects is unexpected because such islands are much stronger than Wh-islands. Not only the oddness induced by a Wh-island constraint violation is less pronounced than that of clausal Subject islands, but also because counterexamples to the former are much easier to find. Compare (7) with the acceptable counterpart in (9).

(9) Which shoes are you wondering [whether you should buy _]?

See Abrusán (2014, Ch.4) for strong evidence that Wh-islands and their exceptions are contingent on subtle semantic-pragmatic factors, not syntax. Indeed, there is growing evidence that many island constraints are at least in part due to non-syntactic factors, including pragmatics and processing biases; see Chaves and Putnam (2020) for a detailed overview. For example, counterexamples have been noted in the literature to all of the island constraints probed by Wilcox et al., with the exception of the Conjunct Constraint and the Left Branch Constraint islands; see Hofmeister and Sag (2010) and references cited. This includes Subject Islands involving VP subjects, as in the attested data in (10). See Huddleston et al. (2002, 1093,1094), Santorini (2007), and Chaves (2013) for more attestations.

(10) a. In his bedroom, which [to describe _ as small] would be a gross understatement, he has an audio studio setup.
[pipl.com/directory/name/Frohwein/Kym]

b. They amounted to near twenty thousand pounds, which [to pay _] would have ruined me. (Benjamin Franklin, William Temple Franklin and William Duane. 1834. Memoirs of Benjamin Franklin, vol 1. p.58)
[archive.org/details/membenfrank01frankrich]

c. The (...) brand has just released their S/S 2009 collection, which [to describe _ as noticeable] would be a sore understatement.
[missomnimedia.com/2009/page/2/?s=art+radar&x=0&y=0]

d. Because this does purport to be a food blog, I will move from the tv topic to the food court itself, which [to describe _ as impressive] would be an understatement.
[phillyfoodanddrink.blogspot.com/2008/06/foodies-food-court.html]

All of these counterexamples involve restrictive relative clauses, suggesting that the Subject Condition is sensitive to pragmatics (Abeillé et al., 2018; Chaves and Dery, 2019).

The point here is a cautionary one: many island constraints are not absolute, and come with a complex array of patterns, many of which are still poorly understood. It cannot be claimed that a given language model has learned an island constraint before showing that both the negative and the positive cases (if any exist) have been correctly learned as well.

Note also that the Gulordava and the Google models did not perform in the same way at learning these island constraints: whereas the Google model failed to learn CNPC islands when the word 'that' appears instead of 'who/what', the Gulordava model failed to learn Wh-Islands. The performance of the Google was not significantly better that Gulordava's even though the former was originally trained with ten times more data than the latter, contained ten times as many hidden units, and used character CNN embeddings. This again suggests that something fundamental about filler-gap dependencies is being missed.

The question then becomes: are these models actually learning filler-gap dependencies or are they simply learning surface-based contingencies that have little to do with the underlying syntactic and semantic mechanisms that cause island phenomena? As Jo and Bengio (2017) demonstrate, neural networks tend to learn surface statistical regularities in the dataset rather than higher-level abstract concepts; for adversarial research showing this to be the case in the language domain see Jia and Liang (2017) and Iyyer et al. (2018), for instance. Indeed, Marvin and Linzen (2018) found that LSTM RNNs fail to learn reflexive pronoun agreement and negative polarity licensing, and Wilcox et al. (2019a) showed that such models learn center-embedding dependencies only imperfectly. In the remainder of this paper the same models, code and licensing interaction approach of Wilcox et al. (2018) is used to provide evidence suggesting that these LSTM RNNs merely capture partial and superficial morphosyntactic properties of filler-gap dependency constraints. The present results are consistent with those of Wilcox et al. (2019a), in which these models are not fully able to suppress expectations for gaps inside at least some island environments and recover them later.

