How well do LSTM language models learn filler-gap dependencies?

Satoru Ozaki, Dan Yurovsky, Lori Levin
Carnegie Mellon University
{ ikazos, yurovsky }@cmu.edu, levin@andrew.cmu.edu

Abstract

This paper revisits the question of what LSTMs know about the syntax of filler-gap dependencies in English. One contribution of this paper is to adjust the metrics used by Wilcox et al. (2018) and show that their language models (LMs) learn embedded *wh*-questions – a kind of filler-gap dependencies – better than they originally claimed. Another contribution of this paper is to examine four additional filler-gap dependency constructions to see whether LMs perform equally on all types of filler-gap dependencies. We find that different constructions are learned to different extents, and there is a correlation between performance and frequency of constructions in the Penn Treebank Wall Street Journal corpus.

1 Introduction

Language models (LMs) that use recurrent neural networks (RNNs, Elman, 1990), especially those adopting the long short-term memory (LSTM, Hochreiter and Schmidhuber, 1997) architecture, achieve outstanding performance in various natural language processing tasks. The fact that the same architecture yields high performance across many tasks seems to suggest that these LMs are learning something fundamental about natural language.

But what does it mean to learn a language, and have neural networks really achieved language acquisition? Much recent work focuses on evaluating neural networks' understanding of various syntactic phenomena that occur in natural language, such as subject-verb agreement (Linzen et al., 2016; Bernardy and Lappin, 2017; Kuncoro et al., 2018; Gulordava et al., 2018), negative polarity item licensing (Futrell et al., 2018; Jumelet and Hupkes, 2018; Marvin and Linzen, 2018), and anaphora (Marvin and Linzen, 2018; Warstadt et al., 2019).[1] These studies typically take a pre-trained LM or train one from scratch, and test the LM's performance on a dataset of artificially constructed linguistic expressions or a curated subset of real-world linguistic utterances, which pertain to particular syntactic phenomena of the researcher's interest.

Following Chowdhury and Zamparelli (2018), Wilcox et al. (2018, 2019b) and others, we focus on English filler-gap dependencies because of their three interesting properties: (a) bijectivity of filler and gap, (b) unboundedness, and (c) sensitivity to island constraints. We will review these in more detail in Section 2.

Wilcox et al. (2018) address whether neural networks know the bijectivity property: fillers are bad without gaps and gaps are bad without fillers. Their LMs detect that a filler is better (less surprising) with a gap than without a gap, but they do not fully capture bijectivity. One contribution of our paper is to experiment with different types of probabilistic metrics. With our changes, we show that Wilcox et al.'s models do in fact fully capture the bijectivity property for one metric.

English has a variety of filler-gap dependency constructions, which share the same three properties. While some linguists have analyzed these constructions as generated from a common abstract syntactic mechanism such as *wh*-movement (Chomsky, 1977), others have analyzed them as a mixture of idiosyncractic constructions (Sag, 2010). Do LMs capture the same properties across all constructions, or does performance vary over constructions? In this paper, we extend the work of Wilcox et al. (2018) to include four additional filler-gap dependency constructions, and examine their behavior collectively and individually to see how they bear on the issues of general mechanisms and specific constructions in human language.

[1] For a summary of the types of syntactic phenomena tested and the list of studies in the literature for each type, see Warstadt et al. (2020).

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 76-88.
Held on-line February 7-9, 2022

This paper is structured as follows. In Section 2, we review the properties of English filler-gap dependencies and previous work on RNNs' acquisition of filler-gap dependencies. In Section 3, we revisit Wilcox et al. (2018)'s experiment, revise their metrics and propose stricter criteria for the acquisition of filler-gap dependencies. We show that for one metric, their LMs understand filler-gap dependencies better than they had previously claimed. In Section 4, we check if LMs learn four other kinds of filler-gap dependency constructions, and their interaction with island constraints and embedding depth. We see that different constructions are learned to different extents. In Section 5, we test if the performance for each type of construction we obtain in Section 4 correlates with the relative frequency of these constructions in a written-text English corpus. Finally, in Section 6, we conclude our findings.

2 Properties of English Filler-Gap Dependencies

(1-a) is an example of a *wh*-question, which is a filler-gap dependency construction in English. The verb *put* is followed by a **gap**, indicated by an underscore, which is an empty position that would canonically be occupied as in *We put the book on the table*. The word *what* is understood to fill the gap and is called the **filler**. The filler and the gap are marked by a common subscript index.

(1) a. **What**$_i$ did you put ___$_i$ on the table?
b. *__What__$_i$ did you put **it**$_i$ on the table?
c. *You put ___$_i$ on the table.
d. You put **it** on the table.

English has several kinds of filler-gap dependency constructions, including comparatives (2-a), *it*-clefts (2-b), topicalization (2-c), embedded *wh*-questions (2-d), *tough*-movement (2-e) and a few others (Chomsky, 1977; Huddleston and Pullum, 2002; Sag, 2010; Chaves and Putnam, 2021, among others).

(2) a. Maryanne read **more books**$_i$ this month than Alfred read ___$_i$ last month.
b. It was **Anna**$_i$ that Kevin talked to ___$_i$.
c. **These movies**$_i$, Antonio wishes he had never seen ___$_i$.
d. Someone figured out **who**$_i$ Margaret was describing ___$_i$.
e. **Thomas**$_i$ was difficult to persuade ___$_i$.

