Using Computational Models to Test Syntactic Learnability

Citation: Wilcox, Ethan Gotlieb, Futrell, Richard and Levy, Roger. 2022. "Using Computational Models to Test Syntactic Learnability." Linguistic Inquiry.

As Published: 10.1162/LING_A_00491

Publisher: MIT Press

Persistent URL: https://hdl.handle.net/1721.1/150009

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

Using Computational Models to Test Syntactic Learnability

Ethan Gotlieb Wilcox, Richard Futrell and Roger Levy

September 2, 2022

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–

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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).[1] 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 only indirect 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—variously characterized 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; Yang and Piantadosi, 2022). 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, 1979); 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), roadside signs (Arcos-García et al., 2018) and the structure of proteins (Rives et al., 2021). 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, 2019)), we inspect the word-by-word predictions of language models in controlled, factorized tests that are designed to draw out what generalizations they have made about individual structural phenomena. For this work, we base all of our tests on an experimental design common in psycholinguistics, where test sentences take on structural properties based on two 'crossed' factors, which are inspired by two separate predictions made by the grammar about the filler–gap dependency. The first prediction stems from the fact that gaps require fillers to be properly licensed. If we take a sentences such as (2-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 (2-b). (In this and subsequent examples the underscores are for presentational purposes only, and are not included in test items.) This grammaticality contrast holds only if the matrix verb is obligatorily transitive: "I know that the lion ate yesterday" is grammatical.

(2) 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 inspect the model's predictions in the adverbial phrase modifier which occurs after the matrix verb, such as "*yesterday*" in (2). 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 have a higher probability in the context of a filler than in its absence. That is, "yesterday" should be have a lower probability in (2-b) than in (2-a). Because this prediction is about the effect of an upstream wh-word when a gap is present, we call it the WH-EFFECT in a +GAP context.

The second prediction made by the grammar is that upstream fillers require downstream gaps. If we take the grammatical sentence (3-a) and replace the that-complementizer with a wh-complementizer as in (3-b), then the rule is violated and the sentence becomes ungrammatical.

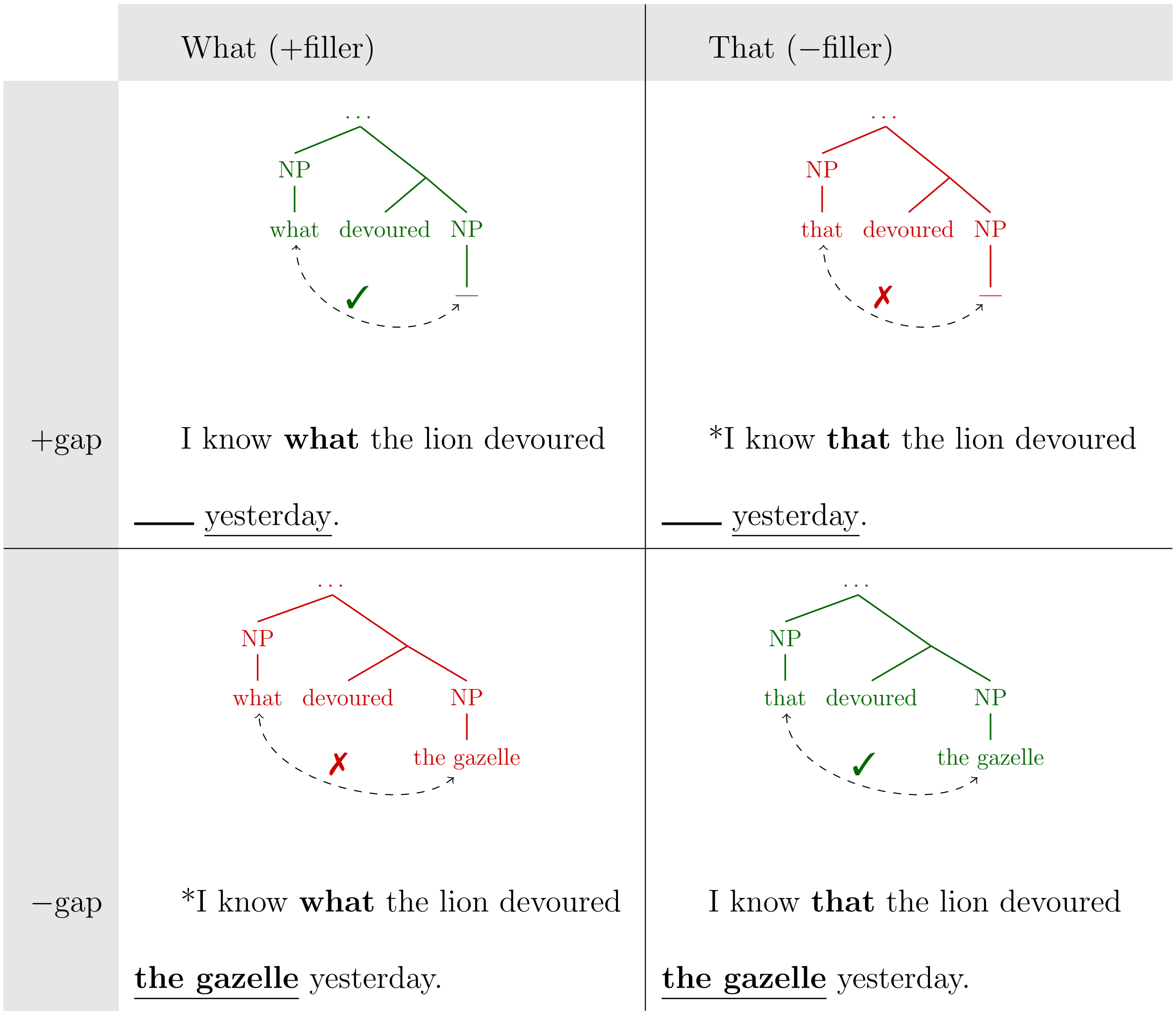
(3) a. I know that the lion devoured the gazelle yesterday.

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 (3-b) than in (3-a).

This effect has been observed in humans and is called the filled-gap effect (Crain and Fodor, 1985; Stowe, 1986). We call this difference the WH-EFFECT in a −GAP context.[2] In order to test that the model has learned both of these constraints simultaneously, we combine these two predictions and look at the interaction between fillers and gaps, a schematic of which is given in Figure 1. If the correct generalization about filler–gap dependencies has been made, we predict that there should be interaction between the presence of a filler and the presence of a gap, whereby target sentence regions (either filled argument structure or post-gap material) are less surprising when both are present than when just one is present.

So far we have focused on measuring the filler–gap dependency in structural positions where it is licensed by the grammar. But more important for our theoretical objective is whether language models learn the restrictions on this dependency, the so-called island constraints. If models are learning that the co-variation between fillers and gaps does not 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. Second, the presence or absence of an upstream filler should not affect the relative surprisal of a filled argument structure position located inside an island, either. Put another way, the wh-effects exemplified by the contrasts (2) and (3) should disappear, and we would not expect an interactive relationship between the presence of a filler and the presence of a gap. In practice, we use two metrics to determine whether a model is showing sensitivity to island constraints: an *absolute metric*, that asks whether wh-effects are at or very close to zero inside island configurations, and a *relative metric*, that inspects the three-way interaction between fillers, gaps, and islands to test whether models are more surprised at gaps inside islands compared to non island-violating



	What (+filler)	That (−filler)
+gap	I know **what** the lion devoured ___ yesterday.	*I know **that** the lion devoured ___ yesterday.
−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.

minimal-pair sentences.

We focus on seven of the most well-studied islands associated with syntactic constraints, identified either by Ross (1967) or Huang (1982). They are:

(4) a. Adjunct Islands

*I know what the patron got mad after the librarian placed ___ on the wrong shelf.

b. Complex NP Islands

*I know what the actress bought the painting that depicted ___ yesterday.

c. Adjunct Islands

*I know what the man bought and ___ at the antique shop.

d. Left Branch Islands

*I know how expensive you bought ___ car last week.

e. Sentential Subject Islands

*I know who for the seniors to defeat ___ will be trivial.

f. Subject Islands

*I know who the painting by ___ fetched a high price.

g. Wh Islands

*I know who Alex said whether your friend insulted ___ yesterday.

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 many 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, 1965a). Section 7 concludes.

2 Background: Modeling Syntactic Islands

Island constraints were first introduced by Ross (1967), who analyzed them as a grammatical phenomenon, specifically barriers to syntactic movement ("islands" being places from which it is difficult to move).[3] Although syntactic islands refer to a series of structural positions that block multiple forms of movement (or, under non-movement approaches, multiple types of dependencies), our focus here is on the filler–gap dependency, although we discuss implications for a broader range of dependencies in Section 6.2. Since their identification, islands 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. Within syntactic theory, many of the most influential proposals for innate, universal constraints have been developed to explain island effects (Chomsky, 1965b, 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.

The filler–gap dependency and its constraints have been studied extensively as a processing phenomenon, as well. One key finding, here, is that after processing a filler, comprehenders expect to find gaps subsequently and are surprised if they encounter filled argument structures instead (the so-called "filled gap" effect). The filled-gap effect was initially identified with increased reading times at NPs and pronouns in English (Crain and Fodor, 1985; Stowe, 1986), and has since been replicated in experiments using different modalities and processing measures (Phillips, 2006; Garnsey et al., 1989; Kaan et al., 2000; Phillips et al., 2005; Traxler and Pickering, 1996), and in different languages (Frazier, 1987; Aoshima et al., 2004; Sekerina, 2003; Schlesewsky et al., 2000). The filled-gap effect is the inspiration behind our WH-EFFECT in the *-gap* condition. As far as we are aware, this paper is the first to propose an analogous effect in *+gap* conditions, although we do not test it experimentally with human subjects. A number of studies have investigated filled-gap effects inside islands, beginning with Stowe (1986) (Experiment 2), who found no increased reading times for pronouns within Complex NPs. Similar studies have shown that subject-verb plausibility relations can affect the Filler–Gap dependency, but that these, too, are attenuated inside island constructions (Traxler and Pickering, 1996; Phillips and Wagers, 2007) (although see Pickering et al. (1994) and Clifton and Frazier (1989) for data that complicates this picture).

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 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 (5).

(5) $[_{CP}$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.[4] 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).[5] 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.[6] One additional point of criticism is that 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 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) assess how well the probabilities assigned to strings by a Recurrent Neural Network language model compare to human acceptability judgements, including for some island phenomena. Instead of looking at sentence regions, they look at whole sentence probabilities, comparing target island sentences to non-island control sentences. They demonstrate that models assign less total-sentence probability to sentences that violate islands, like (6-a) compared to non-question affirmative sentences like (6-b), suggesting that models might have acquired some sensitivity to islands. However, they complicate the story by reporting total-level probability on yes-no question sentences such as (6-c). Even though these sentences have no wh-movement and cannot violate islands, the model assigns them probability in between the island-violating sentences and the assertion control sentences. Chowdhury and Zamparelli conclude that the models have not learned generalized rules, but are sensitive to the "cumulative effect of increasing syntactic complexity, plus position." A follow-up study (Chowdhury and Zamparelli, 2019) finds similar apparent island sensitivity in relative clauses, but again complicates the story by showing similar sensitivity in control conditions, where none would be expected under the hypothesis that they had learned island constraints.