2 Extraction from Relative Clauses

Wilcox et al. (2018) found that evidence suggesting that both the Google and the Gulordava models have learned the CNPC. However, the CNPC is not without principled exceptions. It is well-known that CNPC effects systematically vanish in existential relative clauses (Erteschik-Shir and Lappin, 1979; McCawley, 1981; Chung and McCloskey, 1983) as in (11). See Kush et al. (2013) for experimental evidence that existential relatives are not island inducing syntactic environments.

(11) a. This is the kind of weather that there are [many people who like _]. (Erteschik-Shir and Lappin, 1979)

b. There were several old rock songs that she and I were [the only two who knew _]. (Chung and McCloskey, 1983)

c. John is the sort of guy that I don't know [a lot of people who think well of _]. (Culicover, 1999, 230)

d. Which diamond ring did you say there was [nobody in the world who could buy _]? (Pollard and Sag, 1994, 206)

Such relatives are special in that they express assertions rather than presupposed content, and the extraction is thus arguably acceptable because the referent that is questioned is part of the content that is asserted and at-issue (Goldberg, 2013). It should be relatively easy for the models to use the *there be* sequence as a cue that these constructions are different from other relatives. If Google and Gulordova's RNN models have learned the CNPC rather than superficial contingencies then the existence of a second gap inside an existential relative should not cause a large spike in surprisal and the licensing interaction should be small, or ideally, close to zero. For this purpose 18 experimental items were taken from Kush et al. (2013) and adapted to the present task, using the methodology as Wilcox et al. A sample is in (12).[1]

(12) a. It was known that there were many mathematicians who worked on the project **for** years.
[NO WH-LICENSOR, NO GAP]

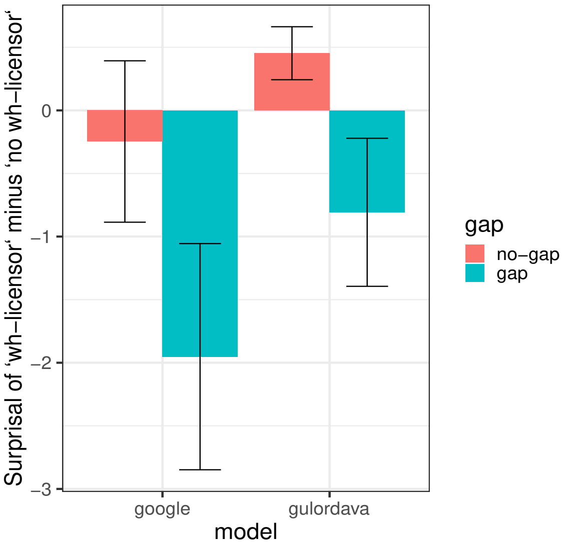

Figure 2: Licensing Interaction in Existential Relatives

b.*This was the problem which there were many mathematicians who worked on the project **for** years.
[WH-LICENSOR, NO GAP]

c.*It was known that there were many mathematicians who worked on _ **for** years.
[NO WH-LICENSOR, GAP]

d. This was the problem which there were many mathematicians who worked on _ **for** years.
1 [WH-LICENSOR, GAP]

Ideally, the no-gap condition interaction S(12b) − S(12a) should be a positive number, and the gap condition interaction S(12d) − S(12c) a negative number. As the graphs in Figure 2 indicate, this is what was found for the Gulordava model, but not for Google's. In the latter, the no-gap condition is indistinguishable from zero (t = -0.75, p = 0.46) suggesting that the latter model overlooks the subject gap. That said, the full wh-licensing interaction values are clearly positive, and in the order of about 1.5 bits. This is much lower than the 4 bits found by Wilcox et al. (2018), but nonetheless suggests that at least some aspects of the filler-gap dependency are detected by the models. Many other attempts were made to arrive at stronger results, with different materials, but the results invariably had similar outcomes, with the 'no-gap' bars either being indistinguishable from zero or negative. I now move on to islands which are not as strongly correlated with surface cues.

[1] Only verbs that strongly require complements were employed, and that-relatives were avoided given that the models have difficulty with them according to Wilcox et al. (2018).

