Using Computational Models to Test Syntactic Learnability

Ethan Gotlieb Wilcox, Richard Futrell and Roger Levy

November 2021

Abstract

We study the learnability of English filler—gap dependencies and the "island" constraints on them by assessing the generalizations made by autoregressive (incremental) language models that use deep learning to predict the next word given preceding context. Using factorial tests inspired by experimental psycholinguistics, we find that models acquire not only the basic contingency between fillers and gaps, but also the unboundedness and hierarchical constraints implicated in the dependency. We evaluate a model's acquisition of island constraints by demonstrating that its expectation for a filler—gap contingency is attenuated within an island environment. Our results provide empirical evidence against the Argument from the Poverty of the Stimulus for this particular structure.

Keywords: Syntactic Islands, Neural Networks, Learnability, Deep Learning, Filler–Gap Dependency

1 Introduction

The English filler–gap dependency is the co-variation between a wh-word or phrase (a filler) and an empty syntactic position (a gap). It is special in that it can span over a potentially unbounded number of nodes in a syntactic tree, yet it is subject to a subtle set of constraints known as *island constraints* (Ross, 1967). For example, in the grammatical sentence in (1-a), the dependency between the filler and the gap spans two sentential embeddings. However, a similar sentence, (1-b), is rendered ungrammatical when the gap site resides within a syntactic 'island', in this case a Complex Noun Phrase.

(1) a. I know what the guide said his friend saw the lion devour ___ last night.
    b. *I know what the guide saw the lion that devoured ___ last night.

A successful theory of the filler–gap dependency and its associated constraints must deal with two interrelated facts: First, despite some inter-language variability, the same set of structures arise as syntactic islands in language after language. Second, despite noisy and primarily negative evidence from caregivers, children within an individual language community tend to coordinate on the same set of islands. This second issue, that of learnability, will be the main focus of this paper.

There are two approaches around which theories about syntactic islands have developed—linguistic nativism and empiricism. Nativism refers to the hypothesis that innate, language-specific constraints aid the child language-learner by reducing their hypothesis space during the acquisition process. Under this approach the ungrammaticality of (1-b) in adult grammars arises from the fact that learners never hypothesize Complex NPs such as `[`$_{NP}$ `the lion [`$_{CP}$ `that ...]]` as host sites for gaps when they are learning the syntax of their language. Historically playing a contrastive role to the nativist theories are a cluster of theories—often referred to as functionalist or empiricist—which posit that language structure is an emergent property of language use. Central to this approach is the hypothesis that some aspects of syntax are acquired due to domain general (i.e. non-linguistic) cognitive learning abilities (Clark and Lappin, 2010). Contributions to this debate have been made in a number of ways: There are language learning experiments in adults and children (Otsu, 1981; Goodluck et al., 1992; Culbertson et al., 2012); cross linguistic analysis of phenomena (Richards, 2001); analytic arguments about learnability (Gold, 1967; Chomsky, 1975); corpus analyses (Pullum and Scholz, 2002); and statistical and computational modeling (Perfors et al., 2006; Pearl and Sprouse, 2013a). Here, we gain traction on this question through computational modeling. Our approach is to study what syntactic generalizations are acquired by state-of-the-art algorithms developed to process natural language text and trained on a childhood's worth of data or more. By asking what can be acquired by data-driven learning algorithms without any obvious domain-specific biases, we map out a lower-bound for learnability and clarify what (if anything) remains as a good candidate for top-down innate constraints.

Our learning algorithms are all Language Models (LMs), models that define joint probabilities over word sequences (Chen and Goodman, 1999). We present learning outcomes for a variety of LMs that have two key properties: First, they are *domain general* (Clark and Lappin, 2010) insofar as their architecture does not limit them to learning generalizations about human languages, and have been successfully deployed to model data as disparate as handwriting recognition, the stock market (Graves and Schmidhuber, 2009), and stop sign detection (Arcos-García et al., 2018). Second, they are *weakly biased* (Lappin and Shieber, 2007) insofar as their initial states are chosen at random, their inputs are all vectors of uniform length, and their internal states consist of large matrices which are capable of approximating arbitrary functions. In order to demonstrate acquisition of the filler–gap dependency, we present a general computational method that could be deployed to test the learnability of many syntactic structures. This method follows the approach of Elman (1990), whose basic insight is to treat statistical models of language like human subjects in a psycholinguistics study. Following Elman (and more recently Linzen et al. (2016) and Futrell et al. (2018)), we inspect the word-by-word predictions of language models. By constructing test suites that mimic classic psycholinguistic experiments used to assess human processing of filler–gap dependencies, we can gain insight into what sorts of syntactic generalizations models make during training, even for models whose internal states are opaque, such as contemporary neural networks. This method can be used to test the learning outcomes of any model that makes incremental predictions over strings, and for a vast array of syntactic, morphological and even semantic rules.

The rest of the paper will proceed as follows: In Section 2 we outline previous computational models of syntactic island acquisition. In Section 3 we present the general methods employed, and introduce our three classes of learning algorithms: Simple n-gram models, which serve as a baseline, and two machine learning architectures, Recurrent Neural Networks and Transformers. In section 4 and 5 we present empirical findings demonstrating that the models tested can learn many aspects of the filler–gap dependency, including its hierarchical restrictions, potential unboundedness, as well as most of the related island constraints. Section 6 discusses implications from our modeling results, including for the most influential analytic argument for linguistic nativism, the Argument from the Poverty of the Stimulus (Chomsky, 1957). Section 7 concludes.

2 Background: Modeling Syntactic Islands

Island constraints were first introduced by Ross (1967), who analyzed them as a grammatical phenomena, specifically barriers to syntactic movement ("islands" being places from which it is difficult to move).[1] Since, they have "been regarded as excellent candidates for innate domain-specific constraints [because] they are hard to learn, and they appear to apply in similar ways across languages." (Phillips, 2013). Indeed, there are about a dozen different island constraints in English and they target a seemingly arbitrary set of syntactic structures: In addition to Complex NPs mentioned above, the islands discussed in this paper will include Adjunct Islands, Subject Islands, Sentential Subject Islands, Coordination Islands, Left Branch Islands and Wh-Islands. Within syntactic theory, many of the most influential proposals for innate, universal constraints have been developed to explain why it is this set of structures that block islands, and not others (Chomsky, 1964, 1973; Huang, 1982; Chomsky, 1986). For example, Chomsky (2005) outlines five candidate constraints that were, in his view, the most successful attempts to formulate various aspects of Universal Grammar from 1955-1977: "the A-over-A Principle, conditions on wh-extraction from wh-phrases (relatives and interrogatives), simplification of T-markers to base recursion [...] and cyclicity [...], later John Robert Ross's (1967) classic study of taxonomy of islands..." Of these, four (all but T-markers) were recruited to explain the distribution of the filler–gap dependency, a strong indication of its central role in Generative Linguistics.

Because islands have long been considered one prime example of linguistic behavior that could stem from innate, language-specific constraints, multiple researchers have asked whether their distribution can be learned by computational models. Below we present an overview of previous models of island acquisition. While these models represent a useful step forward both methodologically and empirically, we argue that no work to-date has demonstrated that the filler–gap dependency and island constraints can be acquired by a model that is both *domain general* and *weakly biased* (in the sense introduced above). We start

[1] There are a group of islands that arise due to *semantic* considerations (e.g. "How many miles didn't you run ___ ?") (Dayal, 2016), which we do not address. Furthermore, competing with the grammatical analysis are "reductionist" accounts (Sprouse and Hornstein, 2013) such as the processing account (Hofmeister and Sag, 2010; Hofmeister et al., 2013) and discourse-structural accounts (Ambridge and Goldberg, 2008; Ambridge et al., 2014). While these approaches are typically associated with empiricist positions, grammatical and reductionist accounts could be articulated from both empiricist and nativist perspectives.

with the model presented in Pearl and Sprouse (2013b). This model identifies islands by tracking the syntactic container nodes, which are all the phrasal category nodes that dominate a given gap site. An example of a syntactic parse and its container-node sequence is given in (2).

(2) $[_{CP}$Who did $[_{IP}$she $[_{VP}$think $[_{CP}$ $[_{IP}$ $[_{NP}$ the gift$]$ $[_{VP}$ was $[_{PP}$ from __$]]]]]]]$?
`start-IP-VP-CP-IP-VP-PP-end`

Pearl and Sprouse propose that children keep track of the linear order of container nodes, and estimate the relative probability of a gap by multiplying the probabilities of individual trigrams along the container node spine.[2] Using this procedure, they estimate these probabilities for gap sites from a portion of the CHILDES corpus, and demonstrate that they broadly track human judgements. Pearl and Sprouse argue that the components of their algorithm are either domain-general but innate (identify trigrams; calculate utterance probability) or domain-specific but learned (parse utterances; use container-node sequence).[3] Because arguments for linguistic nativism must rely on both domain-specific and innate constraints, they argue that their model provides evidence for a shift in the learnability status of island constraints.

There has been some debate, however, about whether this model really constitutes a domain-general learning algorithm (see, especially Phillips (2013)). First, the ontological status of its output is unclear: By comparing the probabilities assigned to sentences to human acceptability judgements, Pearl and Sprouse (2013b) imply that the model assigns a score something akin to grammaticality. However, the relationship between probability and grammaticality is not straightforward (Chomsky, 1957) and, as as pointed out in Phillips (2013), the model will fail to generalize properly to sentences which are grammatical but very unlikely, such as sententially-embedded gap sites of sufficient depth. In addition, the model requires that language learners be able to parse sentences into Penn-Treebank style parses in order to track container nodes. Their model, therefore, does not answer the question about whether islands can be learned via a domain-general algorithm, but rather whether island

[2] Under their model the relative probability of the gap in (2) is P(Start,IP,VP) * P(IP,VP,CP) ... P(VP,PP,end).

[3] See the careful discussion in Section 6 of Pearl and Sprouse (2013a) for a item-by-item analysis of the components of their model.

constraints can be learned by a parsing algorithm (or an algorithm that employs a parser) *without any further domain-specific stipulations*. The result of their study is a resounding *yes*, but instead of resolving the learnability status of islands, it shuttles the debate from island constraints to one particular parse schema, and whether or not it can be learned using only domain-general assumptions.

Other modeling approaches have used connectionist networks, as we do here. Chowdhury and Zamparelli (2018) and Chowdhury and Zamparelli (2019) assess how well the probabilities assigned to strings by a Reucrrent Neural Network language model compare to human acceptability judgements, including some island phenomena. They demonstrate that models assign less total-sentence probability to sentences that violate islands. However, they complicate the story by reporting similar total-level probability on yes-no question sentences that have no wh-movement and therefore cannot violate islands, concluding that the models have not learned generalized rules, but are sensitive to the "cumulative effect of increasing syntactic complexity, plus position." While these results are interesting and demand further investigation, we believe that the equation of whole-sentence probability and grammaticality has lead the authors to too-hasty a conclusion. Just because models show similar sentence-level probability for island-violating wh-extractions and non-island-violating yes-no questions, does not mean that the models have not made accurate generalizations about syntactic islands. It may be the case that models have learned the correct generalization for syntactic islands, and a different set of (perhaps erroneous) generalizations pertaining to polar questions. More fine-grained experimentation, which we present here, suggests that connectionist models can differentiate between different types of syntactic complexity and different orderings of tokens, much like human sentence processors.

3 Methods

In this section, we will present the paradigm which we will use to assess the learning outcomes of our models. Because our methodology is inspired by well-used psycholinguistic methods, we call this the "psycholinguistics paradigm" for model assessment. Below, we explain how we deploy it on filler–gap dependency sentences, introduce our five models, and address a short list of potential concerns.[4]

3.1 Psycholinguistic Assessment of Language Models

Our learning models are all incremental Language Models, which have been trained to produce the probability of a token x_i given its context $x_1 \dots x_{i-1}$ by providing a distribution over $P(x_i|x_1 \dots x_{i-1})$.[5] In order to uncover the models' learning outcomes via behavioral analysis, we follow a simple training/testing setup. Language models are trained on a large corpus, and then tested on a suite of test items, $\mathtt{t}_1 \dots \mathtt{t}_n$ designed to test a particular grammatical phenomena. Models assign probabilities to each word in the test sentences, and these probabilities are used to assess whether the model has successfully learned the phenomena. Test suite items have three crucial components: First, items in a test suite items are hand-crafted and carefully controlled to test for the same grammatical phenomena, and include different content words, so that model behavior can be abstract away from semantics. Second, instead of looking at whole-sentence probabilities we look at probabilities of *critical regions*. Third, we use balanced pairs of grammatical and ungrammatical sentences, and assess the model's relative probabilities between grammatical / ungrammatical conditions

[4] For a lucid introduction to neural network (a.k.a "deep learning") models of language, as well as an insightful overview into the decades-long relationship between generative and computational linguistics, see Pater (2019).

[5] One possible concern for addressing issues of learnability with language models is that maximizing the accuracy of conditional word predictions is a far cry from both parsing and assigning meaning to utterances. We believe the language-modeling objective function is compelling for three reasons. First, it is clear that humans are sensitive to the statistical regularities of their language and use them to make incremental predictions during language learning and online comprehension (Levy, 2008; Hale, 2001). Infants as young as eight months old pay attention to the statistical regularities in short spans of non-linguistic input (Saffran et al., 1996), and in adults, gaze fixation on a token during reading is correlated to its conditional probability (Smith and Levy, 2013; Wilcox et al., 2020). Second, selecting models that produce probabilities over sentence parses or joint distributions over parses and strings would require committing to some syntactic formalism during training. Because language models are supervised solely by surface strings, they can be deployed more effectively in studies where it is important to remain theoretically neutral. Finally, because a well-tuned language model is in principle compatible with multiple structural generalizations, some of them non-hierarchical, language modeling offers a more *distant* form of supervision for syntactic generalization than direct learning of grammatical structures or alternations themselves. Therefore, if models do display humanlike behavior, this is stronger evidence that the proper structural generalizations can be learned from the data.

critical regions, which itself is the same in both conditions.

To give a concrete example, let's adapt the case of subject-verb number agreement from Linzen et al. (2016). Do statistical language models learn that matrix verbs must agree in number with the head noun of the subject NP? In order to test this, we can look at the relative probabilities of the matrix verb *are* (our 'critical region') in the test sentences *The keys to the cabinet are on the table.* and $*$*The key to the cabinets are on the table.* If the model has learned the proper syntactic generalization then "are" should be more likely in the first sentence than in the second. While some studies using this methodology have elected to report accuracy scores (i.e. the proportion of the time the grammatical variant is more probable (Marvin and Linzen, 2018; Hu et al., 2020)), we examine differences in effect sizes between conditions. This means that models' predictions must not only be in the right direction, but they must be substantially different in magnitude. That is, a model cannot appear to be 'right' if it assigns a probability of .4999 to the ungrammatical continuation and .5001 to the grammatical continuation. Crucially, the claim advocated here is not that the model understands that the "The key ... are" is ungrammatical. Rather, the model's incremental predictions are consistent with a generalization that matches English speakers' grammatical generalizations. This is a particularly compelling test because the local bigram is much more likely in the ungrammatical condition ("cabinets are") than in the grammatical condition ("cabinet are"). In order for the model to perform correctly, superficial local *n*-gram frequencies must be overridden by a more global generalization that takes into account a larger linguistic context.