(6) a. What did you see the lion that caught ___ ? [ASSERTION]

b. You saw the lion that caught the gazelle. [WH-QUESTION]

c. Did you see the lion that caught the gazelle ? [Y/N-QUESTION]

While these results are interesting and demand further investigation, we believe that the equation of whole-sentence probability and grammaticality has led the authors to too hasty a conclusion. Just because models show sentence-level probability for yes/no questions that is closer to island-violating sentences than assertions, does not mean that they 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 widely-used psycholinguistic methods, we call this the "psycholinguistics paradigm" for model assessment. Below, we explain how we deploy it on filler–gap dependency sentences and introduce our five models.[7]

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 $x_1 \ldots x_{i-1}$ by providing a distribution over $P(x_i \mid x_1 \ldots x_{i-1})$.[8] 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 designed to test a particular grammatical phenomenon. Models assign probabilities to each word in the test sentences, and these probabilities are used to assess whether the model has successfully learned the phenomenon. 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 phenomenon, 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 critical regions, which itself is the same in both conditions. 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.

One reason why we believe that the psycholinguistic assessment of grammar presents a productive step forward for theoretical linguists is that it is a way to assess the grammatical generalizations made by contemporary Artificial Neural Network (ANN) models while avoiding previous debates about what function, if any, relates probability to grammaticality. By constructing carefully controlled items, we create tests where certain patterns of probability can only be produced if models have in some sense learned the relevant grammatical generalizations. For a technical explanation of this intuition, as well as to explain how we measure our target grammatical phenomenon, we turn to the filler–gap dependency in the next section.

3.2 Measuring the Filler–Gap Dependency

Our goal is to understand the generalization that the models have made about latent structural properties, namely whether a sentence contains a gap or not. The particular generalization of interest is whether or not the model has learned that gaps are grammatical in contexts with fillers, and ungrammatical in contexts without fillers (exemplified by the contrast between (7-a) and (7-b), below); and likewise, that sentences without gaps are ungrammatical in the presence of fillers (exemplified by the contrast between (7-c) and (7-d)).

(7) a. I know what the lion devoured ___ yesterday .

b. *I know that the lion devoured ___ yesterday .

c. I know that the lion devoured the gazelle yesterday.

d. *I know what the lion devoured the gazelle yesterday.

If the model is following grammatical generalizations as humans presumably do, then it should assign higher probability to a word or phrase when that word or phrase is grammatically licensed than when it is not. For example, the adverb "yesterday" should be less surprising in the context of (7-a) where the object of "devoured" may be reasonably expected to have been extracted, compared with (7-b), where the word "yesterday" seems to indicate that the mandatory object of "devoured" is missing even though there is no filler–gap construction to license this.

We can formalize this idea in terms of the ratio of two conditional probabilities: the conditional probability of a word w^+ given a context that licenses a gap, which we notate as $P(w^+ \mid C_{\text{what}})$, and the probability of that same word w^+ given a context that does not license a gap, notated as $P(w^+ \mid C_{\text{that}})$. Here w^+ is what we call a *gap-requiring word*: its presence is only grammatically possible in this context if there is a filler–gap construction. In (7-a)–(7-b), the gap-requiring word is "yesterday" which indicates a gap in the the

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

object position of "devoured". Under reasonable conditions,[9] the *ratio* of these two probabilities should be approximately equal to the probability ratio of the *filler–gap structure itself* given the two contexts:

$$\frac{P(x^+ \mid C_{\text{what}})}{P(x^+ \mid C_{\text{that}})} \approx \frac{P(w^+ \mid C_{\text{what}})}{P(w^+ \mid C_{\text{that}})}, \tag{1}$$

where x^+ represents the existence of a filler–gap construction, so that $P(x^+ \mid C_{\text{what}})$ can be interpreted as the probability assigned by the model to the presence of a filler–gap construction given the preceding context C_{what}.

What this approximate equality means is that the grammatical generalizations that the network has learned should be reflected in word-level probabilities. Specifically, because $P(x^+ \mid C_{\text{what}}) \gg P(x^+ \mid C_{\text{that}})$ in our materials, 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 $P(x^+ \mid C_{\text{that}}) \approx 0$ so the larger this ratio, the more human-like the model's behavior.

A simple transformation of this probability ratio of Eq. (1) deepens its interpretability, namely taking its negative log:

$$-\log_2 \frac{P(x^+ \mid C_{\text{what}})}{P(x^+ \mid C_{\text{that}})} \approx -\log_2 \frac{P(w^+ \mid C_{\text{what}})}{P(w^+ \mid C_{\text{that}})} = -\log_2 P(w^+ \mid C_{\text{what}}) + \log_2 P(w^+ \mid C_{\text{that}}). \tag{2}$$

A negative log probability is also known as a SURPRISAL value, and lies in the range range $[0, \infty)$. A probability 1 event has zero surprisal; as probability decreases, surprisal increases asymptotically toward infinity as probability approaches zero. Introducing the notation $S(y \mid C) \equiv -\log_2 P(y \mid C)$, we can rewrite the right-hand side of Eq. (2) as simply:

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

$$S(w^+ \mid C_{\text{what}}) - S(w^+ \mid C_{\text{that}}). \tag{3}$$

The difference in Equation 3 lies in the range $(-\infty, \infty)$. This value should be *negative* for human-like generalization: the gap-requiring word w^+ should be low-surprisal in the context C_{what} reflecting the fact that the negative log transformation is a monotonically decreasing function.

In fact, we can go further and say that for a well-designed example stimulus where omitting the phrase in the gap position is ungrammatical without an appropriately positioned filler in the context, the smaller this value (i.e., the closer to $-\infty$) the more human-like the generalization, because we should have $P(x^+ \mid C_{\text{that}}) \approx 0$ and $P(x^+ \mid C_{\text{what}}) > 0$.

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

$$S(w^- \mid C_{\text{what}}) - S(w^- \mid C_{\text{that}}). \tag{4}$$

would be indicative of human-like sensitivity to filler–gap structural relationships, because we should have $P(w^-|C_{\text{what}}) < P(w^-|C_{\text{that}})$. 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 $P(w^-|C_{\text{what}}) \approx 0$ in any particular position.

In the introduction, we informally introduced the notion of WH-EFFECTS, which are measures of how well models have learned the target structural generalizations. These differences in surprisal values in Equations (3) and (4) are the technical definitions of WH-effects that will be used throughout the paper. In general, we expect large positive wh-effects in cases where there is no gap, and large *negative* wh-effects in cases where there is a gap.

The formulation in terms of surprisal also offers a link to psycholinguistic theory and data, where surprisal is well established to influence both reading times and brain responses (Hale, 2001; Levy, 2008; Smith and Levy, 2013; Frank et al., 2015; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Heilbron et al., 2021). Although it is not the focus of the present contribution, we also believe this link also facilitates direct quantitative comparisons of model predictions and human responses, such as those described in van Schijndel and Linzen (2018) and Wilcox et al. (2021).

To test that models have learned both wh-effects simultaneously, we predict that, for the full 2 × 2 crossed contrast given in (7) there should be an interaction between the presence of a filler and the presence of a gap, whereby the models should display super-additively 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.[10] 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.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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: 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 Kneser-Ney smoothing trained on the British National Corpus using the SRILM language modeling toolkit (Stolcke, 2002). We present

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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. Neural networks of sufficient depth are universal function approximators, capable of representing arbitrarily complex relationships between input and output given large enough vector representations (Hornik et al., 1989). They can learn a class of logical relationships impossible for simpler learning algorithms, such as 'exclusive or' (Rumelhart et al., 1986). While ANNs have typically been described as 'black box' models, much is now known about their formal representational capabilities, and we touch on these briefly below.

Recurrent Neural Networks (RNNs) (JRNN, GRNN) are a type of ANN that was 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, in practice, SRNs have difficulty learning and maintaining dependencies that involve long linear distances (Rodriguez, 2001).[11] 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). 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., 2014) and has two hidden layers with 8192 units each. It uses the output of a character-level Convolutional Neural Network as input to the LSTM. The second model, originally presented by Gulordava et al. (2018), which we thus we refer to as GRNN, was selected for its previous success at learning number dependencies between subjects and verbs, 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 these models' ability to represent the underlying mathematical formalism assumed by theoretical linguists (Shieber, 1985). We use two Transformer models: GPT-2, which was trained on $\sim$ 8 billion tokens of internet text (Radford et al., 2019),[12] and its successor GPT-3, which was trained on $\sim$ 114 billion words broken into $\sim$ 500 billion subword tokens (Brown et al., 2020).[13]

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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 (8). 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 transitive verbs and balance filler type (*who* vs. *what*) across the items. Bolded text indicates critical regions.

(8) 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, or equivalently log inverse-probabilities). Figure 2 shows these raw surprisal values for the JRNN model. Higher values mean the tokens are less likely under the language model's distribution. Average surprisal starts off at around 10 bits/token in the first region, with surprisal in the +filler condition greater than in the −filler condition (that is, the purple lines are higher than the green lines). This is