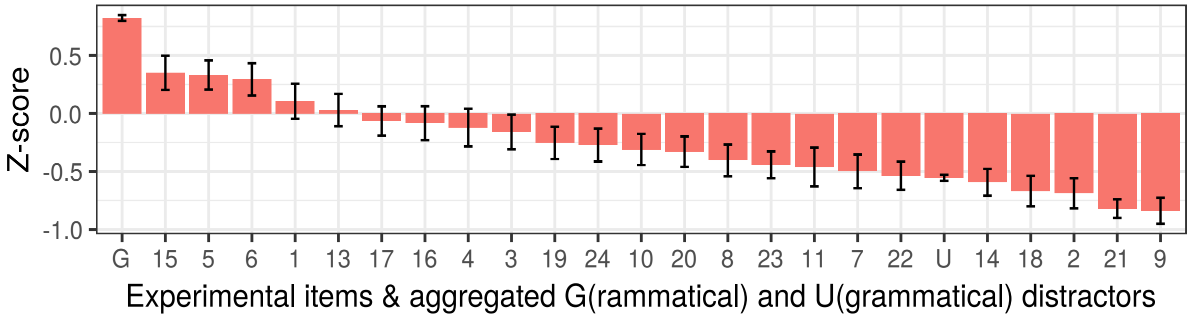

Figure 3: Acceptability ratings by item (with grammatical (G) and ungrammatical (U) distractors aggregated)

3 Extraction from Adjunct Clauses

Wilcox et al. (2018) probed the strongest type of adjunct island (tensed adjuncts), traditionally regarded as exceptionless since Huang (1982). But recent work has revealed that exceptions do exist; see Kluender (1998, 267), Truswell (2011, 175, ft.1), Levine and Hukari (2006, 287), and Goldberg (2006, 144). For example, Sprouse et al. (2016) found no evidence of an island effect in examples like (13), in terms of sentence acceptability rating, but found strong evidence of island effects in other adjunct island examples.

(13) I called the client [who]$_i$ the secretary worries [if the lawyer insults _$_i$].
(Sprouse et al., 2016)

Similarly, Müller (2017) experimentally shows that Swedish conditional adjuncts seem to yield much weaker island effects than causal adjuncts, and Kohrt et al. (2018) found experimental evidence that (non-clausal) English adjunct islands are contingent on semantic factors. In more recent work, Chaves and Putnam (2020) provide experimental evidence suggesting that Mueller's results likely extend to English as well. Chaves and Putnam (2020) report a sentence acceptability experiment with 24 items falling into three conditions, illustrated in (14).

(14) a. Who$_i$ did Sue blush [when she saw _$_i$]? [TEMPORAL ADJUNCT]

b. What$_i$ did Tom get mad [because Phil forgot to say _$_i$]? [CAUSAL ADJUNCT]

c. What$_i$ does Evan get grumpy [if he is told to do _$_i$]? [CONDITIONAL ADJUNCT]

I what follows I briefly describe this experiment in more detail, with the aim of repurposing the items for a counterpart experiment using the Gulordava and Google models. Each item was interspersed and pseudo-randomized with 36 filler phrases, half of which are ungrammatical, as illustrated in (15). The grammatical distractors were immediately followed by Yes/No comprehension questions, and the mean comprehension question accuracy was 86%.

(15) a.*Who does the union identify as having most recently fired from _?

b. What did the editor recommend should be revised _?

Chaves and Putnam analyzed data from 38 English native speakers, who were asked to rate the acceptability of each experimental item on a 5-point Likert scale. There was a wide range of acceptability scores, from fairly high in the acceptability scale to very low, as seen in Figure 3. The (aggregate) ratings for the grammatical (G) and the ungrammatical (U) distractors are included, for comparison. Conditional adjuncts were clustered at the high end of the ratings, temporal adjuncts in the middle, and causal adjuncts at the bottom.