In practice, we will assess the human-likeness of models' predictions in terms of the difference between the *surprisals*, or negative log-probabilities, of matching content in minimally differing contexts (Hale, 2001; Levy, 2008), for which we provide a technical justification in Appendix A. The formulation in terms of surprisals also offers a link to psycholinguistic theory and data: surprisal theory accounts for many well-known sentence processing phenomena, and reading times and brain responses have generally been found to be linear in word surprisal (Smith and Levy, 2013; Frank et al., 2015; Goodkind and Bicknell, 2018; Heilbron et al., 2021). Although this is not the focus of the present contribution, we also believe this link also suggests direct human/model comparisons, such as those described in Van Schijndel and Linzen (2018) and Wilcox et al. (2021).

3.2 Measuring the Filler—Gap Dependency

In order to test filler–gap dependency sentences we use a 2x2 interaction design, inspired by two separate predictions made by the grammar. The first prediction stems from the fact that gaps require fillers to be properly licensed. If we take a sentences such as (3-a) below, which includes both a filler and a gap, and change the filler into a non-wh complementizer like "that", the sentence is rendered ungrammatical, as in (3-b). (In this and subsequent examples the underscores are for presentational purposes only, and are not included in test items.) Note, that this grammaticality contrast holds only if the matrix verb is obligatorily object taking: "I know that the lion ate yesterday" is grammatical.

(3) a. I know what the lion devoured ___ yesterday .
b. *I know that the lion devoured ___ yesterday .

In order to assess whether the models have learned this first prediction, we measure the surprisal in the adverbial phrase modifier and end-of-sentence punctuation which occurs after the matrix verb, such as "*yesterday*" in (3). If the models have learned that gaps are licensed only in the presence of an upstream filler, than the transition *devoured* → *yesterday* which skips over an overt NP object should be less surprising in the context of a filler than in its absence. That is, "yesterday" should be more surprising in (3-b) than in (3-a) and the difference in summed surprisal over that region—$S(b) - S(a)$—should be a large positive number. Because this prediction is about the effect of an upstream wh-word, we call this the WH-EFFECT in the +GAP condition. In Appendix A we present a technical argument for why this difference in surprisals is a natural measure for quantifying the strength of the structural filler–gap dependency in a language model.

The second prediction made by the grammar is that upstream fillers require downstream gaps. If we take the grammatical sentence (4-a) and replace the that-complementizer with a wh-complementizer as in 6, then the rule is violated and the sentence becomes ungrammatical.

(4) a. I know that the lion devoured the gazelle yesterday.
b. *I know what the lion devoured the gazelle yesterday.

If models have learned that fillers require gaps, then after encountering a filler near the start of the sentence, models should expect an empty argument structure position downstream, and the filled object "the gazelle" should be more surprising in 6 than in (4-a). This effect has been observed in humans and is called the filled-gap effect (Crain, 1985; Stowe, 1986). We call the difference in surprisal of the filled argument structure position the WH-EFFECT in the -GAP condition.[6]

Putting together these two predictions made, we derive a 2 × 2 contrast, a schematic of which is given in 1. We predict that there should be interaction between the presence of a filler and the presence of a gap, whereby the models should display supperadditively less surprisal when both are present than when just one is present. To test for this interaction, we run mixed-effects linear regression models, fit on raw surprisal values with sum-coded conditions and by-item random slopes (Barr et al., 2013), and look for significantly *negative* interaction terms.[7] In cases where we investigate the effect of a particular syntactic construction, such as a syntactic island, on the licensing interaction, we include the presence of that structure as a predictor in our model. Here we expect a *positive* coefficient on the three-way interaction term, as gaps should be more surprising inside islands. When presenting materials in this paper, we will give examples of the +filler/+gap variant, even though all four will have been produced to measure two wh-effects. For visualizations, we present the two wh-effects described above, as collapsing across all four conditions may obfuscate important model behavior.

3.3 Models Tested

We assess three types of model classes: *n*-grams, Recurrent Neural Networks and Transformers. As argued in Pearl and Sprouse (2013a), for a model to be relevant for debates about linguistic nativism it must be both *domain general* and *weakly biased*; we discuss each below:

[6] Strictly speaking, this behavioral trace is not necessitated by the grammar, but rather by grammatical constraints combined with statistical knowledge that an object is likely to be gapped for an upstream filler for which a gap has not already been encountered. Unlike in the +gap condition, where the sentence becomes ungrammatical at the word *yesterday*, in the -gap condition sentence only becomes ungrammatical at the period, when the expectation for gaps set up by the filler is not discharged. For an incremental statistical processing model, lack of wh-effects in the -gap condition does not mean that the model has failed to learn the proper generalization. However, strong wh-effects are evidence that the model has learned the generalization that gaps follow fillers and is expecting a gap in that location.

[7] An example of a basic `R` command: `lmer(surprisal ∼ wh * gap + (gap + wh | item-number))`

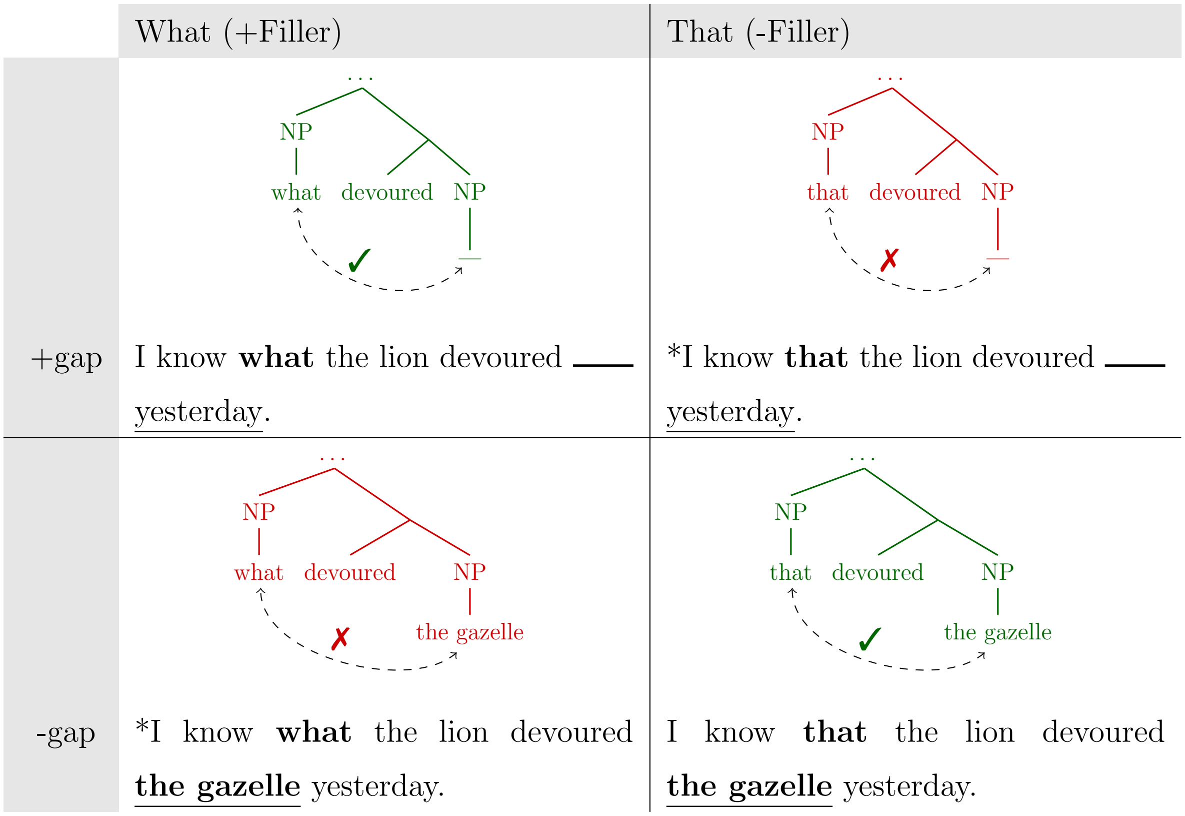
What (+Filler)
That (-Filler)
...
NP
what
devoured
NP
—
✓
...
NP
that
devoured
NP
—
✗
+gap
I know **what** the lion devoured ____ yesterday.
*I know **that** the lion devoured ____ yesterday.
...
NP
what
devoured
NP
the gazelle
✗
...
NP
that
devoured
NP
the gazelle
✓
-gap
*I know **what** the lion devoured **the gazelle** yesterday.
I know **that** the lion devoured **the gazelle** yesterday.

Figure 1: Schematic demonstrating our 2 × 2 interaction design for measuring the filler–gap dependency. The portion of the sentence in which we measure surprisal is underlined.

Following Clark and Lappin (2010), we take *domain general* to mean that the model's architecture does not limit it to making generalizations about just human language. Although we train and test our models on linguistic data, they are capable of representing relationships between arbitrary types of vectorized input, and therefore domain general learners. Following, Lappin and Shieber (2007), we take *weakly biased* models to be "uncomplicated, uniform [and] task-general" while *strongly biased* models are "highly articulated, non-uniform and task-specific." Because our neural networks' internal states are randomly initialized, their input data are all treated the same and their representations consist of large matrices of numbers, we take them as instances of weakly-biased as opposed to strongly-biased models. Consistent with this view, training and testing these architectures on synthetic languages has indicated that their inductive bias does not particularly favor natural language-like generalizations (Ravfogel et al., 2019; White and Cotterell, 2021).

***n*-gram models** are statistical models that assign a probability to a string by taking the product of probabilities of n-token substrings. An n-gram language model has no "memory" outside of its n-gram window, so models can only represent local dependencies between words. We use a 5-gram model with Knesser-Key smoothing trained on the British National Corpus using the SRILM language modeling toolkit (Stolcke, 2002). We present the n-gram model primarily as a baseline.

Artificial Neural Network (ANN) models, or Connectionist Networks, are a class of learning algorithms that map input into one or many intermediate layers of continuous valued vector representation. Each is a universal funciton approximator, capeable of learning arbitrarily complex relationships between input and output (Hornik et al., 1989); thus, they can learn a class of logical relationships impossible for simpler learning algorithms, such as 'exclusive or'. While ANNs have typically been described as 'black box' models, much is now known about their formal representative capabilities, and we touch on these briefly below.

Recurrent Neural Networks (RNNs) (JRNN, GRNN) are a type of ANN, which were introduced in Elman (1990). RNNs consume multiple fixed-size inputs sequentially, an architectural choice that permits them to make generalizations about series that may vary in length, like sentences of human language. Elman (1991) showed that simple recurrent networks (SRNs) are able to model linguistic phenomena from data generated by toy grammar fragments. However, SRNs have difficulty learning and maintaining dependencies that involve long linear distances (Rodriguez, 2001). More contemporary Long Short-Term Memory

(LSTM) models (Hochreiter and Schmidhuber, 1997) are better at maintaining dependencies between distant items (Schmidhuber et al., 2002), and may be particularly suitable for modeling human language because they can implement counters, which they can deploy to process context free expressions (Weiss et al., 2018). In practice, even SRNs can recognize bounded hierarchical languages with reasonable efficiency (Hewitt et al., 2020). We use two LSTM models. The first, originally presented as the BIGLSTM+CNN in Jozefowicz et al. (2016) which we thus refer to here as JRNN, was trained on the training portion of the One Billion Word Benchmark of newswire text (Chelba et al., 2013) and has two hidden layers with 8196 units each. It uses the output of a character-level Convolutional Neural Network as input to the LSTM. The second, originally presented by Gulordava et al. (2018) model which we thus we refer to as GRNN, was selected for its previous success at learning the subject-verb number agreement task, and does not include a CNN embedding network. It was trained on 90 million tokens of English Wikipedia, and has two 650-unit hidden layers.

Transformers (GPT-2, GPT-3) are a type of neural network that have produced many of the most recent state-of-the-art performances on Natural Language Processing tasks. They encode the positional relationships between elements directly into the input representation itself. Information is propagated through the network via self-attention—each input token i is connected to every other token at the next layer, with the relative weights of the connection corresponding to its importance for predicting the token the next layer up. These models have often have large numbers of parameters (some, in the hundreds of billions) and must be trained on large datasets. Theoretical results about the capabilities of transformers paint a mixed picture. Hahn (2020) proves that they are not able to recognize unbounded hierarchical structure, nor even all of the regular languages. Although transformers *are* able to recognize languages up to a bounded depth, this raises questions about model ability to represent the underlying mathematical formalism assumed by theoretical linguists (Shieber, 1985). We use two Transformer models: GPT-2, which was trained on $\sim$ 8billion tokens of internet text (Radford et al., 2019),[8] and its successor GPT-3, which was trained on $\sim$ 114 billion words broken into $\sim$ 500 billion subword tokens (Brown et al., 2020).[9]

[8] We used the version of GPT-2 available through the `Language Modeling Zoo` distribution (https://cpllab.github.io/lm-zoo/index.html#welcome-to-lm-zoo)

[9] This is an estimate derived from the $\sim 1,000,00$words/5.2MB WSJ portion of the Penn Treebank and the reported size of the GPT models' training dataset, which was 40GB (GPT-2) and 570GB (GPT-3).

3.4 Possible Concerns

Data Size and Genre: If we assume that children are exposed to about 30,000 words per day (≈ 11-million words per year) (Hart and Risley, 1995), then GRNN has the linguistic experience comparable to a 8-year-old. If this rate of exposure persists throughout a typical adult lifetime, then JRNN roughly has the linguistic experience of a 80-year-old, GPT-2 that of 10 human lifetimes, and GPT-3 that of 100 human lifetimes. Additionally, all models were trained on newswire or similar adult-directed text, which may differ from child-directed speech in critical ways.

Although some of these models are clearly exposed to vastly more linguistic experience than human children, their learning outcomes may still be relevant in addressing what grammatical generalizations are learnable *in principle*, which can bear on analytic arguments about learnability. In addition, GRNN's data size is approximately equal to that of late childhood, when we would expect robust acquisition of islands. Crucially, though, these differences are made moot by our results, which indicate that data size has little effect on qualitative model success, with our smallest data model performing equivalently to GPT-3. Commenting on genre, it is an open question whether it has a substantial effect on human language learning outcomes. The manner and rate at which adults speak to children varies substantially across cultures, and there are some cultures in which children are spoken to infrequently yet still learn their native language (Weber et al., 2017; Cristia et al., 2019).