Filler-gap dependencies are of interest for at least three reasons. First is the property of bijectivity of filler and gap: there can be no gap without a filler and no filler without a gap. (1-b) is ungrammatical because there is a filler (*what*) but where we would expect a gap, there is a pronoun (*it*). Conversely, (1-c) is ungrammatical because there is a gap, but no filler.

Second, filler-gap constructions are unbounded in the sense that the filler and gap can be separated by a potentially unlimited number of clausal boundaries (three in (3)). This poses a challenge to language modelling, as these dependencies must be modelled robustly across arbitrarily many intervening words.

(3) **What**$_i$ did Rebecca believe [you and Albert said [the professor thought [she already discussed ___$_i$ last week]]] ?

Finally, the availability of filler-gap dependencies is constrained by complex structural restrictions. This is illustrated in (4-a), which is a paraphrase of (4-b). Though the two questions differ only minimally in their structure, (4-a) is ungrammatical while (4-b) is not. On the other hand, (4-c), which has the same structure as (4-a) but no filler-gap dependency, is grammatical. This shows that there is a constraint that disallows filler-gap dependencies across a kind of structure unique to (4-a). The precise identification and characterization of such constraints are challenging for linguists, and the mere existence of such constraints poses a challenge for language acquisition researchers: how do children acquire such complex structural linguistic rules from exposure to positive evidence alone?

(4) a. *__What__$_i$ did Rebecca believe your claim that the professor discussed ___$_i$?
b. **What**$_i$ did Rebecca believe you claimed that the professor discussed ___$_i$?
c. Did Rebecca believe your claim that the professor discussed **this**?

There is much debate on the question of how well RNNs learn filler-gap dependencies. Chowdhury and Zamparelli (2018) claim that GRU and LSTMs produce higher perplexity and cross-entropy loss for ungrammatical, gapless *wh*-questions than for their grammatical, gapped counterparts (e.g. *Which candidate should the students discuss ___/*him?*). However, their performances are heavily affected by sentence processing factors. Wilcox et al. (2018, 2019b, *et seq.*) look at two pre-trained LSTM LMs and define a metric called *wh*-licensing interac-

tion, which measures the extent to which the surprisal of a gapped clause is reduced significantly by the presence of the licensor. Using this metric, they show their LMs learn several structural properties of filler-gap dependencies as well as certain island constraints. On the other hand, Da Costa and Chaves (2020) and Chaves (2020) study the same LMs with respect to number agreement between the head noun and the verb of a relative clause (e.g. *which lawyer I think was/*were ...*) and observe that the LMs become less sensitive to agreement violations as the dependency crosses increasing levels of embeddings. They also claim that the island constraints Wilcox et al. (2018) purport these LMs to learn have certain exceptions, which are not acquired by these LMs.

3 Study 1: Surprisal and grammaticality

Wilcox et al. (2018) use a 2x2 factorial design as in (5), differing by [licensor], i.e. the presence/absence of the licensor, and [gap], i.e. the presence/absence of the gap. Their data consists entirely of embedded *wh*-questions. The filler is called a **licensor** because the gap cannot occur without it.

(5) a. I know that the lion devoured a gazelle at sunrise. [-licensor, -gap]
b. *I know what the lion devoured a gazelle at sunrise. [+licensor, -gap]
c. *I know that the lion devoured ___ at sunrise. [-licensor, +gap]
d. I know what the lion devoured ___ at sunrise. [+licensor, +gap]

They experiment on two pre-trained LSTM LMs. The first is the **Google model** (Jozefowicz et al., 2016). Trained on the One Billion Word Benchmark (Chelba et al., 2013), it consists of two hidden layers with 8196 units each. The second is the **Gulordava model** (Gulordava et al., 2018). Trained on 90 million tokens of English Wikipedia, it consists of two hidden layers with 650 units each.

The metric designed by Wilcox et al. builds on the definition of surprisal in (6) (Hale, 2001; Levy, 2008; Smith and Levy, 2013), where $S(w_k)$ is the surprisal generated by an RNN upon seeing the word w_k in a sentence, and h_{k-1} is the RNN's hidden state after consuming all previous words in the sentence. The probability is calculated from the RNN's softmax activation.

(6) $S(w_k) = -\log_2 \mathbb{P}(w_k|h_{k-1})$

For each experimental item, Wilcox et al. measures surprisal at two places: summed over a region immediately following the potential gap (emphasized in (7-a)), and summed over the entire embedded clause following the potential licensor (emphasized in (7-b)). The former, which we call **local surprisal**, reflects any local effects from the gap's licitness, while the latter, which we call **global surprisal**, reflects global expectations about the general well-formedness of the sentence.

(7) a. I know that/what the lion devoured (a gazelle) *at sunrise* .
b. I know that/what *the lion devoured (a gazelle) at sunrise* .

One can thus extend the definition of surprisal to be a function of experimental items, i.e. sentences. Then, Wilcox et al. define a metric they call *wh*-licensing interaction, as $(S([\text{+licensor, -gap}]) - S([\text{-licensor, -gap}])) - (S([\text{+licensor, +gap}]) - S([\text{-licensor, +gap}]))$.

This metric computes the surprisal difference between the two kinds of sentences that are ungrammatical ([+licensor, -gap] and [-licensor, +gap]) and the two kinds of sentences that are grammatical ([+licensor, +gap] and [-licensor, -gap]). When this metric is positive, we can conclude that the model reflects some understanding of filler-gap dependencies because it finds ungrammatical sentences as a group more surprising than grammatical ones. A model can score high on this metric by knowing the bijectivity of filler-gap dependencies, i.e. if a sentence has a gap it should have a licensor *and* if a sentence has a licensor it should have a gap. However, a model can achieve a large positive score on this metric even if it only encodes one direction of the bijectivity. For instance, if the presence of a licensor reduces the surprisal of a sentence with a gap, but has no impact on a sentence without a gap, the formula will indicate that the model has learned filler-gap dependencies even though it has learned only a single direction of the dependence. In search for stronger evidence, we propose two criteria: (8) and (9).