I now describe how the stimuli from this experiment was repurposed to the same task that Wilcox et al. (2018) employed. The top 5 human-rate rated items (High Acceptability condition) received a mean acceptability of 3.30 ($SD = 0.2$), and the bottom human-rated 5 rated items (Low Acceptability condition) received a mean acceptability of 1.95 ($SD = 0.13$). These 10 items were selected and adapted to the $3 \times 2 \times 2$ factorial licensing interaction methodology of Wilcox et al. (2018). The counterparts of the item in (14c) are shown in (16) and (17) for illustration. In a nutshell, all items were embedded under 'I know' and

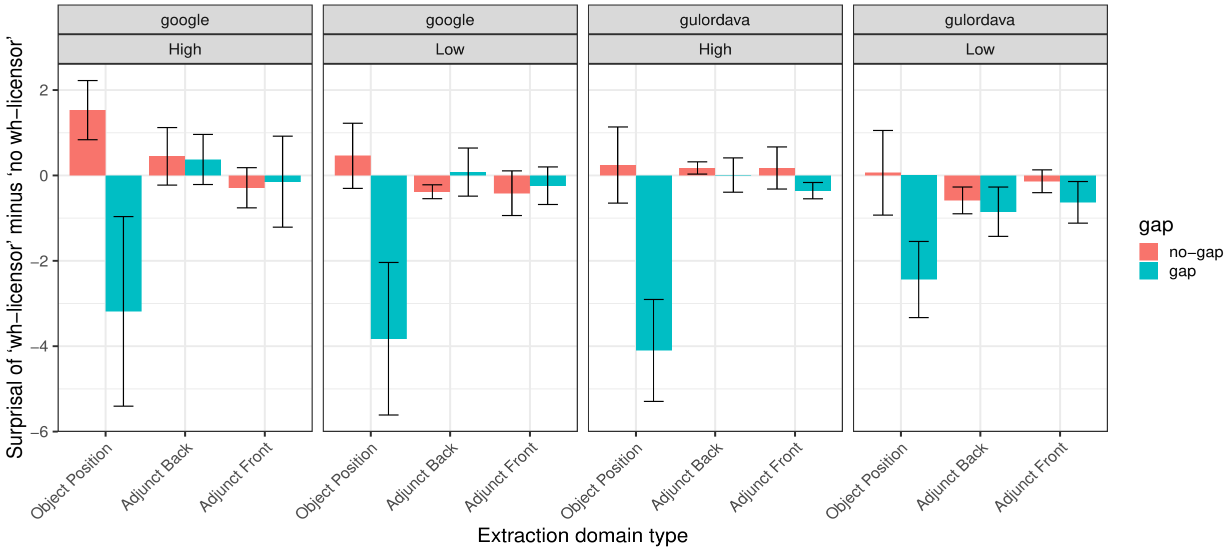

Figure 4: Effect of extraction site on wh-licensing interaction for adjunct islands, across high/low acceptability

all proper names were replaced with pronouns. In the Object condition there is no adjunct clause.

(16) a. I know that they usually are told to do the homework **in** the morning.
[OBJECT, NO WH-LICENSOR, NO-GAP]

b.*I know what they usually are told to do the homework **in** the morning.
[OBJECT, WH-LICENSOR, NO-GAP]

c.*I know that they usually are told to do _ **in** the morning.
[OBJECT, NO WH-LICENSOR, GAP]

d. I know what they usually are told to do _ **in** the morning.
[OBJECT, WH-LICENSOR, GAP]

In the Adjunct back condition there is an adjunct clause at the end of the sentence, as in (17). Following Wilcox et al. (2018), there was a third condition where the adverbial clause is fronted, and appears immediately after the complementizer *that* rather than at the end of the utterance.