Probability and Grammaticality: The last issue, related to the one above, concerns the relationship between the probabilities assigned by statistical language models and judgements of grammaticality. As Chomsky (1957)'s original arguments here have played an important role in the relationship between computational and theoretical linguistic inquiry, we quote them at length below:

> [T]he notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English." It is fair to assume that neither sentence (1) *Colorless green ideas sleep furiously* nor (2) *Furiously sleep ideas green colorless* has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. [...] We cannot, of course, appeal to the fact that sentences such as (1) 'might' be uttered in some

> sufficiently far-fetched context, while (2) would never be, since the basis for this differentiation between (1) and (2) is precisely what we are interested in determining. [...]
>
> If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between order of approximation and grammaticalness. Despite undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances. I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure. (Chomsky, 1957)

When interpreting this line of reasoning, it is important to distinguish between Chomsky's *empirical* and *theoretical* claims. Especially in the first paragraph, the main critique is not that statistical models wouldn't be interesting if they *could* distinguish between unseen grammatical vs. ungrammatical sentences, but that they are unable to do so. This is certainly true for any language model on hand in 1957. However more contemporary statistical models are able to make abstract generalizations about novel strings (Pereira, 2000) and distributional categories (Belinkov et al., 2017; Liu et al., 2019). Thus, these models may be able to answer questions about whether the training data can support the type of abstract categorical representations that humans rely on to adjudicate between two novel sentences. To illustrate the point, one of the modern language models used in this study (GRNN) assigns a summed surprisal of 78 to (1), but 100 to (2), meaning that the ungrammatical unseen sentence is orders of magnitude more likely under its distribution than the unseen grammatical variant.[10]

Chomsky's empirical arguments continue in the third paragraph where he points out that were one to rank every possible permutation of English words in terms of their probability assigned by a language model, the grammatical strings would not necessarily all rank higher than the ungrammatical ones. This is true, at least for the language models tested here, and

[10] Given the fact that (1) has become something of a celebrity sentence and may very well be in the model's training data, we also point out that the wholly novel sentence pairs *Round excitement painted seven attitudes* and *Painted attitudes excitement round seven* received summed surprisal values of 87 and 88 respectively.

largely has to do with issues of length and word frequency. For example, the same neural model that was able to correctly rank (1) higher than (2), fails for the sentence pairs *Snails died the old* (summed surprisal: 59) and *The ancient crustaceans expired* (summed surprisal: 70) because the latter has more low-frequency words, which makes it less likely under the assumptions of the language model. However, our methodology is designed precisely to factor out confounds such as length and vocabulary frequency. All of our results report differences in surprisal between *identical substrings*, whose only difference is the context in which they appear. Thus, our results should not be interpreted as reporting grammaticality judgements, but rather whether models' incremental predictions are consistent with having made the right grammatical generalizations during training.

Turning from empirical arguments, there are at least two theoretical arguments being put forth in this section. The strong argument is that "statistical studies of language ... have no direct relevance to the problem of determining or characterizing the set of grammatical utterances." Put another way, probabilistic models should not be likened to grammars. While we believe there are good arguments both for and against this claim, our goal in this paper is *not* to provide a general probabilistic model of grammaticality, so we set it aside. The second, weaker argument put forward at the end of the paragraph is that "probabilistic models give no particular insight into some of the basic problems of syntactic structure." This weaker claim obviously depends on the point of view of the researcher, however we would like to note that the "basis for differentiation" between grammatical and ungrammatical sentences is something Chomsky puts forward as being of interest. As our experiments determine whether this differentiation can be accomplished by learning algorithms without innate syntactic biases, we believe that under Chomsky's own criteria they do give insight into the problems of syntactic structure.

4 Modeling Results: Base Experiments

4.1 Basic Licensing & Flexibility

One unique feature of the filler–gap dependency is that it is highly flexible. Fillers can license gaps in subject, object and indirect object positions. In order to test these basic properties we created 63 sentences following the three conditions outlined in Example (5).

Short adverbial phrases were added after the filler which, otherwise, would be adjacent to the gap site in the *subject* condition. For this and subsequent experiments we use obligatorily object-taking verbs and balance filler type (*who* vs. *what*) across the items. Bolded text indicates critical regions.

(5) a. I know who without thinking ___ **showed** the slides to the guests after lunch. [SUBJECT]

b. I know what without thinking the businessman showed ___ **to the guests** after lunch. [OBJECT]

c. I know who without thinking the businessman showed the slides to ___ **after lunch.** [PREP. PHRASE]

We will walk through these results in detail, and present summary figures for the remainder of the experiments. Our models output probabilities, which we turn into *surprisal* values (negative log probabilities). Figure 2 shows these raw surprisal values for the JRNN model, with the x-axis corresponding to sentence region and the y-axis indicating average by-word surprisal values for all the tokens in the region. Higher values mean the tokens are less likely under the language model's distribution. Critical regions are highlighted with teal boxes. Average surprisal starts off at around 10 bits/token in the first region, with surprisal the +filler condition greater than in the -filler condition (that is, the purple bars are higher than the green bars). This is because 'that' is a much more common complementizer than wh-words. The average by-word surprisal is even higher in the second region (e.g. '...without thinking...'), as adverbial phrases rarely come before subject position in embedded sentences. For the remainder of the phrase, we see the following behavior: When gaps are absent (left column) +filler conditions are higher in filled argument structure positions (e.g. *the CEO*, *the slides*, *the guests*) than -filler conditions. Crucially, this pattern is reversed when gaps are present. In the -gap conditions (right column) the -filler is higher surprisal than the +filler condition, in the critical regions immediately following a gap site. This reversal corresponds to the interactive effect we expect to find if models had successfully learned the dependency between fillers and gaps.

Now, instead of reporting raw surprisal values, we will focus of the difference between the +filler and -filler conditions (the differences between the purple and green lines in Figure 2), the so-called *wh-effect*. We zoom in on this in Figure 3, still for JRNN, with the same sentence regions on the x-axis, but with the wh-effect plotted on the y-axes. Here, the red

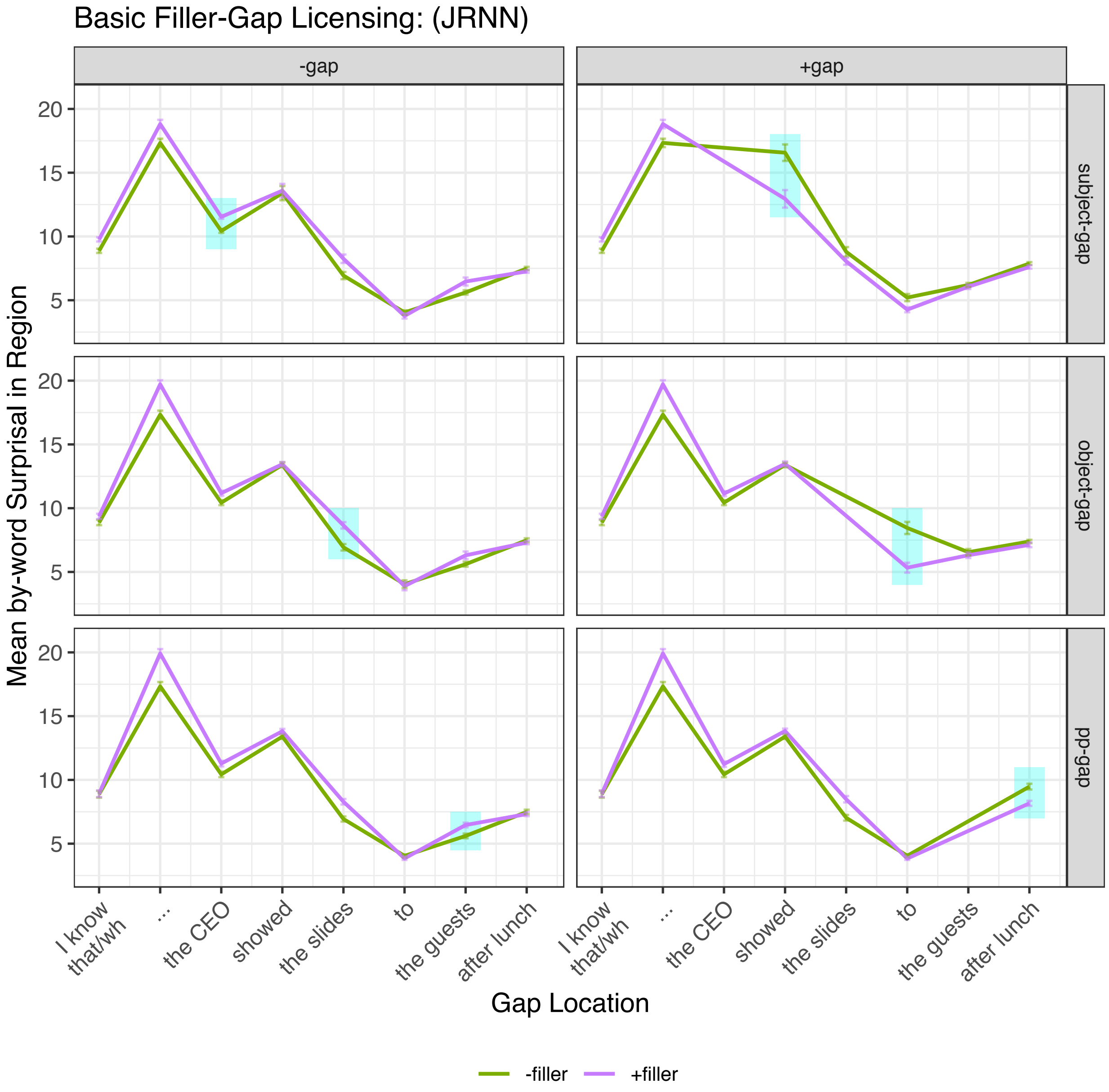

Figure 2: Mean by-word surprisal by sentence region for the JRNN model, with critical regions highlighted in teal. In the absence of a gap (left column) *+filler* conditions hare more surprising. When gaps are present (right column) this trend is reversed, corresponding to our predicted interaction. Slight differences in the first region between tests are due to the different proportion of *who* and *what* in the test materials.

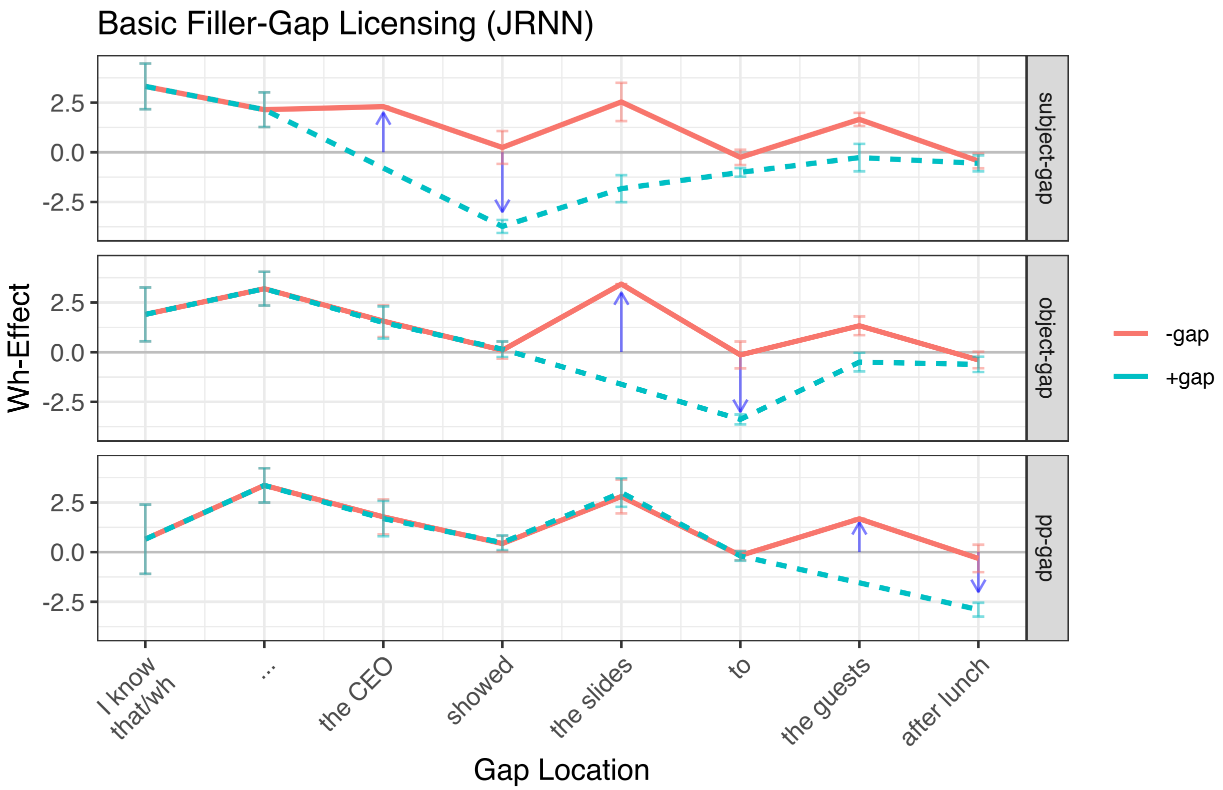

Figure 3: Wh-effect by sentence region, with blue arrows indicating regions of interest in each condition. In the -gap conditions, wh-effects are positive in critical regions. In the +gap conditions, wh-effects are negative. Error bars are 95% confidence intervals across 63 test sentences.