(8) **Does surprisal "flip"?**
Is the surprisal higher for [+licensor] than [-licensor] when [-gap] and is it lower for [+licensor] than [-licensor] when [+gap]?

(9) **Does surprisal "divide" by grammaticality?**
Within the four variants of a filler-gap de-

pendency (e.g. (5)), do grammatical variants have lower surprisals than their ungrammatical counterparts?

A flip in surprisal (8) is stronger than a high *wh*-licensing interaction because the former implies the latter but not the other way around. To see this, consider the previous scenario where *wh*-licensing interaction is high despite the LM not having learned bijectivity. Then, surprisal is not higher for [+licensor] when [-gap], so there is no flip. Surprisal will flip only if LMs learn bijectivity.

A division in surprisal by grammaticality (9) is even more demanding than a flip, but it is a reasonable criterion given that probabilistic measures correlate with acceptability judgments (Lau et al., 2014, 2015, 2017). The method of comparing probabilities within minimal pairs has also driven much other work in the assessment of neural networks' understanding of syntax (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018).

For our first study, we assess Wilcox et al.'s LMs' acquisition of filler-gap dependencies by checking for flips (8) and divisions (9) on three kinds of probabilistic metrics calculated from the data from their first experiment, which we describe now.

3.1 Metrics

Local surprisal Wilcox et al. (2018) always measures local surprisal on the post-gap region regardless of [gap]. This means local surprisal for a [-gap] sentence is measured after the filled gap, e.g. in a [-gap] variant of (7-a), the measurement takes place at *at sunrise* rather than *a gazelle*. However, a spike in surprisal due to illicit filled gaps might occur at the filled gap rather than at the post-gap region (Roger Levy, p.c.). Taking this possibility into account, local surprisal is measured at the filled gap for [-gap] sentences in later work such as (Wilcox et al., 2019b). We follow this practice and measure local surprisal at different regions depending on [gap]. Note that we can no longer check for divisions by grammaticality with local surprisal as we cannot compare surprisals between [+gap] and [-gap] conditions, since [gap] perfectly confounds with region. Nevertheless, this allows us to check for surprisal flips, which only depends on comparisons within [+gap] and [-gap].

Global surprisal We follow Wilcox et al. (2018) in measuring global surprisal. We normalize it by region length, which is otherwise an obvious confound – the embedded clause in [+gap] sentences is shorter and thus likely less surprising than [-gap] sentences.

SLOR The syntactic log-odds ratio (SLOR, Pauls and Klein, 2012) for a sentence s is sentence probability normalized for word frequency and word count, and has been shown to positively correlate with human acceptability judgments (Lau et al., 2017).

We train two unigram models on the training sets for the Google model and the Gulordava model respectively with add-one smoothing, and use the unigram model that matches the LM in our calculation of SLOR.

3.2 Experiments

We use mixed-effects models in all analyses. To check if the metrics flip, we predict the metrics with a fixed effect of [+licensor] on [-gap] and [+gap] sentences separately. To check if the metrics divide by grammaticality (9), we predict the metrics with a fixed effect of the grammaticality variable [+gram], defined as [gram] = NOT ([licensor] XOR [gap]), on all data.[2] We always include a random intercept by sentence, not by variant, i.e. all variants in (5) count as the same sentence.

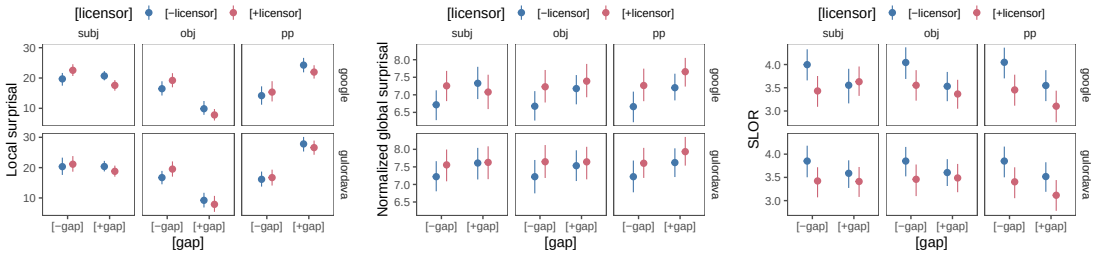
In the plots, points indicate means and error bars indicate 95% confidence intervals thereof computed by non-parametric bootstrapping.

We analyze the data from Wilcox et al. (2018)'s first experiment, which shows that both LMs show positive though different *wh*-licensing interactions for sentences each containing either a subject, object or a prepositional object (PP) gap, indicating that they learn that gaps may occur in all three syntactic positions.

3.3 Local surprisal

Figure 1a shows local surprisal. We see flips in all conditions with significant differences between the [+/-licensor] sentences ($p < 0.05$). This allows us to make a stronger statement about the LMs' acquisition of filler-gap dependencies than Wilcox et al. (2018), whose *wh*-licensing interaction metric can only lead them to conclude that these LMs learn that having a licensor has a different effect on surprisal depending on [gap].

[2] [+gram] iff either [+licensor, +gap] or [-licensor, -gap], i.e. iff the sentence is grammatical insofar as filler-gap dependencies are concerned.