(17) a. I know that the kids get grumpy if they are told to do the homework **in** the morning.
[ADJUNCT BACK, NO WH-LICENSOR, NO-GAP]

b.*I know what the kids get grumpy if they are told to do the homework **in** the morning.
[ADJUNCT BACK, WH-LICENSOR, NO-GAP]

c.*I know that the kids get grumpy if they are told to do _ **in** the morning.
[ADJUNCT BACK, NO WH-LICENSOR, GAP]

d. I know what the kids get grumpy if they are told to do _ **in** the morning.
[ADJUNCT BACK, WH-LICENSOR, GAP]

If the Gulordava and Google models have learned the subtleties of the tensed Adjunct Constraint then the filler-gap dependencies in the High Acceptability condition items should have a significantly lower surprisal than the Low Acceptability condition items. In order to access this, the surprisal of the word after the critical region was measured. Focusing on the object items first, interactions of the type $S(16\text{b}) - S(16\text{a})$ should ideally result in a positive number, however, for both High acceptability or Low acceptability items. This was the case in the Google model, but not for the Gulordava model, as Figure 4 shows; perhaps the latter model discovered that a gap after the preposition in (16b) is not necessarily out of the question. $S(16\text{d}) - S(16\text{c})$ yielded the expected highly negative values, as illustrated by the long teal bars.

Moving on to the Adjunct back items, the interactions of the type $S(17\text{b}) - S(17\text{a})$ should ideally result in a positive number as usual, contrary to fact, and $S(17\text{d}) - S(17\text{c})$ should ideally result in a negative number in the High acceptability condition and cancel out in the Low acceptability conditions. Neither result occurred because the interaction values were centered around zero. The full licensing interaction $(S(17\text{b}) - S(17\text{a})) - (S(17\text{d}) - S(17\text{c}))$ is shown in Figure 5. None of the Adjunct front/back High/Low conditions is statistically distinguishable from zero, although significance is approached ($t = 2.73, p = 0.052$) in the

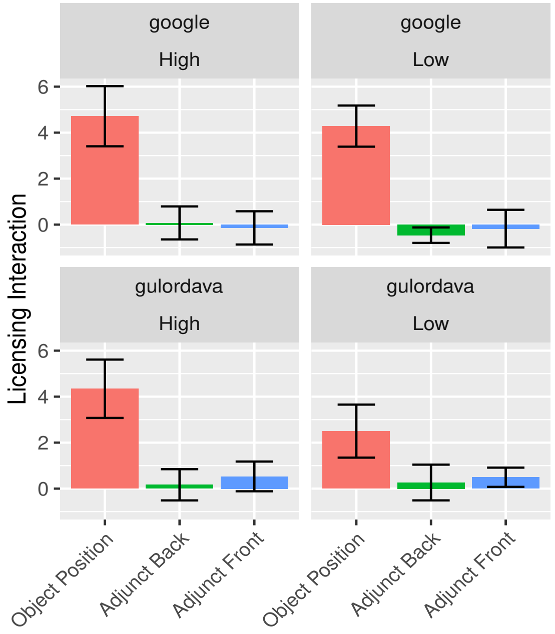

Figure 5: Full licensing interaction for Adjunct Islands

case of Adjunct front High for Gulordava.

In sum, all extractions from clausal adjuncts are ultimately deemed islands environments by the models, contrary to the human judgments.

4 Extraction from Negative Phrases

Negative Islands are perhaps the clearest type of island in which semantic and pragmatic factors play a key role. Consider the examples in (18).

(18) a.*Which country weren't you born in _?

b.*How many kids don't you have _?

c.*How fast didn't John drive _?

The question in (18a) presupposes that the addressee was born in all countries but one, which is contrary to world knowledge, and therefore infelicitous (Kuno and Takami, 1997). Hence, the oddness vanishes if the verb is not a one-time predicate, as in (19).

(19) Which country haven't you visited _ yet?

The oddness of the degree questions in (18b,c) is due to an analogous reason; see Abrusán (2011) for detailed discussion. It is again clear that the oddness is caused by semantic factors, since the introduction of existential modals makes the island effect vanish (Fox and Hackl, 2006):

(20) a. How many kids can't you have _?