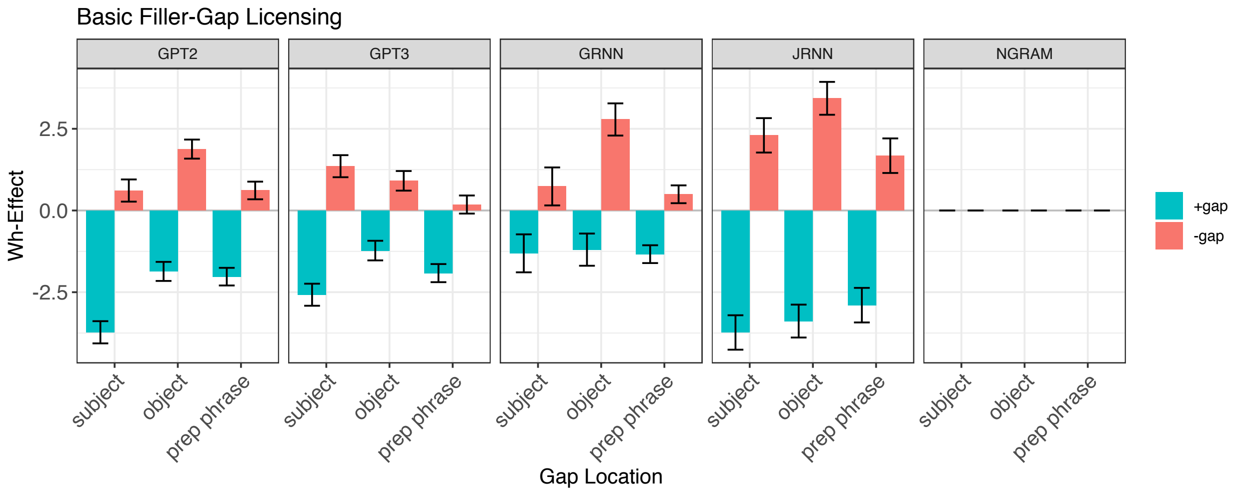

Figure 4: Basic Licensing. Negative wh-effect in the *+gap* condition (blue bars) and a positive wh-effect in the *-gap* condition (red bars) shows that all models are learning the basic filler–gap dependency.

lines are the wh-effect in the -gap condition, and the dotted blue lines are the wh-effect in the +gap condition (the blue line skips over sentence regions that are gapped in that test). In this figure there are two patterns to note: First, the blue dashed lines are all negative in the critical region immediately following a gap, indicated by the downward pointing blue arrows in the figure. This corresponds to the contrast in (3), namely that the presence of a gap is less surprising when it is licensed by an upstream filler. Second, the solid red lines are all positive in the critical regions that correspond to argument structure slots (*the CEO*, *the slides*, *the guests*), indicated by the upwards pointing blue arrows in the figure. This corresponds to the contrast in (4), namely that filled argument structure positions should be more surprising in the presence of a filler.

For the remainder of this paper, we will zoom in even further, showing just the wh-effects in the critical regions, which correspond to the differences highlighted by the blue arrows in Figure 3. These can be seen as bar charts for each of the models tested in Figure 4, which is the presentational paradigm that will be used from here on out. The various experiments are on the x-axis, and the wh-effect is on the y-axis. Error bars represent 95% confidence intervals. Roughly speaking, if the blue bars are well below the zero line and the red bars are well above it, we can expect a significant interaction from our regression model. As expected, in statistical tests we found a significant interaction term for all four of our neural models ($p < 0.001$ in all conditions; except for the n-gram model), indicating that they have learned the basic co-variation between fillers and gaps.

Turning to effect size, in the +gap condition for the *subject* test, the wh-effect is about 4 bits, which corresponds to ≈25% of the average by-word surprisal in the preceding region (Region 2, in Figure 2) and ≈40% of the average by-word surprisal in the following region (Region 5, or "the slides", in Figure 2). That is to say, in +gap conditions, the wh-effect is a large percentage of the average by-word surprisal for a baseline grammatical sentence. However, in the -gap conditions, the wh-effect size is smaller. This pattern is true generally across all of our tests, and not necessarily unexpected. As mentioned in Section 3.2, the presence of a filler sets up an expectation for a gap, but not in a particular argument structure position. Therefore, when encountering a filled subject (e.g. "the CEO" in Figure 2), the sentence is not rendered ungrammatical as a gap could occur in downstream object or indirect object positions. We do not wish to claim that the difference in the +gap conditions corresponds to ungrammaticality, whereas the difference in the -gap condition does not, given that they both show significant licensing effects. What's important for our purposes is that both behaviors correspond to an over-all expectation for gaps in subject, object and indirect object locations.

Comparing, the models' licensing strength to the distribution of gaps in English written text suggests that models have made generalizations beyond the statistics encountered in their training data. An analysis of the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993) reveals that in written English text there are about two times the number of gaps in subject position compared to object position, and two times the number of gaps in object position compared to prepositional phrases inside the VP.[11] But contrary to the distribution of gaps in written English, GPT-2, JRNN and GRNN show the strongest

[11] We find 21,041 Subject Gaps, 11,657 object gaps and 5,179 PP gaps. Corpus data was collected using Tregex (Levy and Andrew, 2006), with the following commands. For gaps in subject position: `NP-SBJ < (-NONE- < ( @* ) )`. This command searches for an NP subject node that dominates a trace. For gaps in object position: `VBD|VBG|VBN|BVP|VBZ $++ (@NP (< (-NONE- < @*)))`. This command searches for NP nodes that dominate a trace and are the immediate right sister of a V node. For gaps in VP-internal prepositional phrases or indirect objects (which the PTB parse schema treats as prepositions): `@PP <+(!@VP) (@NP < (-NONE- < @*) )`. This command searches for a PP node that dominates an NP node, which dominates a trace. The chain between the PP and NP node must not include a VP node. This was done to exclude gaps that are inside relative clauses whose head is dominated by a preposition.

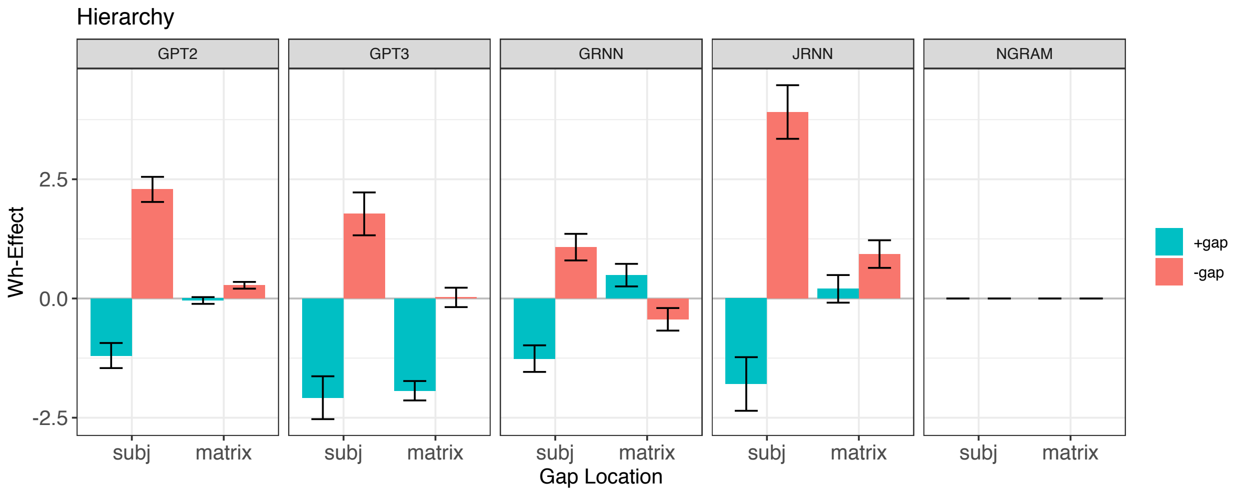

Figure 5: Sensitivity to syntactic hierarchy. All models show reduction (or inversion) of wh-effects in the *matrix* condition.

expectation for filler–gap dependencies in *object* position, with slightly less, but about equal expectation for filler–gaps in *subject* and *PP* conditions. GPT-3's licensing pattern more closely matches the distribution of gaps we would expect to find in a large corpus. These results suggest that model behavior is a result of learned generalizations that goes beyond rote recapitulation of training-corpus statistics.

4.2 Hierarchy

The experiments presented above are compatible with a strictly linear relationship between fillers and gaps. A model with a simple linear heuristic—"If I see a filler, expect a gap downstream where an NP is likely to occur"—would perform well. In reality there are a number of hierarchical constraints implicated in the filler–gap dependency. To a first approximation these state that the filler must *c*-command the gap (although the precise relationship is more complex (see Pollard and Sag (1994)). Example (6) demonstrates this relationship with a filler–gap sentence, where the portion of the sentence c-commanded by the filler *who* is demarcated with brackets. The sentence (6-b) is ungrammatical because the gap site is not in this region.

(6) a. The fact that the reporter knows who [the witness surprised ___ **with his testimony**] surprised the jury during the trial. [SUBJECT GAP]

b. *The fact that the reporter knows who [the witness shocked the jury with his testimony] surprised ___ **during the trial.** [MATRIX GAP]

If models have learned these tree-structural restrictions, we should expect two things: First, when the filler does not *c*-command the gap, its presence should not affect the likelihood of a gap. That is, the wh-effect in the +gap condition should be close to zero. Second, the presence of an upstream filler should not make the filled argument structure position more or less surprising. That is, the wh-effect in the -gap condition should also be close to zero. Putting this all together, we predict that if the models are learning the structural restrictions on the filler–gap dependency, wh-effects should be strong in grammatical conditions like (6-a), but close to zero in the ungrammatical conditions like (6-b).

In order to test this prediction we created 45 sentences following the template in (6). We matched the verbs across the two conditions to control for spurious lexical frequency effects (i.e. *surprised* is in both *subject* and *matrix* conditions in the example). The results from this experiment can be seen in Figure 5. All neural models show proper wh-effects in the *subject* condition, where the gap is properly licensed. Crucially, each also shows a significant reduction (or inversion) of wh-effects in the *matrix* condition, when the gap is no longer *c*-commanded by the filler ($p < 0.001$ for JRNN and GPT-2, and $p < 0.05$ for GRNN and GPT-3). The *n*-gram model show no variation between conditions.

4.3 Unboundedness

One worry is that the previous experiment doesn't test sensitivity to hierarchy *per se*, but merely sensitivity to distance: models demonstrate wh-effects in the *subject* condition because the gap site is closer to the filler. The filler–gap dependency, however, is unbounded insofar as the location of the filler and the gap can be separated by any number of nodes in a parse tree, as in (7), where the final sentence may be difficult to process, but (arguably) remains grammatical despite the intervening material. Importantly, the grammaticality of sentences like (7-c) demonstrates the types of generalizations that language learners make about the filler–gap dependency. An analysis of the constituency parses of the CHILDES Treebank (Pearl and Sprouse, 2019) reveals that of the ∼100,000 sentences in the dataset, 396 contain 2 layers of sentential embedding, 44 contain 3 layers of sentential embedding

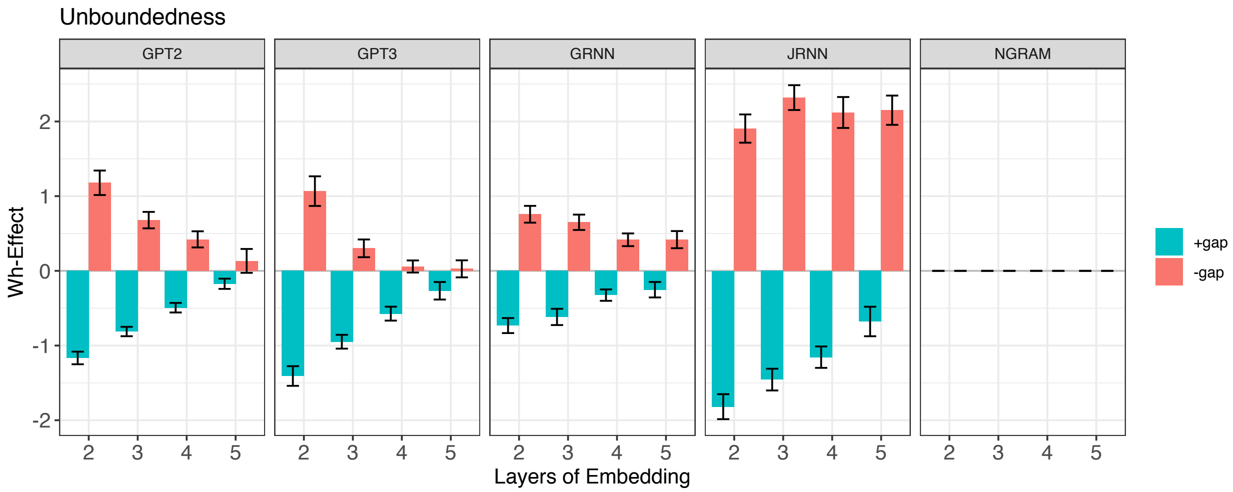

Figure 6: Undboundedness of the Filler–Gap Dependency

and 9 contain 4 layers of sentential embedding.[12]

(7) a. I know who the countess offended ___ at the ball. [1-LAYER]

b. I know who the footman believed the countess offended ___ at the ball. [2-LAYERS]

c. I know who the count remembered the lady reported the guard said the czar thought the footman believed the countess offended ___ at the ball. [5-LAYERS]

In order to test potential unboundedness, and to provide a control for the *hierarchy* tests, we created 54 sentences following (7) with between 2 and 5 layers of sentential embedding.[13] The results for this experiment can be seen in Figure 6. All of the neural models demonstrate proper wh-effects for all conditions ($p < 0.001$ for all models and all layers of embedding). In order to test whether models show a significant reduction in licensing between layers, we fit a linear model with number of layers as an additional predictor and inspect for a significantly positive three-way interaction between the presence of a filler, the presence of a gap and the number of layers. For GPT-2 and GPT-3, we find that there is a significant interaction after one additional layer ($p < 0.01$ for all contrasts); for GRNN the reduction is significant after two additional layers ($p < 0.001$); and for the JRNN the reduction is significant after three

[12] Searches were conducted on Tregex (Levy and Andrew, 2006) and commands were formulated using the following pattern, which matches sentences that contain four layers of sentential embedding: `S << (S << (S << (S << (S << S))))`

[13] We do not include sentences with a single layer of sentential embedding in these tests, as sentences would have the same structural configuration as the *object* tests in Section 4.1.

additional layers ($p < 0.01$), largely driven by reduction in wh-effects in the +gap conditions.

Next, in order to compare the results from this experiment to the results from the hierarchy experiment, we combine data from both experiments, and fit a regression model with two additional predictors: *test type* (*heirarchy* vs. *unboundedness*) and whether the distance between the filler and the gap is *proximal* or *distal* (*subject* and *embed2/3* are proximal, whereas *matrix* and *embed4/5* are distal). We treat filler-to-gap distance as a categorical variable due to the categorical structure of the hierarchy test suites. Here, a negative four-way interaction between *filler*, *gap*, *distal* and *test type* means that the grammatical *unboundedness* tests show stronger licensing effects in the critical regions (they are associated with a reduction in surprisal when both fillers and gaps are present). We find significant interactions for JRNN ($p < 0.0.001$), GRNN ($p < 0.05$), GPT-2 ($p < 0.05$), but not for GPT-3, which generally shows the weakest results for this test.