(a) Local surprisal. (b) Normalized global surprisal. (c) SLOR.

Figure 1: Local, normalized global surprisals and SLOR for sentences containing subject, object and PP gaps.

3.4 Global surprisal

Figure 1b shows global surprisal. Flips are only observed for the Google model with subject gaps ([+gap]: $\beta = -0.24, p < 0.05$, [-gap]: $\beta = 0.53, p < 0.05$). Elsewhere, [-licensor] is less surprising than [+licensor], and [-gap] less surprising than [+gap].

As for divisions, grammatical sentences are less surprising than ungrammatical sentences in all conditions, but this difference is significant only for subject gaps for both models (Google: $\beta = -0.39, p < 0.05$, Gulordava: $\beta = -0.16, p < 0.05$).

3.5 SLOR

While local and global surprisal are negative log-likelihoods, SLOR is based on positive log-likelihood. Thus, we expect grammatical sentences to have higher SLORs than ungrammatical sentences.

Figure 1c shows SLOR. As with global surprisal, we see SLOR flip only for the Google model with subject gaps ([-gap]: $\beta = 0.08, p < 0.05$, [+gap]: $\beta = -0.56, p < 0.05$). In the remaining conditions, SLOR is higher for [-licensor] than [+licensor] and higher for [-gap] than [+gap].

Grammatical sentences have higher SLOR than ungrammatical sentences in all conditions, although the difference is significant only for subject and object gaps for both models ($p < 0.05$).

Contrary to local surprisal, both global surprisal and SLOR display flips and divisions in very few conditions. We tend to observe flips and divisions, if at all, most often in subject gaps, then in object gaps, and least often in PP/goal gaps. This trend is consistent with Wilcox et al. (2018)'s results with *wh*-licensing interaction.

3.6 Summary of the metrics

Why does local surprisal give us the most optimistic assessment of filler-gap dependency acquisition? We note that global surprisal and SLOR suffer from more confounds and thus may reflect grammaticality less purely than local surprisal. For example, the impact of word frequency as a confound on global surprisal and SLOR is greater than that on local surprisal; grammatical combinations of infrequent words can be more surprising than ungrammatical combinations of frequent words, violating the division criterion (9). [3] SLOR attempts to correct sentence probability by unigram frequency but ignores higher order effects, e.g. the [-licensor] bigram *know that* and the [+licensor] bigram *know who/what/where* can have different frequencies, which can correlate with extraction availability (Liu et al., 2019; Richter and Chaves, 2020), and affect the conditional probabilities of all words that follow in an autoregressive model such as LSTMs, potentially drastically affecting sentence-level probability.

4 Study 2: Other constructions

How well do LMs learn other kinds of filler-gap dependency constructions? We generated six sub-datasets that contain five kinds of filler-gap dependencies: `comp-quant` for comparatives (2-a), `cleft-adj` and `cleft-noun` for *it*-clefts (2-b), `topic` for topicalization (2-b),

[3] A reviewer has pointed out that the three metrics are all confounded by word frequency. This is correct, as surprisal measured at any word is influenced by the frequency of said word as well as all words in its context. However, we believe the impact of this confound on global surprisal and SLOR is greater than that on local surprisal simply because the region at which local surprisal is measured is strictly contained by the region at which global surprisal is measured, which in turn is strictly contained by the region at which SLOR is measured, i.e. the entire sentence. Semantic factors can also affect these metrics in a similar way. Metrics that involve surprisals from more words are more heavily impacted by such confounds.

`embwhq` for embedded *wh*-questions (2-d) and `tough` for *tough*-movements (2-e). Each subdataset contains 1200 sentences, or 4800 variants. Like Wilcox et al. (2018), every sentence has four variants generated from combinations of [gap] and [licensor]. For independent reasons, we generate two datasets for *it*-clefts. We describe dataset construction in more detail in Appendix A.

The gap is always an object gap. The gap-containing clause in each sentence may be separated by 0 – 3 levels of **embedding**, of which there are three types: bridges (10-a), complex NP objects (10-b) and interrogatives (10-c). The latter two types of embeddings induce island effects – namely violations of the Complex NP Constraint and the *Wh*-island Constraint (Ross, 1967) – so each sentence is also specified for **islandhood**; a sentence is an island iff it contains a complex NP object or an interrogative embedding.

(10) a. It was Alex$_i$ that **I think that** you met ___$_i$.
b. *It was Alex$_i$ that **I believed his claim that** you met ___$_i$.
c. *It was Alex$_i$ that **I wondered if** you met ___$_i$.

We then assess the LMs' acquisition of these constructions from different perspectives. As we shall see, the LMs learn different constructions to different extents. Embedded *wh*-questions are learned best across the board and topicalizations are learned the worst. The LMs show mixed acquisition for the remaining constructions, with clefts and comparatives learned generally better than *tough*-movement.

We fit mixed-effects models to predict one of the three metrics with a random intercept by sentence for all analyses. We will describe the fixed effect structure for each analysis. To visualize the modeling results and the inferences they license, we plot model effect estimates along with error bars indicating 95% confidence intervals on those estimates.

4.1 Licensor-gap interaction

We first focus on sentences with no embeddings and look at how [licensor] and [gap] affect surprisal and SLOR. For each combination of construction type and LM, we fit for a fixed effect of [+licensor], [+gap] and their interaction. We are specifically interested in the interaction term. A significant licensor-gap interaction that points in the direction of higher probability, i.e. lower surprisal and higher SLOR, means a LM has learned that [+licensor] has a better effect on surprisal / SLOR when [+gap] than when [-gap]. This way of assessing the acquisition of filler-gap dependencies is roughly the same as looking at Wilcox et al. (2018)'s *wh*-licensing interaction, which they obtain by direct calculation from the data instead of from a statistical model.