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

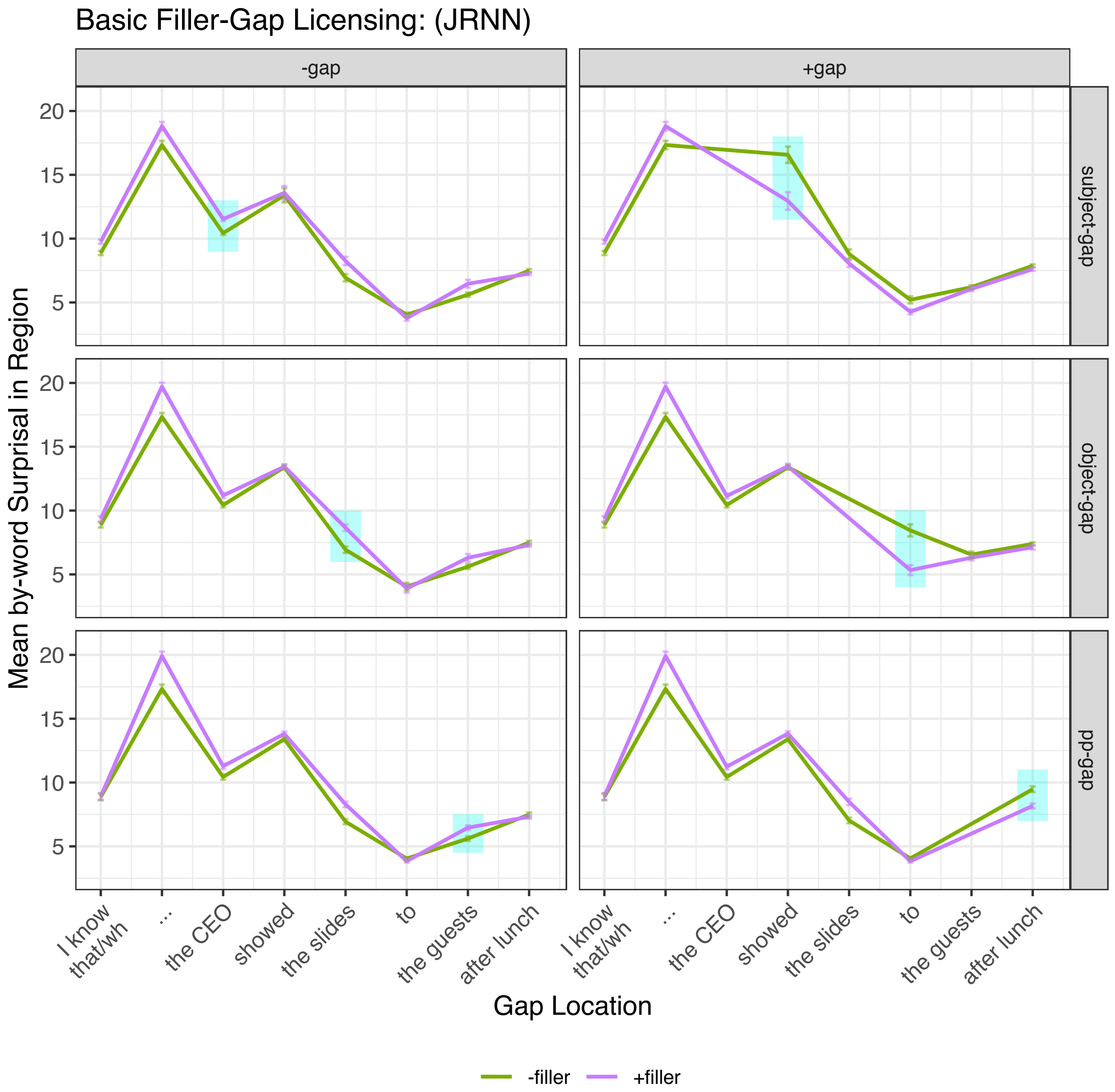


Figure 2: **Filler–Gap Licensing:** Mean by-word surprisal by sentence region for the JRNN model, with critical regions highlighted in teal. In the absence of a gap (left column) +*filler* conditions hare more surprising. When gaps are present (right column) this trend is reversed, corresponding to our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

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. Here, the red 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. This corresponds to the contrast in (2), 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*). This corresponds to the contrast in (3), 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-

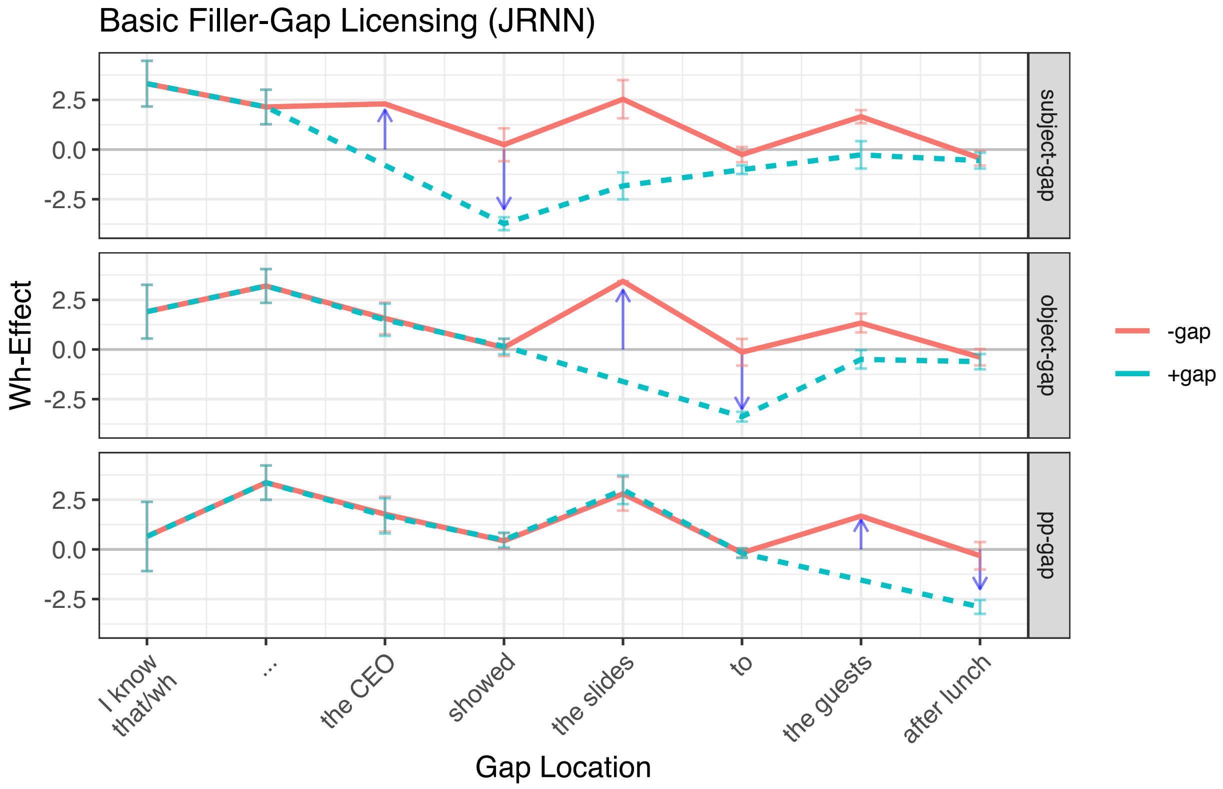


Figure 3: **Wh-effect by sentence region:**, Blue arrows indicate regions of interest in each condition. In the −gap conditions, wh-effects are positive in critical regions. In the +gap conditions, wh-effects are negative. Error bars are 95% confidence intervals across 63 test sentences.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

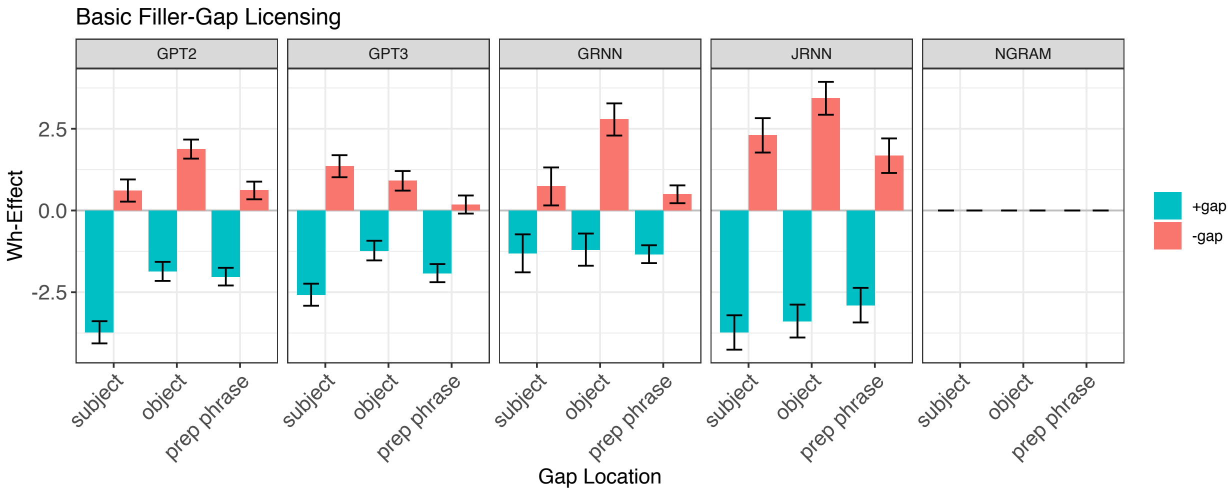


Figure 4: **Basic Licensing:** The various experiments are on the x-axis, and the wh-effect is on the y-axis. Error bars represent 95% confidence intervals. Negative wh-effect in the +*gap* condition (blue bars) and a positive wh-effect in the -*gap* condition (red bars) shows that all models are learning the basic filler–gap dependency.

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. 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

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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.

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.[14] But contrary to the distribution of gaps in written English, GPT-2, JRNN and GRNN show the strongest 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, at least for the former models, behavior is a result of learned generalizations that goes beyond rote recapitulation of training-corpus statistics.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

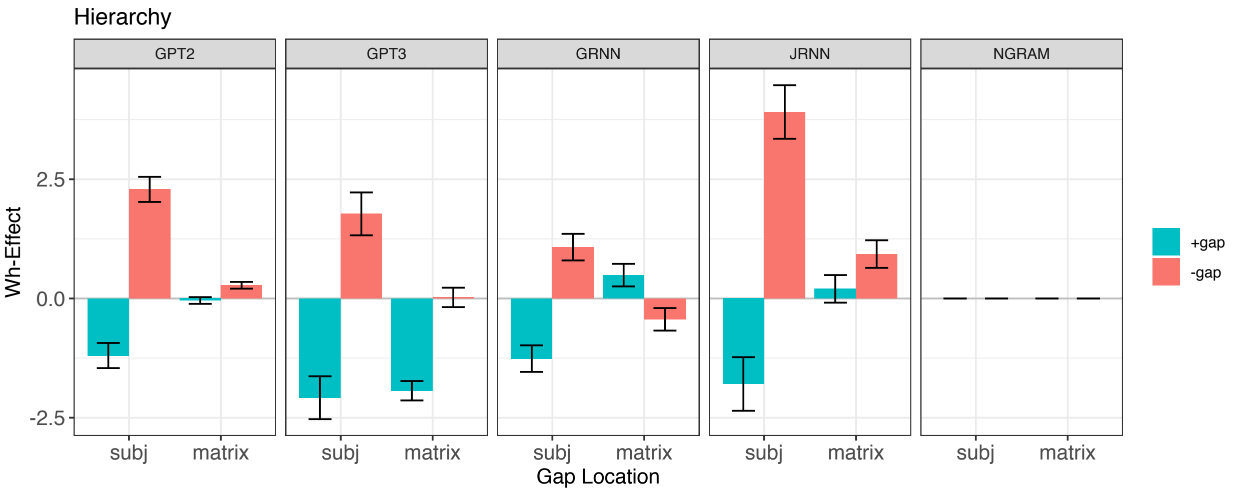


Figure 5: Sensitivity to syntactic hierarchy. All models show reduction (or inversion) of wh-effects in the *matrix* condition.

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, 1994a). Example (9) 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 (9-b) is ungrammatical because the gap site is not in this region. (One other difference between these sentences is the distance between the filler and the gap, an issue we will address in the next section.)