There are two takeaways from this experiment: These results indicate that models can thread wh-expectations through complex syntactic environments, providing the controls needed to conclude that the reduction of wh-effects observed in the *hierarchy* experiment are not due merely to distance, but to a restriction based on their structural relationship.[14]. Second, given the fact that models are likely to have seen very few sentences with four layers of embedding or more, and yet show significant wh-effects in these conditions indicates that they are making some generalization beyond their training data. The fact that we see stronger traces of unboundedness in the RNN models over the Transformers, despite their smaller training vocabularies suggests that their architecture makes them better models of human linguistic competence, at least for this phenomena.[15]

5 Island Effects

5.1 Methodology

In this section, we investigate seven of the most studied island phenomena (Ross, 1967; Huang, 1982). If models are learning that the co-variation between fillers and gaps does not

[14] For a secondary length control that addresses non-hierarchical interveners see Section 3.2 of Wilcox et al. (2018).

[15] See also Da Costa and Chaves (2020) who found similar limitations for another Transformer language model, Transformer-XL (Dai et al., 2019).

hold when gaps are in island positions, then two things should be true: First, when a gap is inside an island construction, the presence or absence of an upstream filler should have no effect on its relative likelihood. That is, in the +gap condition the wh-effect should be close to zero. Second, the presence of absence of an upstream filler should not affect the relative surprisal of a filled argument structure position located inside an island. That is, in the -gap condition, the wh-effect should also be close to zero. This is the same logic that we used to make predictions about model acquisition of hierarchical generalizations in Section 4.2, above. Following the setup in that experiment, here, we contrast the wh-effects in an island condition with a non-island minimal pair counterpart. If the models are successfully learning island effects, we expect strong wh-effects in the non-island condition, and a reduction of licensing effects in the island condition.[16]

There is, however, a potential confound with this approach. It may be the case that reduced wh-effects are not due to generalizations learned about the filler–gap dependency, but rather that any information flow is blocked by sufficiently complex material. In that case, apparent island-like behavior would be an epiphenomenon of a larger inability to thread information through syntactically complex environments. To account this confound we employ controls that utilize expectations for gendered pronouns, set up by morphologically gendered nouns, such as *baron* and *baroness*. We measure the noun's *masculine expectation* by taking the difference between the masculine pronoun *his* and the feminine pronoun *her*, following (8), below.[17] If models have learned the gender dependency, then we expect a strong mascu-

[16] As noted before, in statistical tests this will correspond to a three-way interaciton between +/-*island*, +/-*filler* and +/-*gap* that results in a supperadditive *increase* in surprisal. In our *hierarchy* experiments, we intentionally use the same verb in multiple structural positions to control for potential frequency confounds. We do the same here, except for Complex NP and Subject Islands, for matters of coherence in the testing materials. In these cases, we confirm that the average log frequency of the verbs are not significantly different from each other via a t-test ($p > 0.5$ for both tests; frequency data from Franz and Brants (2006)). Tests are largely adapted from Wilcox et al. (2019) and Wilcox et al. (2018), although additional items have been added.

[17] Unlike the filler–gap dependencies, these do not constitute grammatical/ungrammatical contrasts, as *her* in (8-b) could refer to a sentence-external discourse referent. However, because we needed to find a dependency that could cover the same distribution as the islands–i.e. it could extend into relative clauses and across sentential embeddings—grammatically-determined anaphoric dependencies were not available to us. To counteract the potential confound of exophoricity, we use verbs that are likely to refer to the sentence's subject, like *brushed* in (8); all of our models demonstrate strong gendered expectations effects.

line expectation when the head noun is morphologically masculine and a strongly negative masculine expectation when the head noun is feminine. We set up this 2 × 2 interaction to mimic the experimental design of our island tests.

(8) a. The **baron** said they brushed **her** hair. [MASC, FEM]
b. The **baron** said they brushed **his** hair. [MASC, MASC]
c. The **baroness** said they brushed **her** hair. [FEM, FEM]
d. The **baroness** said they brushed **his** hair. [FEM, MASC]

In addition to these basic licensing conditions, we construct examples where the gendered pronoun is inside an island structure, relative to the head noun. Examples for the *Wh Island* and *Complex NP Island* variants are given in (9), below.

(9) a. The **baron/baroness** said whether they brushed **his/her** hair. [WH ISLAND]
b. The **baron/baroness** said they liked the attendant who brushed **his/her** hair. [COMPLEX NP ISLAND]

If models had learned that complex syntactic structures block all information flow, then we would expect a reduction of gender expectations inside of island configurations. If, however, the models have learned island constraints as a unique feature of the filler–gap dependency, then we should expect no reduction of gender expectations in sentences like (9). We test for significant reduction between the control conditions, like (8) and island conditions using the same statistical procedures as for the filler–gap sentences. What we look for in this section, then, is a reduction of wh-effects between island/non-island conditions, but a *lack* of reduction in gender expectation.

5.2 Island Experiments

Adjunct Islands: Gaps cannot be licensed inside an adjunct clause, as demonstrated by the contrast between (10-a) vs. (10-b).

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

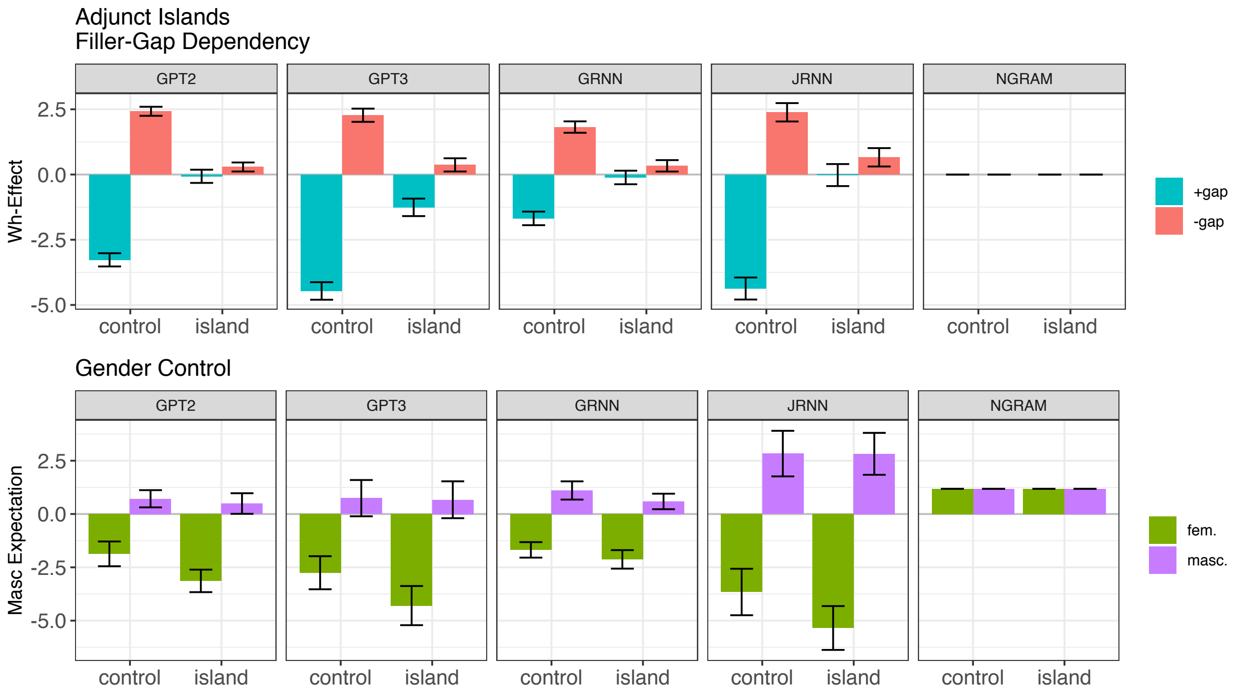

Figure 7: Adjunct Islands

d. The actress got mad after they insulted {**his/her**} performance. [ISLAND, GENDER EXP.]

To test for sensitivity to adjunct islands we created 54 items following the template in (10). The results for this experiment can be seen in Figure 7, with the filler–gap dependency on the top and the gender expectations on the bottom. For the gender dependency, we see the neural models models performing exactly the same in the control conditions as in the island conditions. Turning to the filler–gap dependency, we find a significant reduction of wh-licensing interaction between the control and island conditions for all neural models models ($p < 0.001$), but not for the n-gram model. This is precisely the behavior we would expect if models had learned the selective blocking of filler–gap dependencies inside island constructions.

Complex NP Islands: Gaps are not licensed inside S-nodes that are dominated by a lexical head noun, as demonstrated by the contrast between (11-c) compared to (11-a).

(11) a. I know what the actress bought __ **yesterday**. [CONTROL, FILLER–GAP]

b. *I know what the actress bought the painting that depicted __ **yesterday**. [ISLAND, FILLER–GAP, THAT-COMP]

c. *I know what the actress bought the painting which depicted __ **yesterday**. [ISLAND,

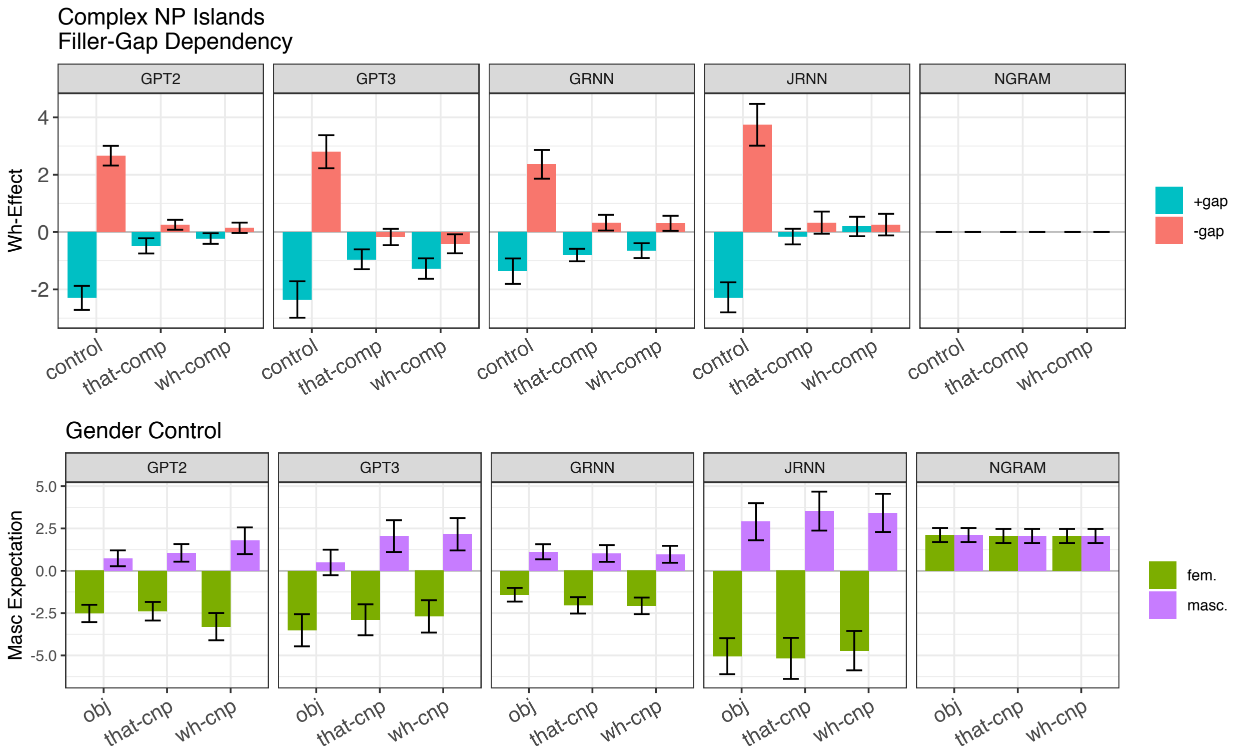

Figure 8: Complex NP Islands

FILLER–GAP, WH-COMP]

d. The actress said they saw her {**his/her**} performance. [CONTROL, GENDER EXP.]

e. The actress said they saw the exhibit {that/which} featured {**his/her**} performance. [ISLAND, GENDER EXP.]

We created items following the examples in (11), with two island conditions: one in which the complex NP is headed by a wh-complementizer and one in which it is headed by a that-complementizer. The results from this experiment can be found in Figure 8. We find a reduction in wh-effects for all neural models ($p < 0.001$ for both *that* and *wh* conditions for JRNN, GPT-2 and GPT-3, $p < 0.01$ for GRNN). We found no reduction in gender expectation between *control* and *island* conditions for any model.

Coordination Islands: The coordination constraint states that a gap cannot occur in one half of a coordinate structure, as demonstrated by the difference in acceptability between (12-b)/(12-c) and (12-a), in which the whole object has been gapped. From an incremental model, however we would only expect a reduction in wh-licensing when the gap is in the second conjunct. This is because when the gap is in the first conjunct the continuation may

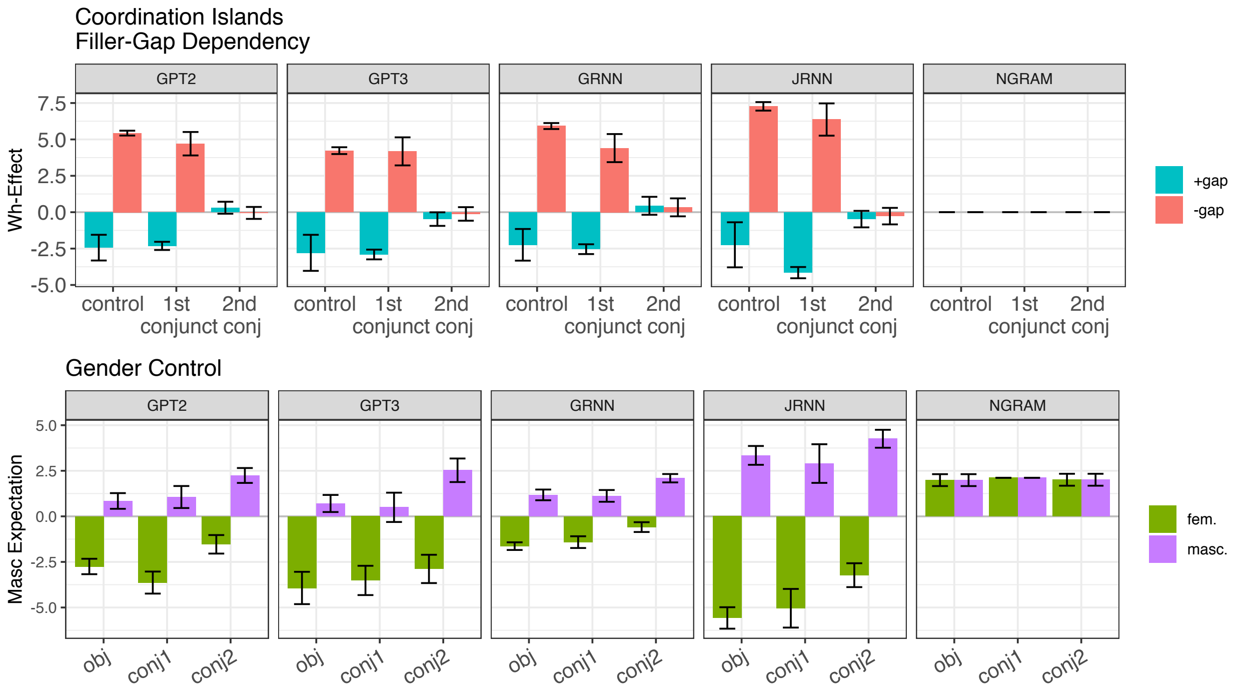

Figure 9: Coordination Islands

end grammatically, such as *I know what the man bought __ and the painting later depicted __* (an example of Across the Board movement). This is not the case when the gap occurs in the second conjunct, which cannot be rescued through ATB movement and is rendered ungrammatical at the gap site.