Figure 2 shows the licensor-gap interactions. For embedded *wh*-questions, the interaction is always significant in the direction of higher probability for both LMs in all three metrics. The Google model shows a highly significant positive interaction for clefts, comparatives and *tough*-movements for all three metrics. However, the Gulordava model shows a significant interaction for clefts and comparatives only for global surprisal and SLOR, and never for *tough*-movements. Both models showed the least understanding of topicalization, here the expected positive interaction was often significantly negative indicating that licensors increased the surprisal of sentences with gaps.

4.2 Flips

Next, we check how often flips occur on the constructions. We first look at sentences with no embeddings. Figure 3 shows the [+licensor] effects. We see that both LMs show flips for embedded *wh*-questions in all three metrics. Clefts flip only in local surprisal for Google and in global surprisal for both LMs. Comparatives only flip in global surprisals for both LMs. No metrics flip for topicalization and *tough*-movement in any condition.

We then look at sentences with one embedding each. The data thus consist of islands and non-islands. Islandhood affects filler-gap dependencies, which are [+gap, +licensor], but not [-gap, -licensor] sentences. [4] We consider the interac-

[4] We expect probabilistic outputs from a human-like LM to be affected by islandhood for [+gap, +licensor] sentences, not for [-gap, +licensor] sentences. This expectation comes from the acceptabilities of these two types of sentences, as illustrated in (i).

(i) I know { that / *who } you believe [the idea that she beat him in the election].

In contrast, Wilcox et al. (2021) expect islandhood to affect surprisals in both [-gap, +licensor] sentences as well as [+gap, +licensor] sentences. This is in line with a view on human sentence processing that humans do not expect a gap in an island, so the filled gap in (i) is equally surprising with or without an upstream licensor (Fodor, 1983; Freedman and Forster, 1985; Stowe, 1986). There is a rich body of literature concerning the debate on the question of how the human parsing mechanism interacts with grammatical constraints such as island constraints, and we believe it would be an interesting research

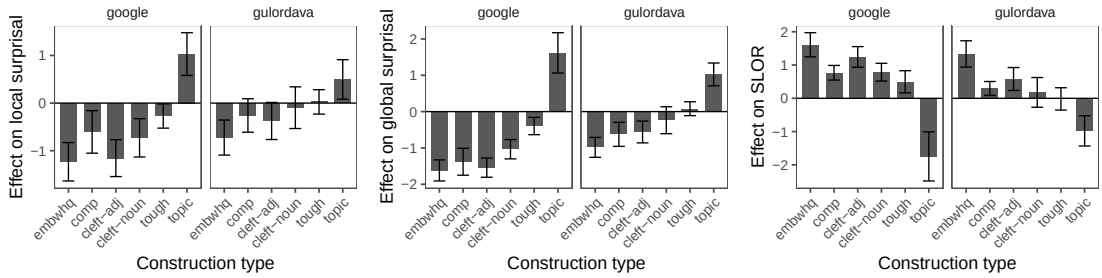


Figure 2: Licensor-gap interaction.

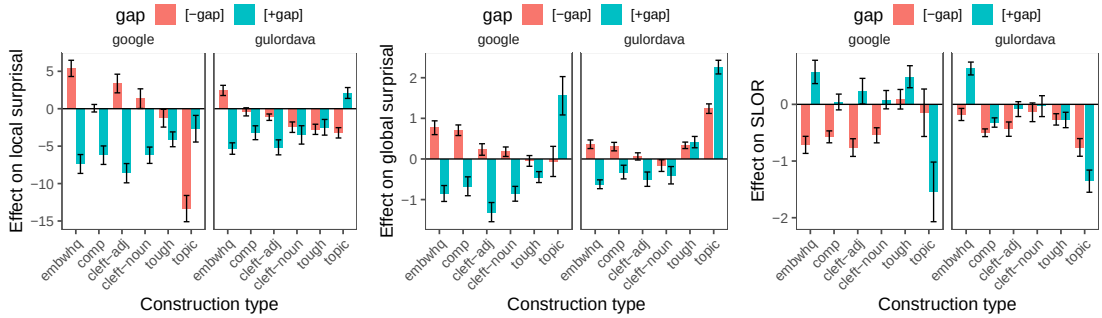


Figure 3: [+licensor] effects on [-gap] and [+gap] sentences.

tion between [licensor] and islandhood for [-gap] and [+gap] sentences separately. Figure 4 shows the [-gap, -licensor, +island] and [+gap, +licensor, +island] effects. When there is neither a licensor or a gap, islandhood does not affect surprisal or SLOR for any of the constructions or either LM. When both are present, islandhood hurts surprisal and SLOR for embedded *wh*-questions for both LMs. Islandhood hurts all three metrics for clefts and comparatives as well except SLOR from the Gulordava model. For *tough*-movement, islandhood does not affect SLOR, but it does associate with lower local surprisal for both LMs and lower global surprisal for the Google model. Islandhood never affects topicalization. From these licensor-islandhood interactions we can conclude that the LMs are learn island constraints to different extents for different constructions.