(9) a. The fact that the reporter knows who [the witness surprised ___ **with his testimony**]

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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 (9-a), but close to zero in the ungrammatical conditions like (9-b). In practice we test this expected reduction in two ways: First, we test the *relative* reduction by inspecting the three-way interaction between filler, gaps and structural position, looking for a significantly positive interaction term which indicates increased surprisal in the *matrix gap* condition. Second, we test the *absolute* reduction by inspecting whether the 95% confidence intervals (CIs) of the wh-effects in the *matrix gap* condition cross the zero line (or if there is an inversion of wh-effects between conditions).

In order to test this prediction we created 45 sentences following the template in (9). 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, many show a re-

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

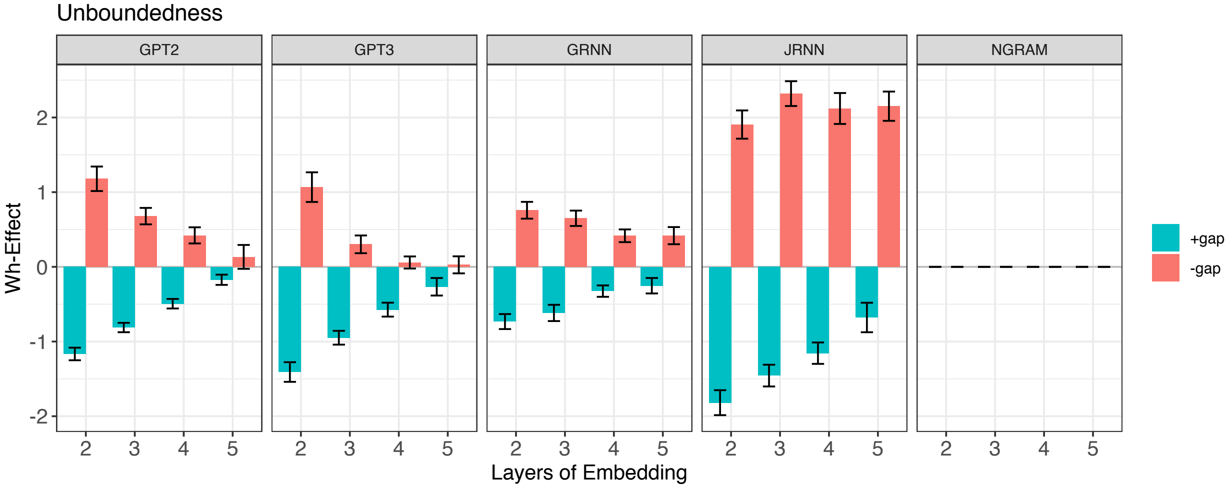


Figure 6: Undboundedness of the Filler–Gap Dependency

duction of wh-effects in the *matrix* condition. For our *relative* metric, we find significant reduction in wh-effects when the gap is no longer *c*-commanded by the filler for all neural models ($p < 0.001$ for JRNN and GPT-2, and $p < 0.05$ for GRNN and GPT-3). However, for GPT-3, this is largely due to reduction in wh-effects in the -gap condition, and behavior in the +gap condition is not consistent with it having learned the correct hierarchical generalizations. For our absolute metric, we find that the 95% CIs cross zero for GPT-2 and JRNN (+gap conditions), and GPT-3 and GRNN (-gap conditions). For GRNN, the 95% CIs do not cross zero, but the effect is in the wrong direction. 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 (10), where the final sentence may be difficult to process, but (arguably) remains grammatical despite the intervening material. Importantly, the grammaticality of sentences like (10-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 and 9 contain 4 layers of sentential embedding.[15]

(10) 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 (10) with between 2 and 5 layers of sentential embedding.[16] 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

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

the reduction is significant after two additional layers ($p < 0.001$); and for the JRNN the reduction is significant after three additional layers ($p < 0.01$), largely driven by reduction in wh-effects in the +gap conditions. As far as absolute licensing, the CIs cross zero only three cases—+gap conditions for GPT-2 at 5 layers of embedding and GPT-3 at 4 and 5 layers of embedding.

Next, in order to be sure that the reduction in our Hierarchy experiment is greater than the reduction observed across multiple layers of embedding, 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* conditions are proximal, whereas *matrix* and *embed4/5* are distal). We treat the filler-to-gap distance as a categorical variable due to the categorical structure of the hierarchy test suites, essentially asking if the prepositional phrase and tensed verb that delineate the end of the relative clause in the hierarchy test is a stronger "blocker" for the filler–gap dependency than a third layer of sentential embedding in the hierarchy test. Here, a negative four-way interaction between *filler*, *gap*, *distal* and *test type* provides a 'yes' answer: 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.[17]. Second, the fact that models are likely to have seen very few sentences with four

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

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. Even though GRNN did show some reduction with additional layers of embedding, RNN models tended to perform better than Transformers, despite their smaller training vocabularies. This suggests that their architecture makes them better models of human linguistic competence, at least for this phenomenon.[18]

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 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. 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. To test these predictions we deploy both a *relative* assessment metric and an *absolute* assessment metric. For our relative metric, we inspect the three-way interaction between +/−island, +/−filler and +/−gap looking for a supperadditive *increase* in surprisal. For our absolute metric, we inspect whether the 95% confidence intervals for the across-item wh-licensing cross the zero line in the island conditions (or invert between island/non-island conditions). 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, typically with gaps in object position.[19]

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 (11), below.[20] If models have learned the gender dependency, then we expect a strong masculine 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.

(11) 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 (12), below.

(12) 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 (12). We test for significant reduction between the control conditions, like (11) 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 (13-a) vs. (13-b).

(13) 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.]

d. The actress got mad after they insulted {**his/her**} performance. [island, gender

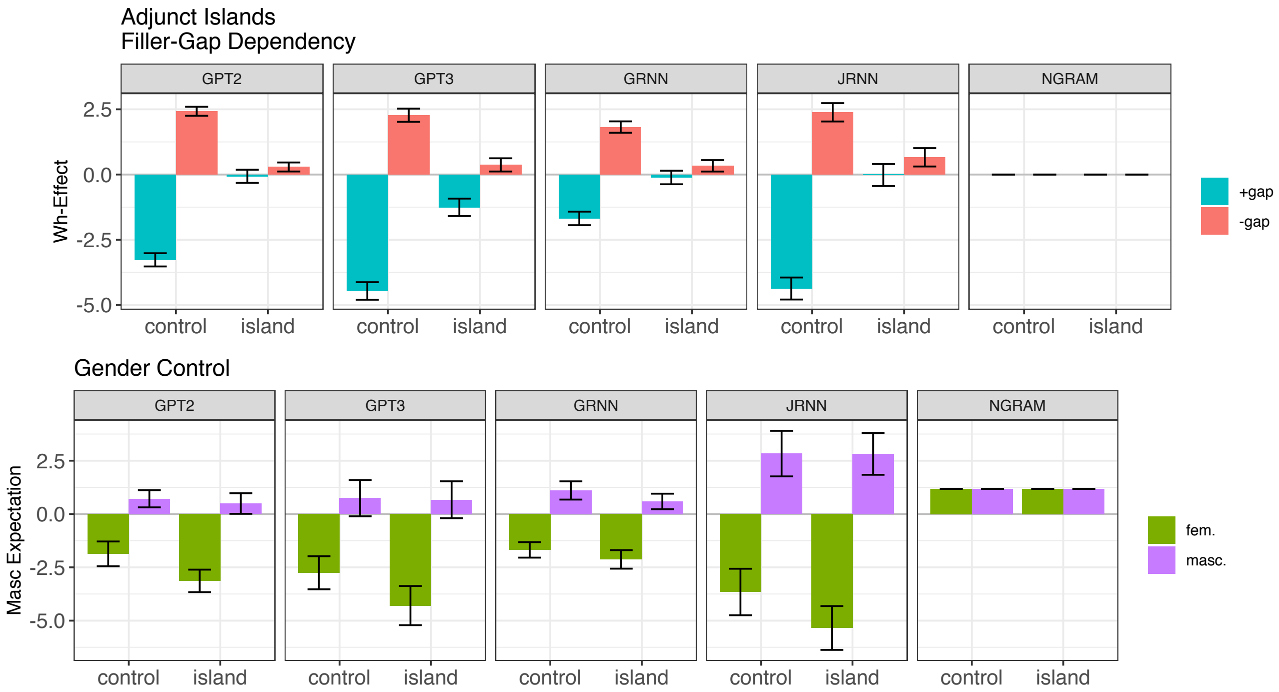


Figure 7: Adjunct Islands

EXP.]

To test for sensitivity to adjunct islands we created 54 items following the template in (13). The results for this experiment can be seen in Figure 7. For the gender dependency, we see the neural models performing exactly the same in the control conditions as in the island conditions. Turning to the filler–gap dependency, we find a relative 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. In terms of absolute metrics, we find that wh-effects are not different from zero for three models in the +gap condition (GPT-2, GRNN and JRNN), but are slightly above zero in the −gap condition for all models.

Complex NP Islands: Gaps are not licensed inside S-nodes that are dominated by a lexical head noun, as demonstrated by the contrast between (14-c) compared to (14-a).

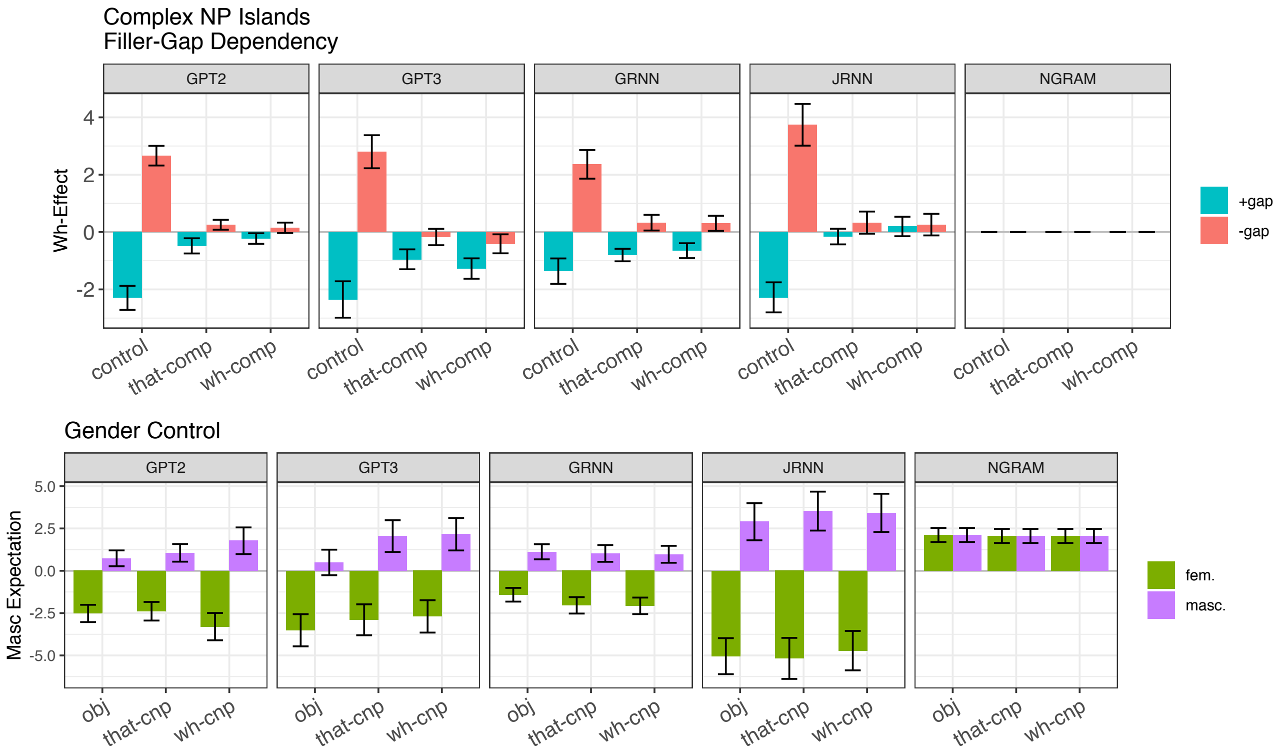


Figure 8: Complex NP Islands

(14) 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, 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 (14), 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