(12) a. I know what the man bought __ **at the antique shop**. [CONTROL, FILLER–GAP]

b. *I know what the man bought __ **and the painting** at the antique shop. [ISLAND, FILLER–GAP]

c. *I know what the man bought the painting and __ **at the antique shop**. [ISLAND, FILLER–GAP]

d. The fireman knows they talked about {**his/her**} performance. [CONTROL, GENDER EXP.]

e. The fireman knows they talked about {**his/her**} performance and the football game. [ISLAND, GENDER EXP., 1ST CONJ.]

f. The fireman knows they talked about the football game and {**his/her**} performance. [ISLAND, GENDER EXP., 2ND CONJ.]

We created experimental items following the examples in (12). In the *-gap* conditions, for the *control* sentences, we sum over a whole conjoined NP (e.g. "the painting and the

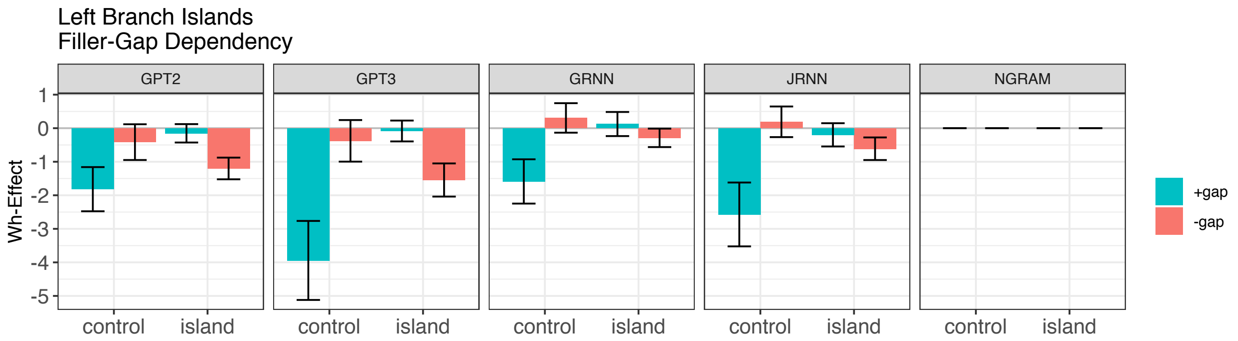

Figure 10: Left Branch Islands

chair"). In addition, we add extra material to the beginings of the *control* and *first conjunct* sentences, to maintain similar linear distance between the filler and the gapsite between all three tests. Results can be seen in Figure 9. We find a significant reduction in wh-effects ($p < 0.001$ for all four neural models), but no such reduction in the gender expectation. The effects here were the largest effect sizes out of all the island configurations tested.

Left Branch Islands: The left-branch constraint states that modifiers which appear on the left branch under an NP cannot be gapped, accounting for the relative ungrammaticality of (14-d) compared to (13-d). We created 20 items following the examples below and measured the wh-licensing interaction in the post-gap material. Because in the island case the moved material consists of an adjective phrase like *how expensive*, we were unable to create minimal pair gender controls for this island and present this test without them. As the *no-gap* variants of these tests are harder to reconstruct than those of the other island experiments we present them below.

(13) a. I know that you bought an expensive **a car** last week.
b. I know that you bought __ **last week**.
c. I know how expensive a car you bought **a car** last week.
d. I know how expensive a car you bought __ **last week**. [WHOLE OBJECT]

(14) a. I know that you bought an expensive **a car** last week.
b. I know that you bought __ **last week**.
c. I know how expensive you bought **a car** last week.
d. *I know how expensive you bought __ **car last week**. [LEFT BRANCH]

The results can be seen in Figure 10. We find that the JRNN, GRNN and GPT-3 models

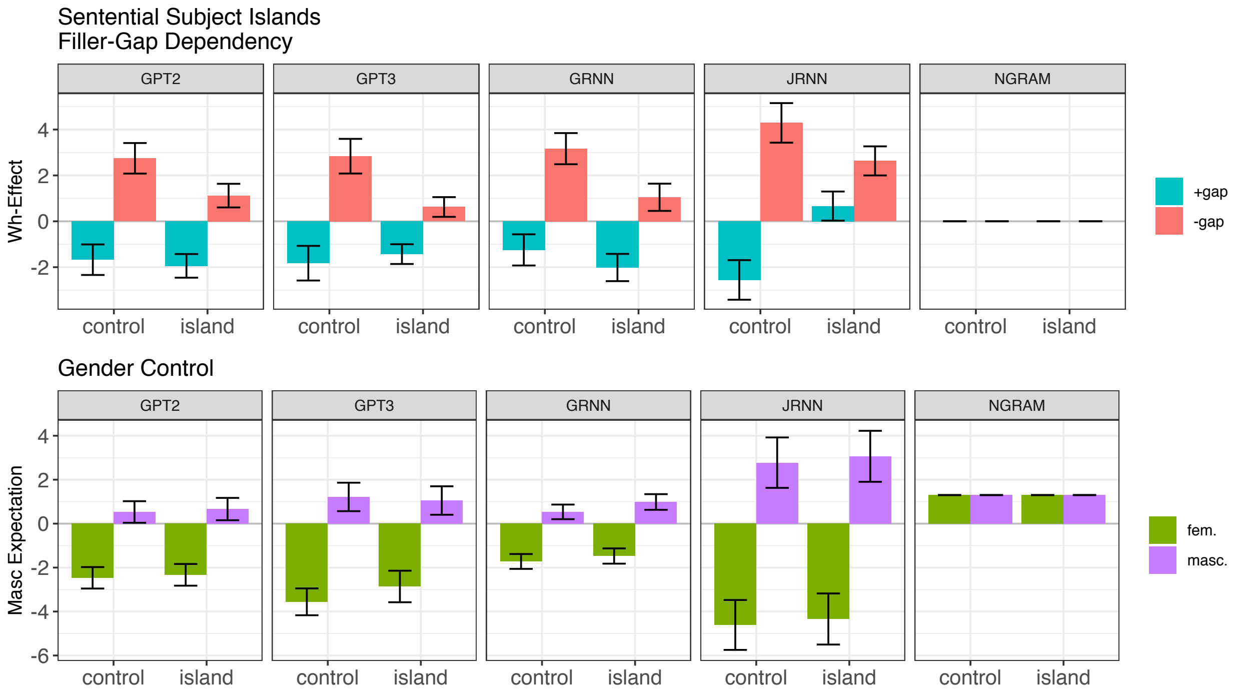

Figure 11: Sentential Subject Islands

show a reduction of wh-effects inside the island ($p < 0.01$ with $p < 0.05$ for JRNN). While the effect is in the right direction for GPT2, the interaction is not significant ($p = 0.12$).

Sentential Subject Islands: The sentential subject constraint states that gaps are not licensed within an S-node that plays the role of a sentential subject. To assess whether the models had learned this constraint we created items following the variants in (15).

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

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

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

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

The results for this experiment can be seen in Figure 11. We find a significant reduction of wh-effects between the island and non-island conditions for JRNN and GPT-3 ($p < 0.05$) but not for GRNN or GPT-2. No model shows reduction in gender expectations inside island configurations.

Subject Islands: Gaps are generally licensed in prepositional phrases, except when they occur attached to subjects. We created experimental items following the examples in (16).

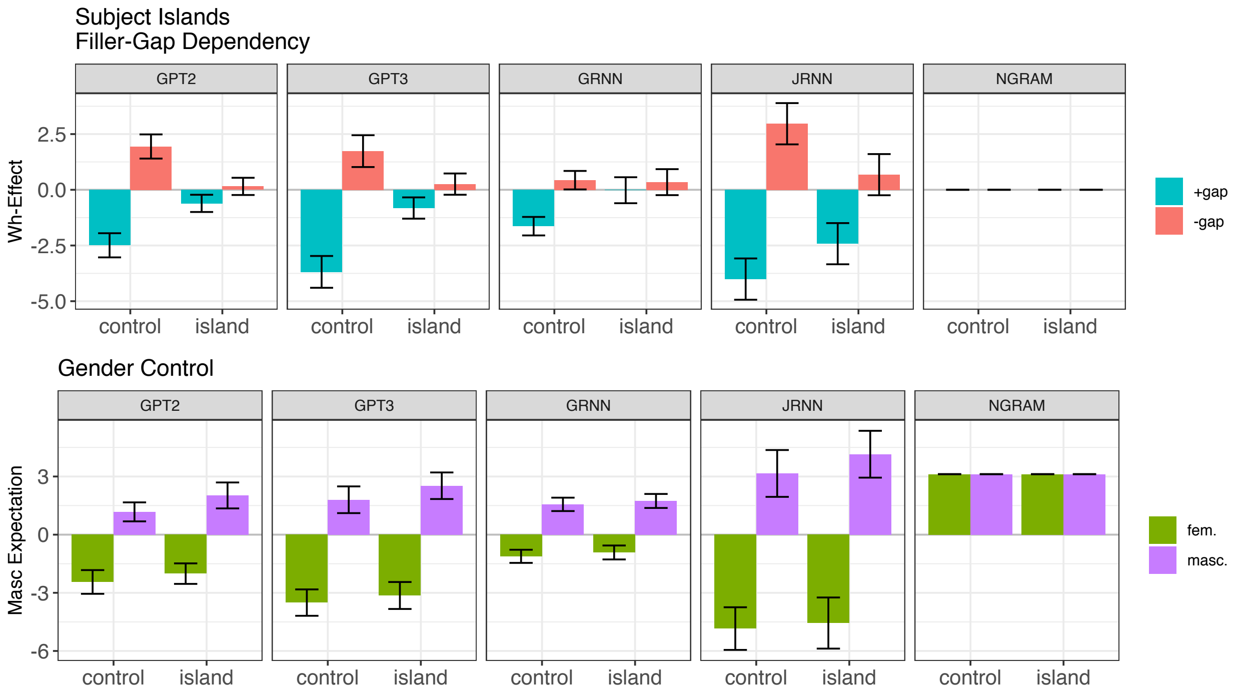

Figure 12: Subject Islands

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

b. *I know who the painting by __ **fetched** a high price. [ISLAND, FILLER–GAP]

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

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

The results from this experiment can be seen in Figure 12. We find a significant reduction of wh-effects for JRNN ($p < 0.05$), GPT-2 and GPT-3 ($p < 0.01$) but not for GRNN. The failure of GRNN in this case is because it does not produce robust wh-effects in the non-island controls, at least relative to the island conditions.

Wh-Islands: The wh-constraint states that the filler–gap dependency is blocked by S-nodes introduced by a wh-complementizer, as demonstrated in the unacceptability of (17-b) compared to (17-a). We created experimental items following the examples in (17) and measured their wh-effects.

(17) a. I know who Alex said your friend insulted __ **yesterday**.[CONTROL, FILLER–GAP]

b. *I know who Alex said whether your friend insulted __ **yesterday**. [ISLAND, FILLER–GAP]

c. The actress said they insulted {**his/her**} friends. [CONTROL, GENDER EXP.]

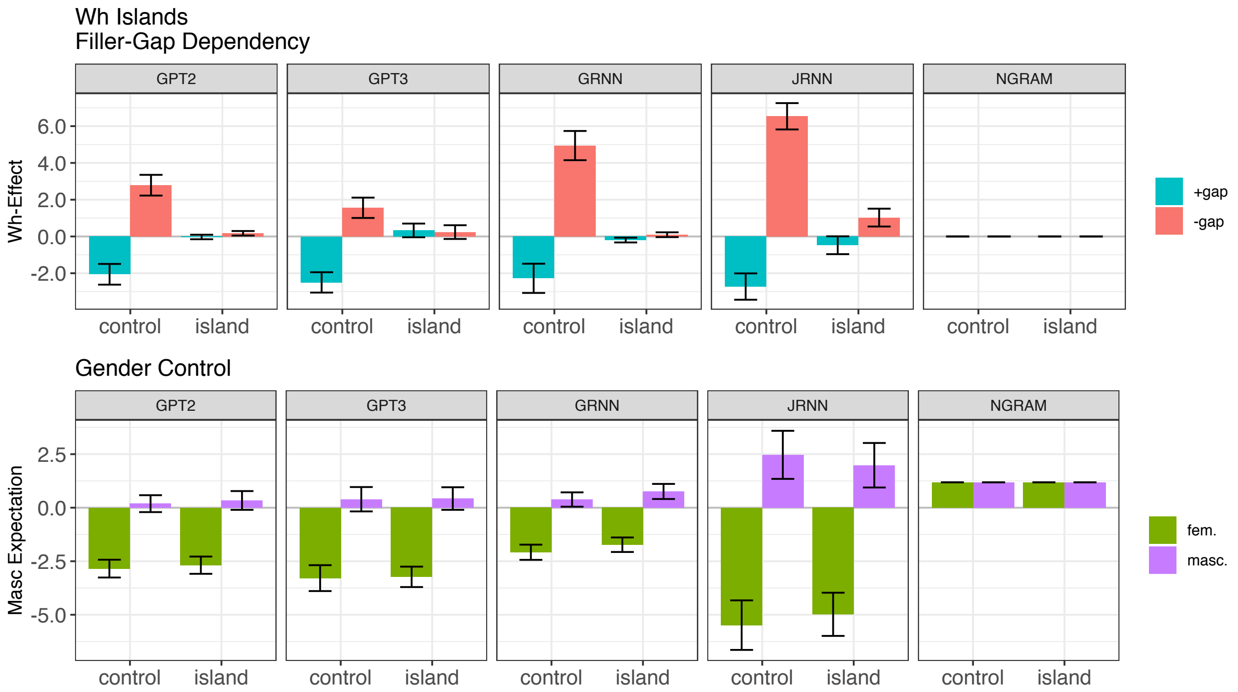

Figure 13: Wh-Island Effects

d. The actress said whether they insulted {**his/her**} friends. [ISLAND, GENDER EXP.]