4.3 Divisions

Finally, we check how often divisions by grammaticality occur on the constructions. We first look at sentences with no embeddings. Figures 5a, 5b show the grammaticality effects. For both LMs, grammatical clefts, comparatives and embedded *wh*-questions have lower global surprisal than their ungrammatical counterparts, whereas for topicalization the grammatical variants are more surprising. Grammatical *tough*-movement is less surprising for Google ($\beta = -0.20, p < 0.05$) but more surprising for Gulordava ($\beta = 0.04, p = 0.55$).

Grammatical clefts, comparatives and embedded *wh*-questions also have higher SLOR, but for Gulordava the difference is non-significant for clefts ($\beta = 0.05, p = 0.48$) and comparatives ($\beta = 0.09, p = 0.24$). Topicalization again is more surprising when grammatical. Grammatical *tough*-movement has slightly higher SLOR for Google ($\beta = 0.20, p = 0.07$) but a non-significant difference for Gulordava ($\beta = -0.005, p = 0.9$).

We then look at all non-island sentences, and consider the interaction between grammaticality and the number of embeddings. Figures 5c, 5d show the interaction effects between grammaticality and the number of embeddings. Increasing number of embeddings is associated with higher global surprisal in all grammatical constructions except topicalization, with non-significant effects in comparatives and *tough*-movement for Gulordava. It is also associated with lower SLOR in the same constructions with non-significant effects in *tough*-movement for both LMs, clefts and comparatives for Gulordava. These patterns suggest that filler-gap dependencies that extend over multiple embeddings are harder for the LMs to process. However,

direction to conduct systematic comparisons between LM and human behaviors in the context of this question (e.g. Wilcox et al. (2019a)).

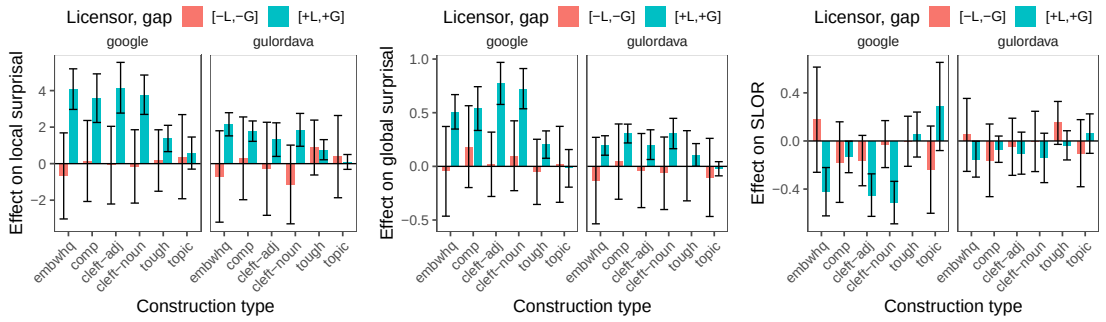
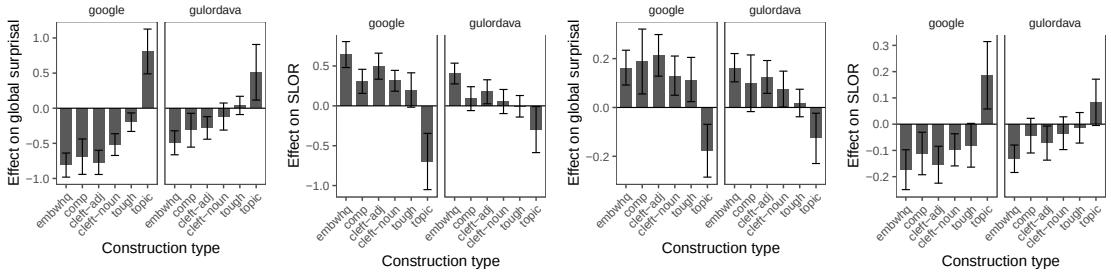


Figure 4: [+island] effects for [-gap, -licensor] and [+gap, +licensor] sentences.



(a) Normalized global surprisal, no embeddings.

(b) SLOR, no embeddings.

(c) Normalized global surprisal, non-islands.

(d) SLOR, non-islands.

Figure 5: Interaction effects between grammaticality and the number of embeddings, fit first on sentences without embeddings then on all non-island sentences.

topicalization has lower surprisal and higher SLOR with more embeddings.

4.4 Summary

We looked at five kinds of filler-gap dependency constructions, and found that the LMs learn different constructions to different extents with respect to licensor-gap interaction, flips, licensor-islandhood interaction, division by grammaticality and the interaction between grammaticality and the number of embeddings. Roughly, the constructions seem to be better learned in the decreasing order of embedded *wh*-questions, clefts and comparatives, *tough*-movement and topicalization.

5 Study 3: Acquisition and frequency

Why do the LMs learn the five filler-gap dependency constructions to different extents? One simple hypothesis is that LMs learn frequent syntactic phenomena better than rare ones (Zhang et al., 2020). To test this, we searched for occurrences of our five constructions in the Brown corpus and the Wall Street Journal corpus from Penn Treebank 3.0 (Marcus et al., 1993) using Tregex (Levy and Andrew, 2006). This gives us an estimate of the relative frequencies of the constructions in a typical written-text corpus, which is what the two LMs were trained on. We choose our licensor-gap interaction from Section 4 to be a quantitative measure of the LMs' acquisition of the constructions. We look for a correlation between licensor-gap interaction and the relative frequency of the constructions; the results are shown in Table 1. We provide the relative frequencies of the constructions and the Tregex scripts we used to search for the constructions in Appendix B.