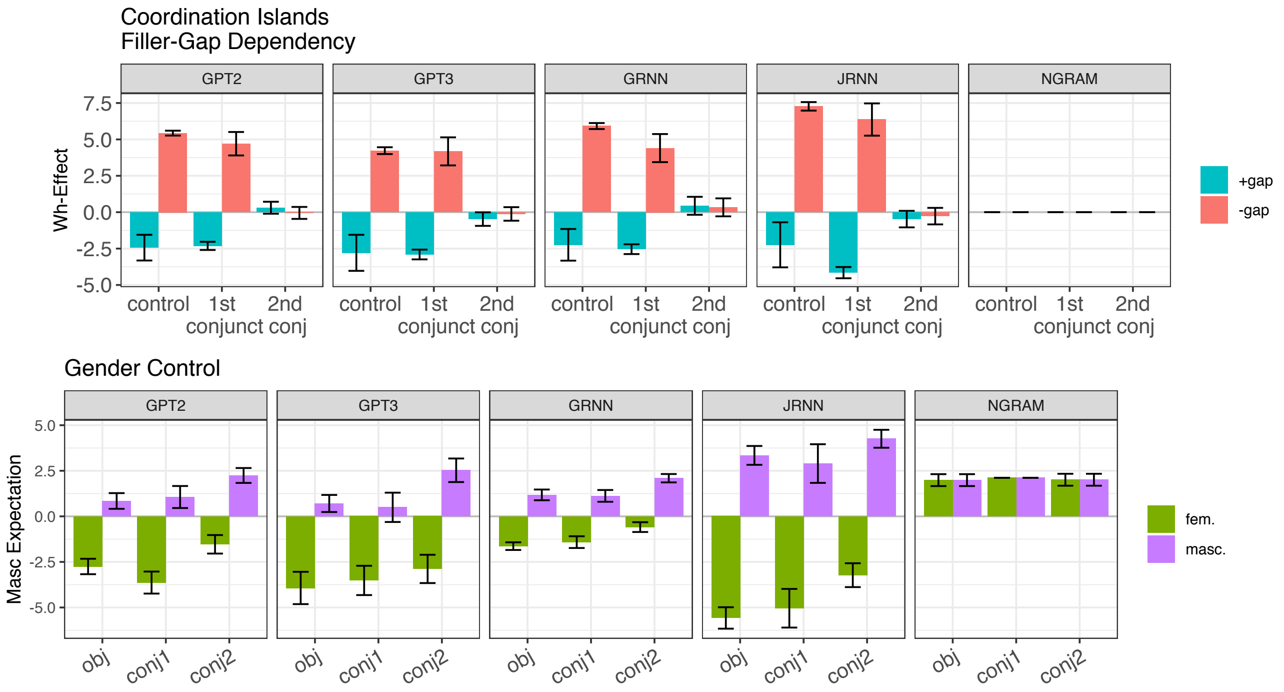


Figure 9: Coordination Islands

find a relative 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). For our absolute metric, we find that 95% CIs cross zero in the +gap condition for JRNN and GPT2 (wh-comp), and for JRNN (that-comp). For the −gap condition, 95% CIs cross zero for all models (wh-comp) and for JRNN (that-comp). 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 (15-b)/(15-c) and (15-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 end grammatically, such as *I know what the man bought __ and the painting*

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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.

(15) 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 (15). In the *-gap* conditions, for the *control* sentences, we sum over a whole conjoined NP (e.g. "the painting and the chair"). In addition, we add extra material to the beginnings 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. In absolute terms, 95% CIs cross the zero line for all neural models as well. The effects here were the largest effect sizes out of all the island configurations tested.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

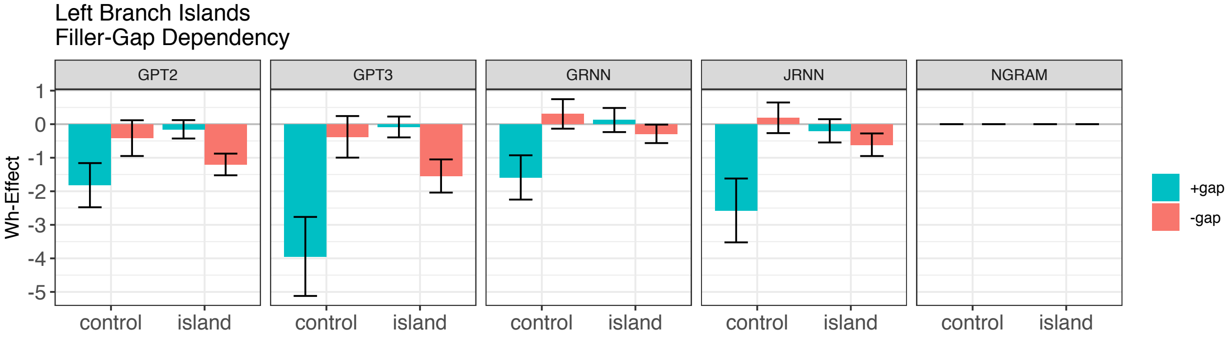


Figure 10: Left Branch Islands

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 (17-d) compared to (16-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.

(16) 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]

(17) 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.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

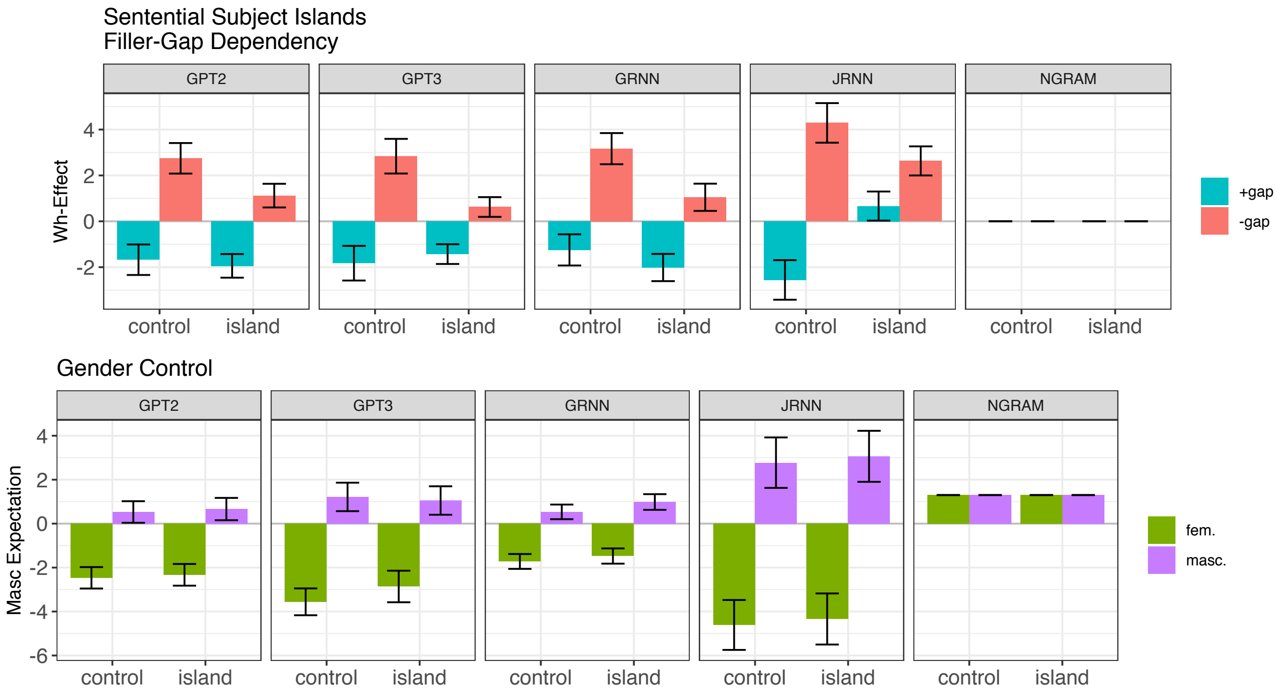


Figure 11: Sentential Subject Islands

d. *I know how expensive you bought __ **a car last week**. [LEFT BRANCH]

The results can be seen in Figure 10. We find that the JRNN, GRNN and GPT-3 models 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 ($p = 0.12$). In terms of absolute licensing, wh-effects are not different from zero for all models in the +gap condition, and all *below zero* in the −gap condition.

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 (18).