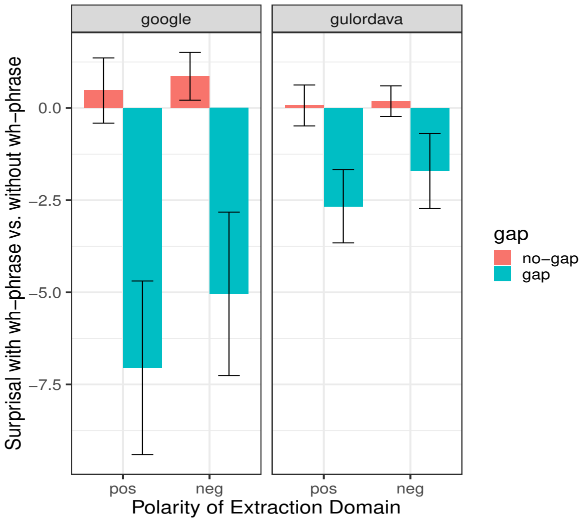

Figure 6: Wh-licensing in negative phrases

b. How fast is John required not to drive _?

In order to evaluate whether RNNs are sensitive to such effects 14 items were constructed in a 2×2×2 design, as illustrated in (21). The verb is negated in items in the negative (NEG) condition.

(21) a. I wonder if the owner of the truck has (not) driven at this speed **during** the race. [NO WH-LICENSOR, POS/NEG, NO GAP]

b.*I wonder how fast the owner of the truck has (not) driven at this speed **during** the race. [WH-LICENSOR, POS/NEG, NO GAP]

c.*I wonder if the owner of the truck has (not) driven at _ **during** the race. [NO WH-LICENSOR, POS/NEG, GAP]

d. I wonder how fast the owner of the truck has (*not) driven at _ **during** the race. [WH-LICENSOR, POS/NEG, GAP]

The results are shown in Figure 6. The interaction $S(21\text{b}) - S(21\text{a})$ should have resulted in a moderate-to-large positive numbers, regardless of the presence of negation. In other words, the red bars should be positive and not overlap with zero. This was not true of either model, especially for Gulordava. Conversely, $S(21\text{d}) - S(21\text{c})$ should have yielded a moderate-to-large negative number in the pos(itive) condition but obtain a significantly higher value in the neg(ative) condition (ideally, close to zero). However, there was no statistically significant difference between the interaction values across the two island conditions (pos and neg) for the Google model ($t = 0.3$, $p = 0.73$)

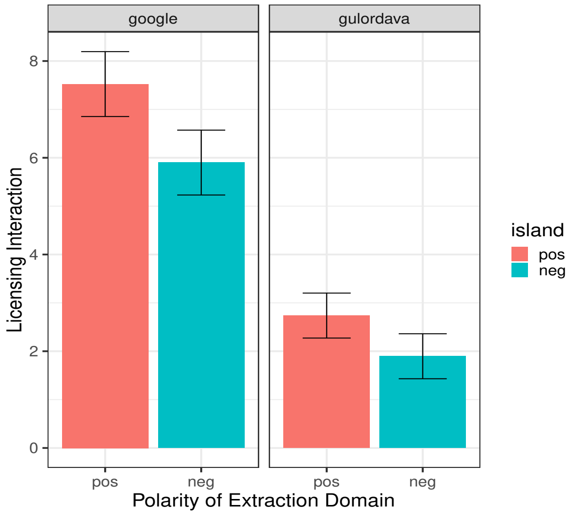

Figure 7: Full licensing interaction for negative islands

nor for the Gulordava model ($t = 1.11, p = 0.27$). The full interactions are shown in Figure 7. Had Negative Islands been learned, the teal bars would be centered around zero, like those in in Figure 5.

5 Discussion

The claim that sate-of-the-art LSTM RNNs models have learned filler-gap dependencies and islands is premature on both linguistic and experimental grounds. First, the linguistic constraints in question are far more complex than what extant studies consider. Second, there is evidence that these models only learn partial contingencies about filler-gap dependencies, which suggests that the actual linguistic mechanism that underlies such long-distance phenomena is not accessible to the model.