LM	Metric	r (Brown)	r (WSJ)
Google	global	-0.20	-0.67
	local	-0.32	-0.75
	slor	0.13	0.65
Gulordava	global	-0.32	-0.73
	local	-0.52	-0.82
	slor	0.52	0.86

Table 1: Pearson's r between the licensor-gap interaction and frequency of the filler-gap dependency constructions.

Significance testing is not performed due to the lack of data – there are only five kinds of constructions. The correlation between licensor-gap interaction for the Gulordava model and frequency

in the WSJ corpus seems strong, potentially due to a domain similarity between English Wikipedia, which Gulordava was trained on, and WSJ, which consists solely of newspaper articles. The Brown corpus however covers a wider range of older texts. Overall, acquisition of a construction seems to be correlated with its frequency, but more constructions need to be tested in order for the correlation to be non-anecdotal and for this conclusion to be supported.

6 Discussion

We have shown that Wilcox et al. (2018)'s LSTM LMs learn the bijectivity of certain English filler-gap dependency constructions. For embedded *wh*-questions, gaps are less surprising with a licensor than without, and filled gaps are more surprising with a licensor than without. However, this sign of acquisition is stronger for local surprisal than for global surprisal and SLOR. Compared to local surprisal, global surprisal and SLOR are more heavily impacted by confounds such as sentence length and word frequency, which makes the latter two unfair metrics for assessing LMs' syntactic understanding – probability is not all about grammaticality.

As has been correctly pointed out by two reviewers, the connections between probability, categorical grammaticality and gradient acceptability are not innocent. While probability seems to be correlated with acceptability for sentences constructed with round-trip translation, it seems less so with grammaticality for sentences constructed by linguists (Lau et al., 2017; Sprouse et al., 2018). This suggests that probability is a good indicator of unacceptability caused by coarse lexical and syntactic errors introduced by machine translation, but it cannot be used to distinguish between linguist-constructed minimal pairs that often vary very subtly in surface structure. The experimental items in our study are also constructed from a linguistic standpoint. With this in mind, the failure of global and SLOR to indicate correspondence with grammaticality supports the present claim in the literature.

We also see that the LMs learn different filler-gap dependency constructions to different extents, in terms of licensor-gap interaction, flips and divisions, as well as islandhood and the number of embeddings. Moreover, the more frequent a construction is in written English, the more the licensor-gap interaction improves the probability of a filler-gap dependency. While this does not tell us much about what the neural networks have learned, this is a human-like behavior in that frequency affects human language acquisition (Ambridge et al., 2015) and sentence processing (Ellis, 2002).

A systematic investigation with a wider coverage of filler-gap dependency constructions is in order. In this study, we were able to adopt Wilcox et al. (2018)'s 2x2 design because for each of our five constructions, we could either take the filler to be the licensor, or find a construction with a minimal surface difference that does not license gaps, and take that to be our [-licensor] variant. For example, we construct [-licensor] variants of comparatives (... *than* ...) by turning them into coordinate structures (... *and* ...). In other filler-gap dependency constructions, this is much more challenging. For example, infinitival relative clauses (11) are filler-gap dependencies, but it is not obvious how to construct [-licensor] variants for them.

(11) a. Here are **some options**$_i$ for you to choose from ___$_i$.
b. She was the first **person**$_i$ ___$_i$ to point out the mistake.

The experimental paradigm needs to be revised in a way to cover such constructions as well. In future research, we wish to collect human acceptability judgments for our data, and compare our results with the probabilistic outputs from LMs to check the connection between probability and acceptability.

We provide our data and code at `https://github.com/ikazos/scil2022-fgd`. We also thank Roger Levy and a reviewer for pointing out SyntaxGym (Gauthier et al., 2020) to us, which we plan to contribute our data to in the near future.

References

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Jean-Philippe Bernardy and Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*.

Rui Chaves. 2020. What don't RNN language models learn about filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 1–11, New York, New York. Association for Computational Linguistics.

Rui P. Chaves and Michael T. Putnam. 2021. *Unbounded Dependency Constructions: Theoretical and Experimental Perspectives*, volume 10. Oxford University Press, USA.

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. *arXiv preprint arXiv:1312.3005*.

Noam Chomsky. 1977. On Wh-Movement. In Peter W. Culicover, Thomas Wasow, Adrian Akmajian, et al., editors, *Formal Syntax*, pages 71–132.

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.

Jillian Da Costa and Rui Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 12–21, New York, New York. Association for Computational Linguistics.

Nick C Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2):143–188.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Janet Dean Fodor. 1983. Phrase structure parsing and the island constraints. *Linguistics and Philosophy*, 6(2):163–223.

Sandra E Freedman and Kenneth I Forster. 1985. The psychological status of overgenerated sentences. *Cognition*, 19(2):101–131.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. Comlex syntax: Building a computational lexicon. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of lstms to understand negative polarity items. *arXiv preprint arXiv:1808.10627*.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. Lstms can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628, Beijing, China. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

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

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: Tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

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.

Yingtong Liu, Rachel Ryskin, Richard Futrell, and Edward Gibson. 2019. Verb frequency explains the unacceptability of factive and manner-of-speaking islands in english. In *CogSci*, pages 685–691.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Stephanie Richter and Rui Chaves. 2020. Investigating the role of verb frequency in factive and manner-of-speaking islands. In *CogSci*.

John Robert Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, Cambridge, MA.

Ivan A. Sag. 2010. English filler-gap constructions. *Language*, 86(3):486–545.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

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

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ethan Wilcox, Richard Futrell, and Roger Levy. 2021. Using computational models to test syntactic learnability. *lingbuzz/006327*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190, Florence, Italy. Association for Computational Linguistics.