(18) 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]

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

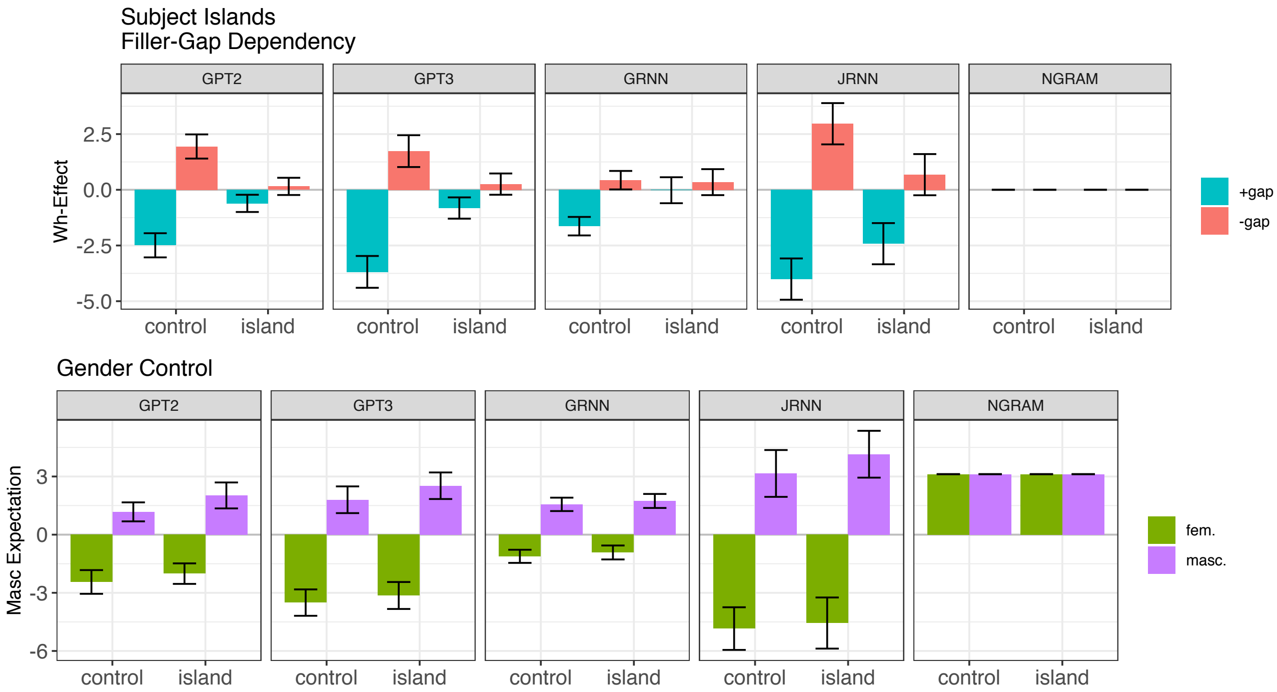


Figure 12: Subject Islands

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. In absolute terms, each model shows a non-zero wh-effect in both +gap and −gap conditions. 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 (19).

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

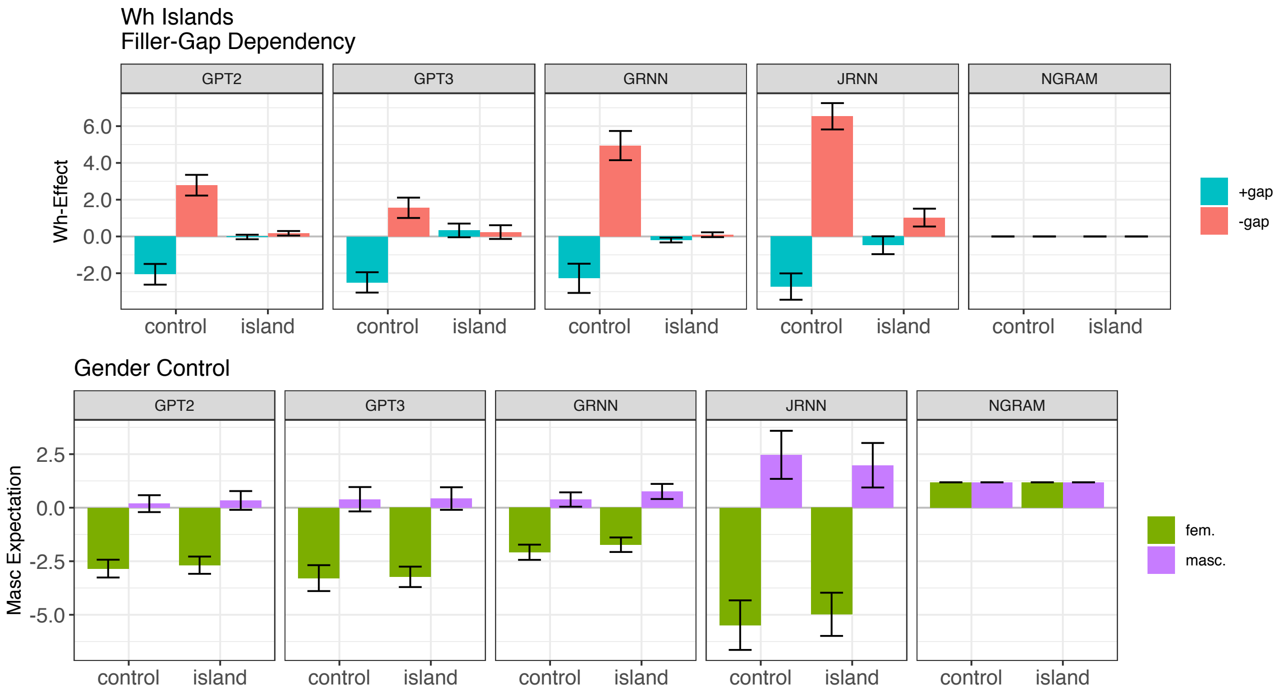


Figure 13: Wh-Island Effects

(19) 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. In absolute terms, we find wh-effects that are not different from zero for all models in the −gap island condition, but are different from zero for all models but GRNN in the +gap island condition.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

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 (20-b) compared to (20-a). We created experimental items following the examples in (20) and measured their wh-effects.

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

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

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

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

The results for this experiment can be seen in Figure 13. We find a relative reduction in licensing effects for all four neural models ($p < 0.001$), with no reduction for the n-gram model. 95% CIs cross zero in the +gap condition for all models, as well as for all models except JRNN in the −gap condition. We find no difference between the island and non-island conditions in the gender controls.

5.3 Discussion

Figures 14 and 15 show summary results for our island tests. Figure 14 highlights model performance on our *absolute* metric by showing the percent reduction in wh-effects between island and non-island conditions. Figure 15 highlights model performance on our *relative* metric, by reporting significance of the three-way interaction term from our statistical tests. For both summary figures, we compare model behavior on island tests to

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

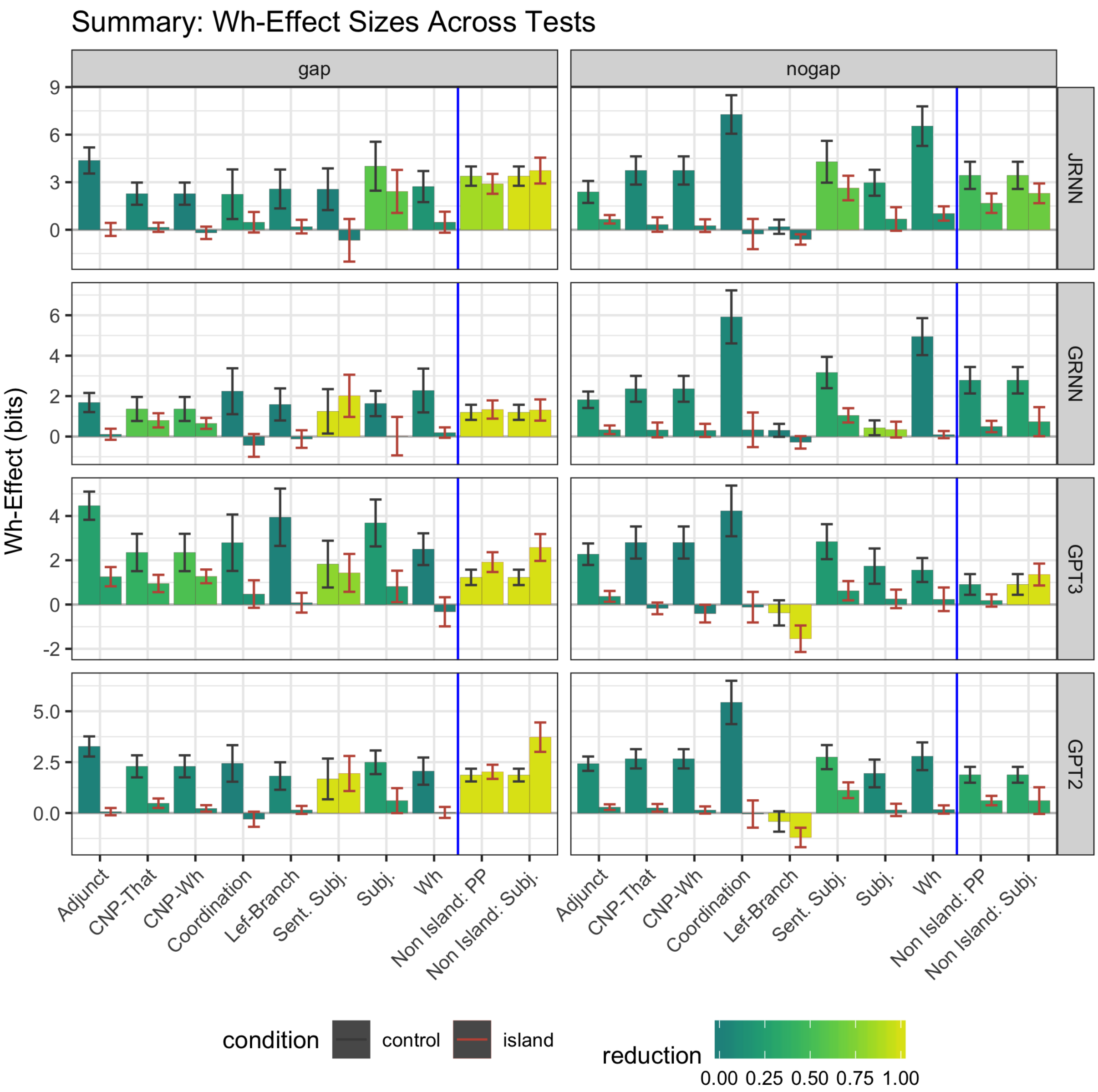


Figure 14: **Summary of results for island tests:** Bars with black 95% CIs are wh-effects for control conditions, bars with red 95% CIs are wh-effects for island conditions. Color indicates the percent reduction in wh-effect between island/non-island conditions. If models are learning islands, we expect substantial reduction, and wh-effects that are not different from zero for islands (i.e. red CIs cross the zero line). To the right of the vertical blue line, we show effects for argument-structure controls, where we expect no decrease in wh-effects between conditions.