The problem is arguably not due to a lack of data. The training datasets for Gulordava and Google are unrealistically large when compared to the amount of linguistic input the average child is exposed to (Atkinson et al., 2018). Similarly, the problem is not likely to be due to lack of expressivity, since this kind of model is Turing-complete; see Siegelmann and Sontag (1995) and Siegelmann (1999, 29–58) for proofs and examples, as well as Hornik et al. (1989) and Lu et al. (2017) for detailed discussion about Cybenko's universal approximation theorem.

The present findings suggest that model size and training regimen yield diminishing returns, and that there is a more fundamental factor preventing such systems to learn filler-gap dependencies. The problem likely stems from the fact that filler-gap dependencies are not merely surface string patterns: they involve rich morphological, syntactic and semantic dependencies which crucially interact with pragmatics and world knowledge, thus far absent from training. Most crucially, many island phenomena seems to be sensitive to semantic and pragmatic constraints, including the Subject Constraint (Chaves and Dery, 2019; Abeillé et al., 2018), the Adjunct Constraint (Truswell, 2011; Müller, 2017; Kohrt et al., 2018; Goldberg, 2013), the Complex NP Constraint (Erteschik-Shir and Lappin, 1979; Goldberg, 2013), the Coordinate Structure Constraint (Kehler, 2002, Ch.5), Wh-Islands Abrusán (2014, Ch.4), Negative Islands (Abrusán, 2011), among others. See Chaves and Putnam (2020) for extensive discussion of these and other island effects.

In sum, it not clear how current neural models can learn island constraints from stringsets alone, precisely because of the subtle semantic and pragmatic properies that underpin the phenomena in question. The present findings are consistent with the fact that Marvin and Linzen (2018) found that LSTM RNNs fail to learn other complex phenomena such as reflexive pronoun agreement, negative polarity licensing, and center-embedding dependencies (Wilcox et al., 2019a).

All experimental items and statistical analysis scripts are made available online at https://github.com/RuiPChaves/LSTM-RNN-unbounded-dependency-experiments. The code to run the models is the same as Wilcox et al. (2018).

References

Anne Abeillé, Barbara Hemforth, Elodie Winckel, and Edward Gibson. 2018. A construction-conflict explanation of the subject-island constraint. 31th Annual CUNY Conference on Human Sentence Processing.

Márta Abrusán. 2011. Presuppositional and negative islands: A semantic account. *Natural Language Semantics*, 19:257–321.

Márta Abrusán. 2014. *Weak island semantics*. Oxford University Press, Oxford.

Emily Atkinson, Matthew W. Wagers, Jeffrey Lidz, Colin Phillips, and Akira Omaki. 2018. Developing incrementality in filler-gap dependency processing. *Cognition*, 179:132 – 149.

Jean-Philippe Bernardy and Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, 15(2):1–15.

Rui P. Chaves. 2013. An expectation-based account of subject islands and parasitism. *Journal of Linguistics*, 2(49):285–327.

Rui P. Chaves and Jeruen E. Dery. 2019. Frequency effects in subject islands. *Journal of Linguistics*, page 147.

Rui P. Chaves and Michael T. Putnam. 2020. *Unbounded Dependency Constructions: theoretical and experimental perspectives*. Oxford University Press, Oxford.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144.

Sandra Chung and James McCloskey. 1983. On the interpretation of certain island facts in GPSG. *Linguistic Inquiry*, 14:703–714.

Peter W. Culicover. 1999. *Syntactic Nuts: Hard Cases in Syntax*. Volume One of Foundations of Syntax. Oxford: Oxford University Press.

Jillian K. Da Costa and Rui P. Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *SCil*, page 10.

Alex Drummond. 2013. Ibex 0.3.7 manual. spellout.net/latest_ibex_manual.pdf.

Elisabet Engdahl. 1983. Parasitic gaps. *Linguistics and Philosophy*, 6:3–34.