The results for this experiment can be seen in Figure 13. We find a reduction in licensing effects for all four neural models ($p < 0.001$), with no reduction for the n-gram model. We find no difference between the island and non-island conditions in the gender controls.

Table 14 summarizes the results for our island experiments. Each cell corresponds to a single test, with the label reports two results: On the top is the level of significance achieved in our statistical tests, with stars indicating the level of confidence. Tests that reach the 0.05 confidence threshold are in green, otherwise the text is grey with "N.S." (not significant). Below, we report the fitted interaction term from the statistical model, corresponding to the estimated additional surprisal (in bits) due to island effects. Cell colors are within-model rank of effect size, with stronger island effects in darker color. Thus, dark cell colors with green text indicates strong island effects, and lighter cell colors with grey text indicate lack of island effects.

The most important take-away from this figure is that, in the vast majority of our tests (24/28), we find statistically significant effects of island phenomena. Despite cross-model architectural variation, we find strongest effects for Coordination, Adjunct and Complex NP

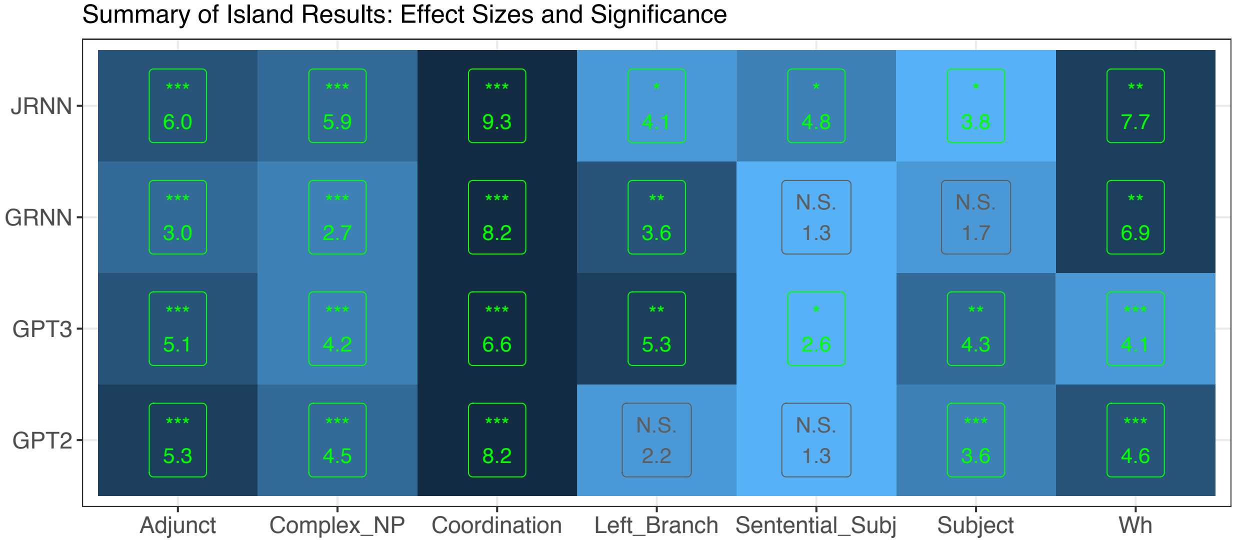

Figure 14: Summary of Model Performance on Islands tests. Green text indicates a statistically significant islandhood effect, with stars indicating the level of confidence. Numbers are the effect terms from our linear fits, corresponding to the estimated additional surprisal (in bits) due to islands. Colors are within-model rank of the effect size, with larger effects in darker colors.

islands, with weaker effects for Left-Branch Islands, Subject, and Sentential Subject islands.[18] Turning to differences between models, we find that JRNN shows successful behavior on all islands tested, outperforming GPT-2 and GPT-3, which was trained on orders of magnitude more data. This matches the findings from Hu et al. (2020) which suggest that when controlling for training data, GPT-2 does not perform better than RNNs at many grammatical generalization tasks. In summary, we take the results from this section to indicate that general-purpose pattern recognizers can acquire island effects from string data alone, and

[18] This may be due to the type of sentential subject employed in our experimental stimuli. In the syntactic literature, many examples are given using sentences with *that*-headed sentential subjects, such as *I know who that the count will duel ___ surprised us.*. But these sentences are difficult to translate into our 2x2 testing paradigm because both of the -*filler* conditions involve two *that*-complimentizers in a row, which may be difficult for humans and models to process. Instead, we constructed we constructed Sentential Subject Island violations using *for*-headed sentential subjects, such as *I know who for the count to duel ___ will surprise us.* Many islands have graded acceptability (Kush et al., 2013; Chaves, 2020), and it may be the case that *for*-headed sentential subject violations are judged as less ungrammatical than their *that*-headed counterparts. Further investigation is needed.

that these generalizations are possible when the training data is equivalent to the linguistic experience of an older child (GRNN). The fact that all models show lower power in the sentential subject and subject experiments, despite their different architectural arrangements and training regimes, raises the intriguing possibility that rules governing subject/object asymmetry may prove especially challenging to learn from data.

6 General Discussion

In this section, we first discuss the issue of *how* and *why* the models are learning what they do, arguing that the similarities in models' generalizations despite their architectural differences can be best understood through their objective function, which is to increase data liklihood. We then turn to the implications of the models' success for one influential argument in favor of linguistic nativism, the Argument from the Poverty of the Stimulus.

6.1 How Are the Models Learning?

We start with a set of questions adapted from Pearl and Sprouse (2013a) addressing their own computational model of island acquisition: (1) Why does the system attempt to learn dependencies between tokens or sequences of tokens? (2) Why does the system attempt to learn constraints on those dependencies? and (3) Why does the system treat wh-dependencies as separate from other dependencies, like RC dependencies and binding dependencies?

Before we begin answering these questions, we need to recapitulate three properties of our models: First, as discussed in Section 3.3, all of our models are weakly-biased and domain general learners, with no evidence for natural-language like syntactic biases. Second, the different neural networks vary widely in their assumptions about how the pieces of input data are related to each other and how the input data should be represented: GRNN represents each word as a single token, GPT-2 and GPT-3 break words down into sub-word units, and JRNN processes each word as a character string using a separate convolutional neural network component. In terms of the relationship among pieces of data, the two LSTM models understand the individual token-to-token relationships as unfolding through incremental state updates, which is intended to model time, whereas Transformers maintain a fully-connected network of relationships where each token is assigned a time 'index.' Third, neural networks in general, and our models in particular, are well-established to acquire representations that correspond to various levels of linguistic abstraction (Belinkov et al., 2017; Belinkov and Glass, 2019). LSTM models have been shown to capture syntactic part of speech (Blevins et al., 2018), and for transformers (including GPT-2 models) different attention heads specialize at processing different parts of speech as well as different dependencies (Vig and Belinkov, 2019).

With these pieces in place, we can turn to the questions posed above, which we address by articulating the horns of the learnability debate: Because the abstractions on which these dependencies operate are justified given the models' inductive bias by the distributional structure of the data, and with these abstractions available, co-variation between these categories is learned either because (a) it is favored by the inductive bias, or (b) because learning these dependencies improves data likelihood. Since we know that no model architecture has an obvious strongly natural language syntax oriented inductive bias, and we find similar results with multiple architectures that make different choices about both input representation and the relationship between these inputs, the most plausible conclusion is that data likelihood drives learning. Thus, in response to the first two questions, our hypothesis is that each model learns the constrained co-variation between fillers and gaps because it is a generalization available within the hypothesis space determined by the model architecture that that allows it to successfully shift probability mass away from unlikely sentences and place more weight on sentences that are more likely to occur.

The response to the third question ("Why does the system treat wh-dependencies as separate from other dependencies") is similar. Given previous work demonstrating that individual featural dependencies are tracked by individual neurons (Lakretz et al., 2019; Giulianelli et al., 2018), it is likely that the respective distributions that characterize each type of dependency are sufficiently distinct that they are represented differently in the network. Given this, the inductive bias either does not favor generalizing constraints from one dependency to others, or such generalization hurts data likelihood. In sum, the type of answer we propose for the explaining the success of our models stems from the nature of their objective function of maximizing data likelihood. If this is correct, then the model success can be related to the success of other, more transparent models that use data likelihood maximization to tackle issues of learnability, such as the model proposed for the acquisition of subject-auxiliary inversion in Perfors et al. (2006).

6.2 Argument from the Poverty of the Stimulus

We now discuss the implications of the models' success. The question we wish to gain traction on is the following: Where on the nativist-to-empiricist spectrum should we place island constraints? Nativist approaches posit that island constraints are learned by language-specific innate constraints, which guide the child's learning process. *Language-specific* refers to abstract rules or operations stated in terms of syntactic (or linguistic) structures, and are not used in other domains of cognition. These constraints *guide* the learning process by eliminating or down-weighting certain structures from the hypothesis space. *Innate*, as argued in Mameli and Bateson (2011), is often discussed as a single property, but it is more properly understood as a cluster (or less charitably a 'clutter') of interrelated concepts. Here, we use the term to describe representations that do not need to be learned and are sufficiently developmentally robust, or canalized. On the opposite end of the spectrum, empiricist approaches posit that island constraints are learned by applying domain general principles—such as pattern prediction—to the problem of efficient communication. As mentioned in Perfors et al. (2006), empiricist approaches to language learning do not deny that grammatical representations are present in the learner's hypothesis space. The difference is that empiricist approaches posit that these structures are selected with respect to some criteria (they do a good job predicting the data; they are simple) rather than because they are favored or disfavored *a priori*.

How can computational modeling provide evidence for or against nativist or empiricist approaches? To be clear, the performance of a given model does not provide *direct* evidence for or against the learnability of a particular structure. Given the vast distance between any computational model available today and the human brain, model success does not mean that the structure is necessarily learned and model failure does not mean that the structure is not learnable. Rather, we use computational modeling as an empirical corollary to analytic arguments about learnability, specifically the Argument from the Poverty of the Stimulus. Since it was first introduce in Chomsky (1957), the APS has been formulated differently by different people. The version presented here is a blend of the argument as construed in Clark and Lappin (2010) and Laurence and Margolis (2001). As the APS has typically been treated as an argument advocating linguistic nativism, it is important to situate it in the discourse of these debates. For this reason, as advocated in Clark and Lappin (2010) the way we will frame the APS is as a logical argument about the necessity of *language-specific* biases.[19]

Argument from the Poverty of the Stimulus:

1. Learners acquire target language L_0 through domain general learning algorithms or through algorithms with a strong language-specific learning biases.
2. The primary linguistic data are consistent with an infinite number of linguistic generalizations $L_0 \dots L_n$.
3. Language learners consistently acquire the actual generalizations of their target language, L_0.
4. There are no domain-general learning algorithms that favor L_0 over $L_1 \dots L_n$
5. Therefore, children acquire L through algorithms with a strong language-specific bias.

This formulation of the APS is useful because it provides a clear role for computational modeling experiments. If an algorithm with no language specific learning bias is able to acquire the proper linguistic generalizations, then Premise 4 does not hold and the argument falls apart. As stated by Lappin and Shieber (2007) in the context of the role of connectionist networks in debates about the APS: "[T]o the extent that a given machine learning experiment is successful in acquiring a particular phenomenon, it shows that the learning bias that the model embodies is sufficient for acquisition of that phenomenon. If, further, the bias is relatively weak, containing few assumptions and little task specificity, the experiment elucidates the key question by showing that arguments against the model based on its inability to yield the relevant linguistic knowledge are groundless."

Now, note that the learning algorithm must be domain-general (in our formulation of the APS) or weakly-biased and not task-specific (in the formulation of Lappin and Shieber (2007)). The two neural architectures we employed—RNNS and Transformers—were selected

[19] The APS has been formulated, for example by Pullum and Scholz (2002) and Legate and Yang (2002) as being about whether the learning algorithm employs *data-driven* induction or not. We avoid this term here, as data-driven induction can support either nativist or non-nativist positions, as introduced above. For example, connectionist networks (McClelland et al., 1987; Elman, 1990) and unsupervised parsers that assume an underlying Context Free Grammar (Charniak, 1996) are both data-driven algorithms, but, if they are to be taken as models of cognition, make very different commitments about what must be learned.

precisely for this purpose. Furthermore, based on the experiments presented in Section 4 and Section 5, these models demonstrate behavior that is consistent with learning the basic dependency between fillers and gaps, their hierarchical restrictions, and island phenomena (Coordination Islands, Complex NP islands, Wh-Islands and Adjunct Islands for all models; Left Branch, Subject and Sentential Subject for a subset of the models). Therefore, we conclude that their success provides a strong refutation of the APS for islands, and neutralizes it as an argument in favor of linguistic nativism, at least for these structures.

Properly understood, our results bear on questions of nativism not as *argument*, but as *counterargument* against the APS. If we remove the APS from the picture, then what are we left to conclude about the filler–gap dependency and islands? One of the most striking features of syntactic islands is that they appear in language after language and in unrelated languages (Richards, 2001) (see also Phillips (2013) Section 2.1.5 for shorter overview of crosslinguistic variation). For a brief overview of the various claims made about the crosslinguistic distribution of the islands explored here, see Appendix B. If innate, domain-specific biases must be recruited to explain their distribution in the world's languages, then the filler–gap dependency and islands may still stand as good evidence for the nativist position. However, we want to note that relative success of grammatical, processing and discourse-structural accounts for explaining crosslinguistic variation is still very much an area of active research. Grammatical accounts, which have traditionally been linked more strongly to nativist positions, have long sought to ground variation in other properties of the grammar (Bošković, 2005; Richards, 2001), and contemporary accounts have used this approach successfully, for example, for subject/object asymmetries (Stepanov, 2007; Nunes and Uriagereka, 2000; Omaki et al., 2020). However, recent empirical work has found that many reported crosslinguistic differences are eliminated when testing materials are properly controlled (Sprouse et al., 2016; Abeillé et al., 2020). And discourse structural accounts (Ambridge and Goldberg, 2008; Ambridge et al., 2014), potentially grounded in empiricist assumptions, may capture these data just as successfully. Given our results, which demonstrate that English islands can be learned by domain-general, weakly-biased algorithms, we posit that the strongest arguments for or against linguistic nativism will hinge on data about the similarities and differences between languages, rather than their learnability within a given language.

7 Conclusion

As mentioned in the introduction, we believe that one key feature of this paper is its methodological contributions and hope that the methodology deployed here can be extended beyond the case of the filler–gap dependency. The approach taken in this paper involves assessing the capabilities of Artificial Neural Network models by testing them similarly to how one would test a human subject in a psycholinguistic experiment. Constructing test suites that mimic online processing experiments in humans makes it possible to test any model that makes incremental predictions about language, even ones whose internal states are opaque, such as RNNs and Transformers. Furthermore, this method can be used to test learning outcomes over a wide array of syntactic structures. Our tests reveal that these weakly-biased models acquire impressively sophisticated generalizations regarding the filler–gap dependency and island constraints from even a childhood's quantity of linguistic input, though in some cases we find acquisition failures. It is our hope that this method gains traction among psycholinguists studying incremental models of processing, as well as syntacticians who are more concerned with grammatical representations.