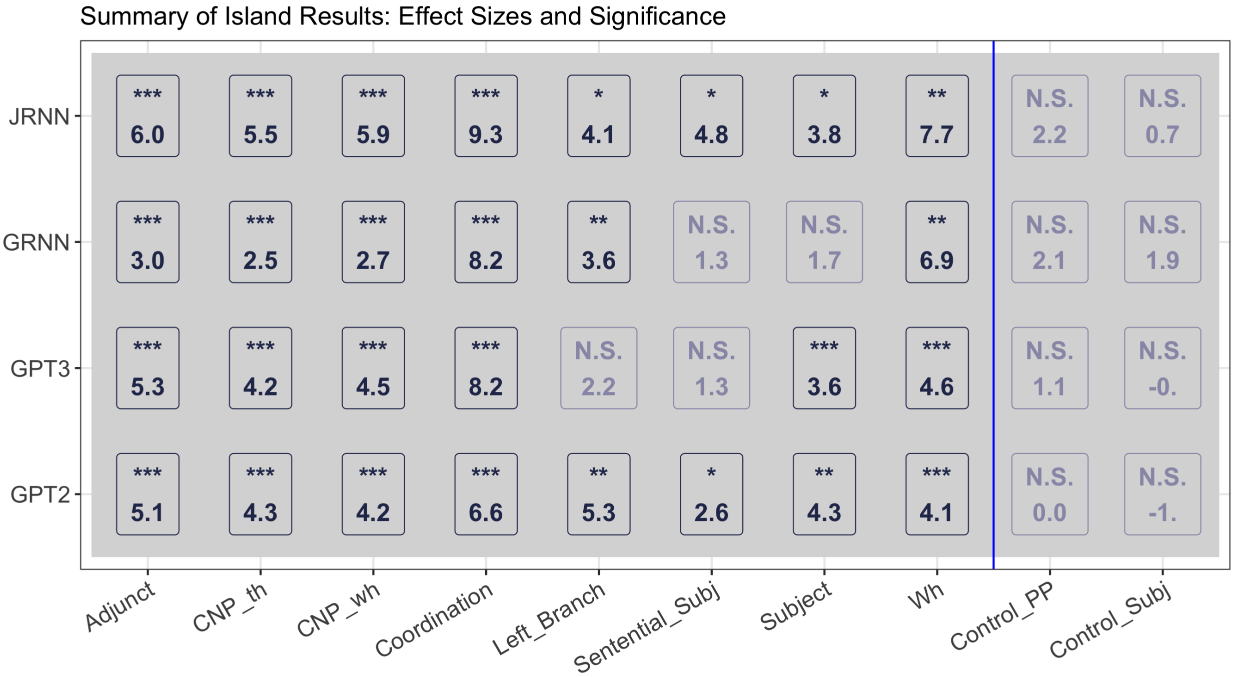


Figure 15: **Reduction of Island Effects:** Each cell corresponds to a statistical test and shows, for the three-way interaction term, the level of significance achieved (top row), and the effect size in bits (bottom row). Tests that reach 0.05 confidence threshold are dark blue. To the right of the vertical blue line are argument structure controls, where we expect no significant effects.

behavior on argument structure tests reported in section 4. As with island tests, gaps in matrix-level object position are controls while gaps in subject position and in prepositional phrases are targets. If models are learning the correct generalizations, we expect these tests to pattern differently from islands.

Looking first at Figure 14, we find that our absolute criteria for island effects is met in 42/64 islands (and 1/8 argument structure controls), with some variance between models, experimental conditions and island structures themselves. At the model level, JRNN matches expectations in the most cases (its 95% CIs are within zero for 12/16 island tests, including both +gap and −gap conditions), followed by GRNN (11/16 island conditions), GPT-2 (10/16), and finally GPT-3 (9/16). Interestingly, by this metric, both of the recurrent models outperform the transformers, even though the latter were trained on orders of magnitude more data. Thus, these results suggest that, within limits, model architecture, more than training data size, may determine the extent to which the models' generalization behavior is consistent with humans (for similar conclusions in a more controlled setting see Hu et al. (2020)). Turning to experimental conditions, we find that models show approximate equal sensitivity in the +gap condition (where 22/32 of islands have 95% CIs within zero) and −gap condition (20/32). However, we do find that models show some reduction for argument-structure controls in −gap conditions, which is not evident in +gap conditions. As far as island-by-island variation, GRNN, GPT-2 and JRNN show human-like behavior for Adjunct Islands, CNP Islands, Coordination Islands and Wh Islands. However, there is more variation between models for Left-Branch islands, where models fail to show the proper licensing behavior in control conditions, as well as Subject Islands and Sentential Subject Islands.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

Turning to Figure 15, we find statistical evidence for island effects in 28/32 tests across all models, and in 0/8 argument structure controls. In this case, we find that JRNN and GPT-2 demonstrate the expected sensitivity to islands for all structures, whereas GRNN and GPT-3 demonstrated the expected sensitivity for 6/8 structures. Despite cross-model architectural variation, again we find strongest effects for Coordination, Adjunct and Complex NP islands, with weaker effects for Left-Branch Islands, Subject, and Sentential Subject islands.[21]

In summary, we take the results from this section to indicate that general-purpose pattern recognizers can acquire many, although not all, of the island effects from string data alone, with models showing the expected behavior in 65% of cases for our absolute metric and 88% for our relative metric. 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 likelihood. Next, we address four possible concerns about the applicability of our

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

modeling results for addressing questions of learnability.[22] Finally, we 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 tested here has an obvious strongly natural language syntax oriented inductive bias,[23] and, despite some differences, we find many 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 Possible Concerns

The first possible concern has to do with data size: If we assume that a typical child in a native English environment is exposed to about 30,000 words per day (≈ 11-million words per year) (Hart and Risley, 1995), then GRNN has a quantity of linguistic experience comparable to an 8-year-old. If this rate of exposure persists throughout a typical adult lifetime, then JRNN roughly has a quantity of linguistic experience comparable to that of an 80-year-old, GPT-2 that of 10 human lifetimes, and GPT-3 that of 100 human lifetimes. Although some of these models are clearly exposed to vastly more linguistic experience than human children, we believe that their learning outcomes are still relevant to debates about learnability. First, their behavior 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, our largest data model.

The second possible concern is about data genre: All of our models were trained on newswire or similar adult-directed text, which may differ from child-directed speech in critical ways, such as average sentence length and size of the lexicon. Commenting on genre, we would like to point out that 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).

The third possible concern has to do with the effect of parasitic gaps on the logic we use to test for island effects. Parasitic gaps (Engdahl, 1983) are instances where gaps in adjunct islands can exist parasitically on the presence of an upstream host-gap, as in (21), below.

(21) I know what the spy burned ___ after reading ___ last night.

The concern is that, even if a model has acquired adult-like distribution of gaps, it should, nonetheless, posit some gaps in adjunct clauses, thus muddying our prediction for the absence of wh-effects inside islands. While the example of parasitic gaps is extremely interesting, we do not believe that they invalidate our logic. Generally, parasitic gaps need to

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

be preceded by a host gap, which do not exist in any of our test items. Thus, in the absence of upstream host gaps, gaps inside adjuncts are still ungrammatical.[24]

The last possible concern we would like to address is one of coverage. We have chosen to study models' generalization of the filler–gap dependency by looking at embedded wh-questions. We have done so to rule out possible confounds associated with other forms of syntactic movement such as do-support for matrix level questions ("The lion ate the gazelle yesterday. → What *did* the lion eat __ yesterday?"). Just because the models tested have learned many aspects of the filler–gap dependency in the embedded question context does not mean they have learned island constraints for other dependencies, such as relative clauses, root-level questions and topicalization. However, it is possible that the models might learn a non-unitary representation of the wh-dependency, but learn the correct constraints on one of its parts (i.e. embedded questions). In this case, model learning would still provide good empirical evidence that these cosntraints are learnable *in principle*. Additionally, this possible limitation of the current work is alleviated by a recent study (Ozaki et al., 2022) finding that our models *do* learn island constraints for other wh-dependencies, and that these constraints are sometimes even stronger than for embedded wh-questions, providing additional support for our conclusions.

6.3 Argument from the Poverty of the Stimulus

We now discuss the implications of our modeling results. 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 by humans. 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 introduced in Chomsky (1965a), the

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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.[25]

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 $L_0, L_1, L_2, \ldots$.
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 $L_1, L_2, \ldots$.
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 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 some of the island phenomena (stronger learning outcomes for Coordination, Complex NP, Wh and Adjunct Islands and weaker learning outcomes for Left Branch, Subject and Sentential Subject Islands). Therefore, we conclude that their success provides an empirical refutation of the APS for these structures, and neutralizes it as an argument in favor of linguistic nativism.

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

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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 many of the English islands can be learned by domain-general, weakly-biased algorithms, and given the rapid advance in machine learning in the past decade, we suspect 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.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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.

Notes

All authors would like to thank two anonymous reviewers for their insightful comments and suggestions, which have significantly improved the quality of this work. This work was supported by an NVIDIA GPU Grant to RF. RPL gratefully acknowledges support from National Science Foundation award BCS-2121074, the MIT–IBM Artificial Intelligence Research Laboratory, and a Google Faculty Research Award. EGW would like to thank the Mind Brain Be-

havior Interfaculty Initiative at Harvard for funding, as well as Emmanuel Dupoux and the Laboratoire de Science Cognitive et Psycholinguistique for hosting me during the initial stages of this work. Thanks also to audiences at the MM Lab at Harvard, CPL Lab at MIT, HSP, CogSci as well as Tal Linzen, Yonatan Belinkov and Stuart Shieber for valuable feedback and discussion.

[1] This phenomenon has been referred to by multiple names, including wh-dependencies, wh-movement and the filler—gap dependency. We elect to use the filler—gap terminology because it is relatively theory-neutral, and it highlights the entities that play the biggest role in our behavioral tests, the filler and the gap. ‘Island constraints’ were first articulated from within theories that view the filler—gap dependency as one of syntactic movement, but have since been used to describe restrictions on the dependency even by those who do not subscribe to the ‘movement’ analysis (Pollard and Sag, 1994b)

[2] 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.

[3] There are a group of islands that arise due to *semantic* considerations (e.g. “How many miles didn’t you run ___ yesterday?”) (Dayal, 2016), which we do not address. Furthermore,

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

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.

[4] Under their model the relative probability of the gap in (5) is $P(\text{Start}, \text{IP}, \text{VP}) \times P(\text{IP}, \text{VP}, \text{CP}) \times \cdots \times P(\text{VP}, \text{PP}, \text{end})$.

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

[6] Within their modeling setup, Pearl and Sprouse (2013b) avoid this concern by comparing the log-odds of island violating sentences with minimal-pair counterparts of similar length.

[7] 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).

[8] 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 prob-

abilities 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.

[9] The approximation is exact iff (1) the probability of the gap-requiring word w^+ is zero when no filler–gap construction is present, and (2) the probability of the gap-requiring word is the same across the two contexts when assuming that a filler–gap construction is present. For the full technical argument, see Appendix A.

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

[11] In theory, even SRNs can recognize hierarchical languages with reasonable efficiency (Hewitt et al., 2020).

[12] 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)

[13] 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).

[14] 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 sub-

ject 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.

[15] 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))))`

[16] 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.

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

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

[19] 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. (2019a) and Wilcox et al. (2018), although additional items

have been added.

[20] Unlike the filler–gap dependencies, these do not constitute grammatical/ungrammatical contrasts, as *her* in (11-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 (11); all of our models demonstrate strong gendered expectations effects.

[21] 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 2×2 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.

[22] We thank one of our anonymous reviewers for raising the last two.

[23] There are model architectures that *do* encode syntax-oriented inductive biases, both explicitly (Dyer et al., ????) and implicitly (Shen et al., 2018). For performance of these models on the tests discussed here see Wilcox et al. (2019b) and Hu et al. (2020).

[24] One possible exception are sentences that involve both topicalization and parasitic gaps

such as “I know what, without reading __, the spy burned __.”, however, in written text, the topic must be introduced with a comma or an m-dash, which are absent in our sentences, ruling them out as a possible confound. One other way to rule out this potential confound would be to look at the whole post-gap region.

[25] 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.