Nomi Erteschik-Shir and Shalom Lappin. 1979. Dominance and the functional explanation of island phenomena. *Theoretical Linguistics*, 6:41–86.

Danny Fox and Martin Hackl. 2006. The universal density of measurement. *Linguistics and Philosophy*, 29:537–586.

Adele E. Goldberg. 2006. *Constructions at Work: the nature of generalization in Language*. Oxford: Oxford University Press.

Adele E. Goldberg. 2013. Backgrounded constituents cannot be extracted. In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Island Effects*, pages 221–238. Cambridge: Cambridge University Press.

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

John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001, Pittsburg, PA*, pages 159–166. ACL.

Philip Hofmeister and Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language*, 86(2):366–415.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Cheng-Teh James Huang. 1982. *Logical relations in Chinese and the theory of grammar*. Ph.d. thesis, MIT.

Rodney D. Huddleston, Geoffrey Pullum, and Peter Peterson. 2002. *The Cambridge Grammar of the English Language*, chapter 12: Relative clause constructions and unbounded dependencies. Cambridge: Cambridge University Press.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL-HLT 2018*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Jo and Yoshua Bengio. 2017. Measuring the tendency of CNNs to learn surface statistical regularities. *CoRR*, abs/1711.11561.

Rafal Jozefowicz, Vinyals Oriol, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*.

Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. Stanford, CA: CSLI Publications.

Robert Kluender. 1998. On the distinction between strong islands and weak islands: a processing perspective. In Peter W. Culicover and Louise McNally, editors, *Syntax and Semantics 29: The Limits of Syntax*, pages 241–279. New York, NY: Academic Press.

Annika Kohrt, Trey Sorensen, and Dustin A. Chacón. 2018. The real-time status of semantic exceptions to the adjunct island constraint. In *Proceedings of WECOL 2018: Western Conference on Linguistics*.

Susumu Kuno and Ken-ichi Takami. 1997. Remarks on negative islands. *Linguistic Inquiry*, 28:553–576.

David Kush, Akira Omaki, and Norbert Hornstein. 2013. Microvariation in islands? In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Island Effects*, pages 239–264. Cambridge: Cambridge University Press.

Robert D. Levine and Thomas E. Hukari. 2006. *The unity of unbounded dependency constructions*. Stanford, CA: CSLI Publications.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqian Hu, and Liwei Wang. 2017. The expressive power of neural networks: A view from the width. In *Neural Information Processing Systems*, pages 6231–6239.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

James D. McCawley. 1981. The syntax and semantics of english relative clauses. *Lingua*, 53:99–149.

Christiane Müller. 2017. Extraction from adjunct islands in Swedish. *Norsk Lingvistisk Tidsskrift*, 35(1):6785.

Colin Phillips. 2006. The real-time status of island phenomena. *Language*, 82:795–823.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press and Stanford: CSLI.

John R. Ross. 1967. *Constraints on Variables in Syntax*. Ph.d. dissertation, MIT, Cambridge, Massachusetts. [Published in 1986 as *Infinite Syntax!* Norwood, NJ: Ablex Publishing].

Beatrice Santorini. 2007. (Un?)expected movement. University of Pennsylvania. http://www.ling.upenn.edu/ beatrice/ examples/movement.html. Accessed: Jun 14 2019.

Hava T. Siegelmann. 1999. *Neural Networks and Analog Computation: Beyond the Turing Limit*. Progress in Theoretical Computer Science. Birkhäuser, Boston, MA.

Hava T. Siegelmann and E. Sontag. 1995. On the computational power of neural nets. *Journal of Computer and Systems Sciences*, 50(1):132–150.

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.

Robert Truswell. 2011. *Events, Phrases and Questions*. Oxford: Oxford University Press.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of Blackbox NLP at ACL*, page pp.10.

Ethan Wilcox, Roger P. Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? In *Proceedings of the Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Ethan Wilcox, Roger P. Levy, Takashi Morita, and Richard Futrell. 2019b. What syntactic structures block dependencies in RNN language models? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci)*.