References

Abeillé, A., Hemforth, B., Winckel, E., and Gibson, E. (2020). Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition*, 204:104293.

Ambridge, B. and Goldberg, A. E. (2008). The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics*, 19(3):357–389.

Ambridge, B., Pine, J. M., and Lieven, E. V. (2014). Child language acquisition: Why universal grammar doesn't help. *Language*, 90(3):e53–e90.

Arcos-García, Á., Alvarez-Garcia, J. A., and Soria-Morillo, L. M. (2018). Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods. *Neural Networks*, 99:158–165.

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

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017). What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Blevins, T., Levy, O., and Zettlemoyer, L. (2018). Deep rnns encode soft hierarchical syntax. *arXiv preprint arXiv:1805.04218*.

Bošković, Ž. (2005). On the locality of left branch extraction and the structure of np. *Studia linguistica*, 59(1):1–45.

Bošković, Ž. (2015). From the complex np constraint to everything: On deep extractions across categories. *The Linguistic Review*, 32(4):603–669.

Bošković, Ž. (2020). On the coordinate structure constraint and the adjunct condition. *Syntactic architecture and its consequences II*, page 227.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Charniak, E. (1996). *Statistical language learning*. MIT press.

Chaves, R. P. (2020). What don't RNN language models learn about filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics*, 3(1):20–30.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. (1964). *Current issues in linguistic theory*, volume 38. Walter de Gruyter.

Chomsky, N. (1973). Conditions on transformations. *A festschrift for Morris Halle.*

Chomsky, N. (1975). The logical structure of linguistic theory.

Chomsky, N. (1986). *Barriers*, volume 13. MIT press.

Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, 36(1):1–22.

Chowdhury, S. A. and Zamparelli, R. (2018). Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144.

Chowdhury, S. A. and Zamparelli, R. (2019). An lstm adaptation study of (un) grammaticality. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 204–212.

Clark, A. and Lappin, S. (2010). *Linguistic Nativism and the Poverty of the Stimulus.* John Wiley & Sons.

Corver, N. F. M. (1991). The syntax of left branch extractions.

Crain, S. (1985). How can grammars help parsers? *Natural language parsing.*

Cristia, A., Dupoux, E., Gurven, M., and Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: a time allocation study. *Child development*, 90(3):759–773.

Culbertson, J., Smolensky, P., and Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122(3):306–329.

Da Costa, J. K. and Chaves, R. P. (2020). Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. *Proceedings of the Society for Computation in Linguistics*, 3(1):189–198.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Dayal, V. (2016). *Questions*. Oxford.

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

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2-3):195–225.

Engdahl, E. (1982). Restrictions on unbounded dependencies in swedish. *Readings on unbounded dependencies in Scandinavian languages*, pages 151–174.

Engdahl, E. et al. (1997). Relative clause extractions in context. *Working papers in Scandinavian syntax*, 60:51–79.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Franz, A. and Brants, T. (2006). All our n-gram are belong to you. *Google Machine Translation Team*, 20.

Futrell, R., Wilcox, E., Morita, T., and Levy, R. (2018). Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Georgopoulos, C. (1991). *Syntactic variables: Resumptive pronouns and A binding in Palauan*, volume 24. Springer Science & Business Media.

Giulianelli, M., Harding, J., Mohnert, F., Hupkes, D., and Zuidema, W. (2018). Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Goodluck, H., Foley, M., and Sedivy, J. (1992). Adjunct islands and acquisition. In *Island constraints*, pages 181–194. Springer.

Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pages 545–552.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138.*

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Hart, B. and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and Lange, F. P. d. (2021). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv.*

Hewitt, J., Hahn, M., Ganguli, S., Liang, P., and Manning, C. D. (2020). Rnns can generate bounded hierarchical languages with optimal memory. *arXiv preprint arXiv:2010.07515.*

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

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2013). Islands in the grammar? standards of evidence. *Experimental syntax and island effects*, 42.

Hofmeister, P. and Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, 86(2):366.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692.*

Huang, C.-T. J. (1982). Logical relations in chinese and the theory of grammar. phd diss.

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

Kiss, K. É. (2013). *Configurationality in Hungarian*, volume 3. Springer Science & Business Media.

Kuno, S. (1973). Constraints on internal clauses and sentential subjects. *Linguistic Inquiry*, 4(3):363–385.

Kush, D., Omaki, A., and Hornstein, N. (2013). 11 microvariation in islands? *Experimental syntax and island effects*, page 239.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., and Baroni, M. (2019). The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*.

Lappin, S. and Shieber, S. M. (2007). Machine learning theory and practice as a source of insightinto universal grammar. *Journal of Linguistics*, 43(2):393–427.

Laurence, S. and Margolis, E. (2001). The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, 52(2):217–276.

Legate, J. A. and Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 18(1-2):151–162.

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

Levy, R. and Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

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

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Maling, J. M. (1978). An asymmetry with respect to wh-islands. *Linguistic Inquiry*, pages 75–89.

Mameli, M. and Bateson, P. (2011). An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1563):436–443.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank.

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

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1987). *Parallel distributed processing*, volume 2. MIT press Cambridge, MA:.

Nunes, J. and Uriagereka, J. (2000). Cyclicity and extraction domains. *Syntax*, 3(1):20–43.

Oda, H. (2017). Two types of the coordinate structure constraint and rescue by pf deletion. In *Proceedings of the North East Linguistic Society*, volume 47, pages 343–356.

Omaki, A., Fukuda, S., Nakao, C., and Polinsky, M. (2020). Subextraction in japanese and subject-object symmetry. *Natural Language & Linguistic Theory*, 38(2):627–669.

Otsu, Y. (1981). *Universal grammar and syntactic development in children: Toward a theory of syntactic development*. PhD thesis, Massachusetts Institute of Technology.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.

Pearl, L. and Sprouse, J. (2013a). Computational models of acquisition for islands. *Experimental syntax and islands effects*, pages 109–131.

Pearl, L. and Sprouse, J. (2013b). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68.

Pearl, L. S. and Sprouse, J. (2019). Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition: Supplementary material. *Language*, 95(4).

Pereira, F. (2000). Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253.

Perfors, A., Regier, T., and Tenenbaum, J. B. (2006). Poverty of the stimulus? a rational approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.

Phillips, C. (2013). On the nature of island constraints ii: Language learning and innateness. *Experimental syntax and island effects*, pages 132–157.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar.* Center for the Study of Language and Information, Stanford, CA.

Pullum, G. K. and Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The linguistic review*, 18(1-2):9–50.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Ravfogel, S., Goldberg, Y., and Linzen, T. (2019). Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics.

Richards, N. (2001). *Movement in language: Interactions and architectures.* Oxford University Press, USA.

Rizzi, L. (1982). 1982: Issues in italian syntax. dordrecht: Foris.

Rodriguez, P. (2001). Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation*, 13(9):2093–2118.

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

Sabel, J. (2002). A minimalist analysis of syntactic islands.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Schmidhuber, J., Gers, F., and Eck, D. (2002). Learning nonregular languages: A comparison of simple recurrent networks and lstm. *Neural Computation*, 14(9):2039–2041.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. In *Philosophy, Language, and Artificial Intelligence*, pages 79–89. Springer.

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

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

Sprouse, J. and Hornstein, N. (2013). *Experimental syntax and island effects*. Cambridge University Press.

Stepanov, A. (2007). The end of ced? minimalism and extraction domains. *Syntax*, 10(1):80–126.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

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

Van Schijndel, M. and Linzen, T. (2018). A neural model of adaptation in reading. *arXiv preprint arXiv:1808.09930*.

Vig, J. and Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Weber, A., Fernald, A., and Diop, Y. (2017). When cultural norms discourage talking to babies: Effectiveness of a parenting program in rural senegal. *Child Development*, 88(5):1513–1526.

Weiss, G., Goldberg, Y., and Yahav, E. (2018). On the practical computational power of finite precision rnns for language recognition. *arXiv preprint arXiv:1805.04908*.

White, J. C. and Cotterell, R. (2021). Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Wilcox, E., Levy, R., and Futrell, R. (2019). Hierarchical representation in neural language models: Suppression and recovery of expectations. *arXiv preprint arXiv:1906.04068*.

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

Wilcox, E., Vani, P., and Levy, R. P. (2021). A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Yoshida, M. (2006). *Constraints and mechanisms in long-distance dependency formation.* University of Maryland, College Park.

A Quantifying grammatical generalization strength with log-probabilities

If a language model's word predictions reflect implicit grammar-like structural generalizations, then its predictions for a word w given a context C can be productively decomposed as $\sum_X P(w|X, C)P(X|C)$ where X ranges over possible structure-level continuations that might apply in a context C. Consider two contexts that superficially are only minimally different, but that should have large differences in the expectation as to whether there will be a gap, such as "I know {that/what} the lion devoured…". If we take X to be whether a gap happens or not, then for a model making human-like generalizations, $P(X|C_{\text{that}})$ and

$P(X|C_{\text{what}})$ should be very different. But $P(w|X, C_{\text{that}})$ and $P(w|X, C_{\text{what}})$ may be very similar, reflecting word frequencies, relations with preceding individual words, and so forth. Let us denote the possible values of X as x^+ for presence of a gap and x^- for absence of a gap, and denote by w^+ a word that would require a gap, such as *yesterday* in the given example. We make explicit the marginalization over possible X explicit and consider the ratio of conditional word probabilities:

$$\begin{aligned}\frac{P(w^+|C_{\text{what}})}{P(w^+|C_{\text{that}})} &= \frac{\sum_X P(w|X, C_{\text{what}})P(w^+|C_{\text{what}})}{\sum_X P(w^+|X, C_{\text{that}})P(X|C_{\text{that}})} \quad (1)\\ &= \frac{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{what}}) + P(w^+|x^-, C_{\text{what}})P(x^-|C_{\text{what}})}{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{that}}) + P(w^+|x^-, C_{\text{that}})P(x^-|C_{\text{that}})} \quad (2)\end{aligned}$$

For a model that has successfully learned human-like generalizations, $P(w^+|x^-, C_{\text{what}})$ and $P(w^+|x^-, C_{\text{that}})$ should both be extremely small, because gap-requiring words w^+ require gaps (x^+), allowing us to drop the second term of the summands in both the numerator and denominator:

$$\frac{P(w^+|C_{\text{what}})}{P(w^+|C_{\text{that}})} \approx \frac{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{what}})}{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{that}})} \tag{3}$$

Furthermore, we can expect that $P(w^+|x^+, C_{\text{that}}) \approx P(w^+|x^-, C_{\text{what}})$—that is, the presence or absence of a gap is the factor that *mediates* the probability of gap-requiring words across these contexts. This allows us to cancel the first remaining terms in the numerator and denominator, giving us:

$$\frac{P(w^+|C_{\text{what}})}{P(w^+|C_{\text{that}})} \approx \frac{P(x^+|C_{\text{what}})}{P(x^+|C_{\text{that}})} \tag{4}$$

We should have $P(x^+|C_{\text{what}}) \gg P(x^+|C_{\text{that}})$ so in turn the probability ratio for any specific gap-requiring word in the two contexts should be much larger than one. Indeed, one might argue that for truly human-like linguistic generalization, we should have $P(x^+|C_{\text{that}}) \approx 0$ so the larger this ratio, the more human-like the model's behavior. Equivalently, we can take the base-2 log of the inverse of this probability ratio, giving us the difference, measured in bits, of the SURPRISAL VALUES $S(w)$ of the in the two different contexts:

$$\log_2 P(w^+|C_{\text{that}}) - \log_2 P(w^+|C_{\text{what}}) \tag{5}$$

which can range between $(-\infty, \infty)$, with a large positive value indicating human-like linguistic generalization.

Conversely, for material w^- that indicates there is *no* gap (such as *the gazelle*), equivalent logic implies that a *negative* value for the quantity

$$\log_2 P(w^-|C_{\text{that}}) - \log_2 P(w^-|C_{\text{what}}) \tag{6}$$

would be indicative of human-like sensitivity to filler–gap structural relationships. Here, however, the logic of "the larger the more human-like" does not go through: there are generally multiple possible locations for a gap given a filler, so we do not necessarily expect $P(x^-|C_{\text{what}}) \approx 0$ in any particular position.

These differences of log-probabilities in Equations (5) and (6) are the WH-EFFECTS described in Section 3.2 and used throughout the paper in our tests of neural language models' acquisition of human-like filler–gap dependencies.

B Crosslinguistic Distribution of Islands

Table 1 gives a sample of the crosslinguistic variation of islands tested in this paper. This is not intended to be an exhaustive list, rather to highlight some of the important empirical conclusions from the past 50 years. For the purposes of our argument two things are important: First, the same islands appear in multiple unrelated languages, and second, there is variation between languages regarding the status of each island.

Island	Yes Extraction	No Extraction
Adjunct Islands	Korean, Japanese, Malayalam (Yoshida, 2006)	Hungarian, Russian, Spanish, Basque (Yoshida, 2006) English, Italian, Portuguese, French, German (Sprouse and Hornstein, 2013)
Complex NP Islands	Swedish, Danish, Norwegian (Engdahl, 1982; Engdahl et al., 1997), Japanese (contested) (Kuno, 1973)	Serbo-Croatian, Greek, English (Bošković, 2015); Italian, Spanish, Portuguese, French, German, Russian, Hungarian Sprouse and Hornstein (2013)
Coordination Islands	None (but proposed whole conjunct extraction for: Japanese, Korean, Serbo-Croatian, Russian, Old English, Latin (Oda, 2017))	All Languages (Bošković, 2020) (extraction out of conjuncts)
Left Brach Islands	Russian, Polish, Czech (Corver, 1991)	English, Dutch (Corver, 1991)
Sentential Subject Islands	Palauan, Malagasy, Chamorro, Japanese, Akan, Tuki (Sabel, 2002)	English
Subject Islands	Japanese, Navajo, Turkish, Russian (Stepanov, 2007), Hungarian (Kiss, 2013), Palauan (Georgopoulos, 1991)	Norwegian, Italian (Sprouse et al., 2016), Eglnish, French (Sprouse and Hornstein, 2013)
Wh-Islands	Italian (Rizzi, 1982), Spanish, Portuguese (Sprouse and Hornstein, 2013) Scandanavian Langs (Maling, 1978)	English, German, Russian (Sprouse and Hornstein, 2013)

Table 1: Some Crosslinguistic Variation of Islands