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

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020. When do you need billions of words of pretraining data?

A Dataset construction

In this section, we discuss how we constructed the [-licensor] and [-gap] variants of the 5 filler-gap dependency constructions.

[-gap] variants are always created by filling the gap with the filler. The creation of [-licensor] differs over constructions.

A.1 Comparatives

We look at comparative constructions with a comparative quantifier *more* modifying an object NP with *than* leading a subordinating clause containing the gap. We consider the filler to be *a lot of* + matrix object NP. For the [-licensor] variant, we replace *more* with *a lot of* and replace *than* with *and*, giving us a coordinate structure that does not license a gap.

(12) a. [+licensor, +gap]
Mary bought **more books** this month **than** John bought ___ last month.

b. [+licensor, -gap]
*Mary bought **more books** this month **than** John bought **a lot of books** last month.

c. [-licensor, +gap]
*Mary bought **a lot of books** this month **and** John bought ___ last month.

d. [-licensor, -gap]
Mary bought **a lot of books** this month **and** John bought **a lot of books** last month.

A.2 Clefts

We look at clefts with object gaps of the structure *It was ... that ....* The two subdatasets

`cleft-adj` and `cleft-noun` are created with different strategies for creating the [-licensor] variants. In `cleft-adj`, the [-licensor] variants are created by replacing the filler with an adjective that take an extraposed sentential subject, e.g. *apparent*. In `cleft-noun`, the [-licensor] variants are created by replacing the filler with a noun that take an extraposed sentential subject, e.g. *a fact*. We collected a list of such adjectives and nouns from COMLEX Syntax (Grishman et al., 1994).

(13) a. [+licensor, +gap]
It was **books** that Mary bought ___ last month.
b. [+licensor, -gap]
*It was **books** that Mary bought **books** last month.
c. [-licensor, +gap] (`cleft-adj`)
*It was **apparent** that Mary bought ___ last month.
d. [-licensor, +gap] (`cleft-noun`)
*It was **a fact** that Mary bought ___ last month.
e. [-licensor, -gap] (`cleft-adj`)
It was **apparent** that Mary bought **books** last month.
f. [-licensor, -gap] (`cleft-noun`)
It was **a fact** that Mary bought **books** last month.

A.3 Embedded *wh*-questions

We look at embedded *wh*-questions with object gaps. The matrix verb selects for either a sentential or an interrogative complement; we gathered a list of such verbs from VerbNet (Schuler, 2005) and from Wilcox et al. (2018)'s data. Following Wilcox et al. (2018), we replace the *wh*-phrase leading the interrogative complement with *that* for the [-licensor] variants.

(14) a. [+licensor, +gap]
Clara knows **what** Mary bought ___ last month.
b. [+licensor, -gap]
*Clara knows **what** Mary bought **books** last month.
c. [-licensor, +gap]
*Clara knows **that** Mary bought ___ last month.
d. [-licensor, -gap]
Clara knows **that** Mary bought **books** last month.

A.4 Topicalization

We look at topicalization (also known as complement preposing (Huddleston and Pullum, 2002)) with object gaps. Unlike the other constructions, topicalization does not allow subject gaps – this is one of the reasons why we exclusively generate object gaps throughout all constructions. The filler is always a definite NP, which helps with a focus interpretation. For the [-licensor] variants, we simply delete the filler and the comma. For the [-gap] variants, we fill the gap with the filler directly instead of e.g. a referential pronoun, because that would give us left-dislocation for [+licensor, -gap], which is a grammatical construction.

(15) a. [+licensor, +gap]
These books, Mary bought ___ last month.
b. [+licensor, -gap]
***These books**, Mary bought **these books** last month.
c. [-licensor, +gap]
*Mary bought ___ last month.
d. [-licensor, -gap]
Mary bought **these books** last month.

A.5 *Tough*-movement

We look at *tough*-movement with object gaps. We select matrix adjectives that license hollow *to*-infinitivals (Huddleston and Pullum, 2002). For the [-licensor] variants, we replace the filler with *it*.

(16) a. [+licensor, +gap]
These books are impossible to finish ___ in a day.
b. [+licensor, -gap]
***These books** are impossible to finish **these books** in a day.
c. [-licensor, +gap]
***It** is impossible to finish ___ in a day.
d. [-licensor, -gap]
It is impossible to finish **these books** in a day.

B Data and Tregex scripts for Study 3

The relative frequencies for each construction type in the Brown corpus and the WSJ corpus are listed in Table 2. Here are the Tregex scripts used to search for the occurrences for each construction.

B.1 Clefts

This covers both `cleft-adj` and `cleft-noun`. The script is: `S-CLF`

Construction	Freq (in Brown)	Freq (in WSJ)
Clefts	108	65
comp-quant	9	41
embwhq	280	146
topic	119	14
tough	36	79

Table 2: Relative frequency for each construction type in the Brown corpus and the WSJ corpus.

B.2 Comparatives

This covers comp-quant. The script is: `@NP << more & !<< (@ADJP << more) < (PP|SBAR < (__ < than) & < @S)`

B.3 Embedded *wh*-questions

This covers embwhq. The script is: `VP < (SBAR < (/WH*/ << what|who))`

B.4 Topicalization

This covers topic. We first look for occurrences of `/NP-TPC-?/ !<< ''`, then subtract the number of occurrences of `'' $+ (/NP-TPC-?/ !<< '')` to rule out false positives.

B.5 *Tough*-movement

This covers tough. The script is: `ADJP-PRD < (SBAR < /WHNP-*/)`.