References

Abeillé, Anne, Barbara Hemforth, Elodie Winckel, and Edward Gibson. 2020. Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition* 204:104293.

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

Ambridge, Ben, Julian M Pine, and Elena VM Lieven. 2014. Child language acquisition: Why universal grammar doesn’t help. *Language* 90:e53–e90.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

Aoshima, Sachiko, Colin Phillips, and Amy Weinberg. 2004. Processing filler-gap dependencies in a head-final language. *Journal of memory and language* 51:23–54.

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

Barr, Dale J, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68:255–278.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 861–872. Vancouver, Canada: Association for Computational Linguistics.

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

Blevins, Terra, Omer Levy, and Luke Zettlemoyer. 2018. Deep rnns encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 14–19. Melbourne, Australia: Association for Computational Linguistics.

Bošković, Željko. 2005. On the locality of left branch extraction and the structure of NP. *Studia Linguistica* 59:1–45.

Bošković, Željko. 2015. From the complex NP constraint to everything: On deep extractions across categories. *The Linguistic Review* 32:603–669.

Bošković, Željko. 2020. On the coordinate structure constraint and the adjunct condition. In *Syntactic architecture and its consequences ii: Between syntax and morphology*, ed. Jamie Douglas Sten Vikner András Bárány, Theresa Biberauer, 227–258. Berlin, Germany: Language Science Press.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Charniak, Eugene. 1996. *Statistical language learning*. Boston: MIT press.

Chaves, Rui P. 2020. What don't RNN language models learn about filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics*, ed. Gaja Jarosz Allyson Ettinger and Max Nelson, volume 3, 20–30. Society for Computation in Linguistics.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005* .

Chen, Stanley F, and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13:359–394.

Chomsky, Noam. 1957. *Syntactic structures*. Berlin, Germany: Walter de Gruyter.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

Chomsky, Noam. 1965a. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1965b. Current issues in linguistic theory. In *The structure of language: Readings in the philosophy of language*, ed. Jerrold Katz Jerry Fodor. New York: Prentice Hall.

Chomsky, Noam. 1973. Conditions on transformations. In *A festschrift for morris halle*, ed. Stephen Anderson and Paul Kiparsky, 232–286. New York: Holt, Rinehart Winston.

Chomsky, Noam. 1979. The logical structure of linguistic theory. *Synthese* 40:317–352.

Chomsky, Noam. 1986. *Barriers*. Boston: MIT press.

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

Chowdhury, Shammur Absar, and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, ed. Akbik, Alan and Blythe, Duncan and Vollgraf, Roland, 133–144. Santa Fe, NM: Association for Computational Linguistics.

Chowdhury, Shammur Absar, and Roberto Zamparelli. 2019. An LSTM adaptation study of (un) grammaticality. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, ed. Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, 204–212. Association for Computational Linguistics.

Clark, Alexander, and Shalom Lappin. 2010. *Linguistic nativism and the poverty of the stimulus*. Hoboken, NJ: John Wiley & Sons.

Clifton, Charles, and Lyn Frazier. 1989. Comprehending sentences with long-distance dependencies. In *Linguistic structure in language processing*, ed. Greg Carlson and Michael Tanenhaus, 273–317. Dordrecht: Springer Netherlands.

Corver, Norbert Ferdinand Marie. 1991. The syntax of left branch extractions. Doctoral Dissertation, Tilburg University.

Crain, Stephen, and Janet Dean Fodor. 1985. How can grammars help parsers? In *Natural language parsing: Psychological, computational, and theoretical perspectives*, ed. David Dowty, Lauri Karttunen, and Arnold Zwicky, 94–128. Cambridge, UK: Cambridge University Press.

Cristia, Alejandrina, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: a time allocation study. *Child development* 90:759–773.

Culbertson, Jennifer, Paul Smolensky, and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition* 122:306–329.

Da Costa, Jillian K, and Rui P Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *Proceedings of the society for computation in linguistics*, ed. Allyson Ettinger, Gaja Jarosz, and Joe" Pater, volume 3, 189–198. New York: Association for Computational Linguistics.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*. Advance publication. https://doi.org/10.1162/ling_a_00491

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ed. Marta R Costa-jussà and Enrique Alfonseca, 2978–2988. Florence, Italy: Association for Computational Linguistics.

Dayal, Veneeta. 2016. *Questions*. Oxford, UK: Oxford University Press.

Dyer, Chris, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. ???? Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA: Association for Computational Linguistics.

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

Elman, Jeffrey L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning* 7:195–225.

Engdahl, Elisabet. 1982. Restrictions on unbounded dependencies in swedish. In *Readings on unbounded dependencies in scandinavian languages*, ed. lisabet Engdahl and Eva Ejerhed, 151–174. Stockholm: Almqvist & Wiksell.

Engdahl, Elisabet. 1983. Parasitic gaps. *Linguistics and philosophy* 6:5–34.

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

Frank, Stefan L, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language* 140:1–11.

Franz, Alex, and Thorsten Brants. 2006. All our n-gram are belong to you. *Google AI Blog* .

Frazier, Lyn. 1987. Syntactic processing: evidence from dutch. *Natural Language & Linguistic Theory* 5:519–559.

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

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

Garnsey, Susan M, Michael K Tanenhaus, and Robert M Chapman. 1989. Evoked potentials and the study of sentence comprehension. *Journal of psycholinguistic research* 18:51–60.

Georgopoulos, Carol. 1991. *Syntactic variables: Resumptive pronouns and a binding in palauan*. Dordrecht, NL: Kluwer.

Giulianelli, Mario, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, ed. Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. Brussels, Belgium: Association for Computational Linguistics.

Gold, E Mark. 1967. Language identification in the limit. *Information and control* 10:447–474.

Goodkind, Adam, and Klinton Bicknell. 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)*, ed. Asad Sayeed, Cassandra L Jacobs, Tal Linzen, and Marten van Schijndel, 10–18. Salt Lake City, UT: Association for Computing Machinery.

Goodluck, Helen, Michele Foley, and Julie Sedivy. 1992. Adjunct islands and acquisition. In *Island constraints: Theory, acquisition and processing*, ed. Helen Goodluck and Michael Rochemont, 181–194. Dordrecht: Springer Netherlands.

Graves, Alex, and Jürgen Schmidhuber. 2009. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, volume 21, 545–552.

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

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), ed. Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Meza-Ruiz, 1195–1205. New Orleans, Louisiana: Association for Computational Linguistics.

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

Hale, John. 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*, 1–8. Pittsburgh, PA: Association for Computational Linguistics.

Hart, Betty, and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young american children.*. Baltimore, MD: Paul H Brookes Publishing Co.

Heilbron, Micha, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. 2021. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences* 119:e2201968119.

Hewitt, John, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. 2020. RNNs can generate bounded hierarchical languages with optimal memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ed. Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 1978–2010. Online: Association for Computational Linguistics.

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

Hofmeister, Philip, Laura Staum Casasanto, and Ivan A Sag. 2013. Islands in the grammar? Standards of evidence. In *Experimental syntax and island effects*, ed. Jon Sprouse and Norbert Hornstein, 42–63. Cambridge, UK: Cambridge University Press Cambridge.

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

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

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ed. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 1725–1744. Online: Association for Computational Linguistics.

Huang, C-T James. 1982. Logical relations in chinese and the theory of grammar. Doctoral Dissertation, MIT.

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

Kaan, Edith, Anthony Harris, Edward Gibson, and Phillip Holcomb. 2000. The p600 as an index of syntactic integration difficulty. *Language and cognitive processes* 15:159–201.

Kiss, Katalin É. 2013. *Configurationality in hungarian*. Berlin: Springer.

Kuno, Susumu. 1973. Constraints on internal clauses and sentential subjects. *Linguistic Inquiry* 4:363–385.

Kush, Dave, Akira Omaki, and Norbert Hornstein. 2013. Microvariation in islands? In *Experimental syntax and island effects*, ed. Jon Sprouse and Norbert Hornstein, 239 – 264. Cambridge, UK: Cambridge University Press.

Lakretz, Yair, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in lstm language models. 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)*, ed. Jill Burstein, Christy Doran, and Thamar Solorio, 11–20. Minneapolis, Minnesota: Association for Computational Linguistics.

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

Laurence, Stephen, and Eric Margolis. 2001. The poverty of the stimulus argument. *The British Journal for the Philosophy of Science* 52:217–276.

Legate, Julie Anne, and Charles D Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review* 18:151–162.

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

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

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

Maling, Joan M. 1978. An asymmetry with respect to wh-islands. *Linguistic Inquiry* 9:75–89.

Mameli, Matteo, and Patrick Bateson. 2011. An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366:436–443.

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics* 19:313 – 340.

Marvin, Rebecca, and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, ed. Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 1192–1202. Brussels, Belgium: Association for Computational Linguistics.

McClelland, James L, David E Rumelhart, PDP Research Group, et al. 1987. *Parallel distributed processing, volume 2: Explorations in the microstructure of cognition: Psychological and biological models*. Cambridge, MA: MIT press.

Nunes, Jairo, and Juan Uriagereka. 2000. Cyclicity and extraction domains. *Syntax* 3:20–43.

Oda, Hiromune. 2017. Two types of the coordinate structure constraint and rescue by pf deletion. In *Proceedings of the North East Linguistic Society*, ed. Katerina Tetzloff Andrew Lamont, volume 47, 343–356. Amherst, MA: Graduate Linguistic Student Association, University of Massachusetts.

Omaki, Akira, Shin Fukuda, Chizuru Nakao, and Maria Polinsky. 2020. Subextraction in japanese and subject-object symmetry. *Natural Language & Linguistic Theory* 38:627–669.

Otsu, Yukio. 1981. Universal grammar and syntactic development in children: Toward a theory of syntactic development. Doctoral Dissertation, Massachusetts Institute of Technology.

Ozaki, Satoru, Dan Yurovsky, and Lori Levin. 2022. How well do lstm language models learn filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2022*, ed. Tim Hunter Allyson Ettinger and Brandon Prickett, 76–88. online: The Society for Computation in Linguistics.

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

Pearl, Lisa, and Jon Sprouse. 2013a. Computational models of acquisition for islands. In *Experimental syntax and island effects*, ed. Jon Sprouse and Norbert Hornstein, 109–131. Cambridge, UK: Cambridge University Press Cambridge.

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

Pearl, Lisa S, and Jon Sprouse. 2019. Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. *Language* 95:583–611.

Perfors, Andrew, Terry Regier, and Joshua B Tenenbaum. 2006. Poverty of the stimulus? a rational approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.

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

Phillips, Colin. 2013. On the nature of island constraints ii: Language learning and innateness. In *Experimental syntax and island effects*, ed. Jon Sprouse and Norbert Hornstein, 132–157. Cambridge, UK: Cambridge University Press.

Phillips, Colin, Nina Kazanina, and Shani H Abada. 2005. Erp effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research* 22:407–428.

Phillips, Colin, and Matthew Wagers. 2007. Relating structure and time in linguistics and psycholinguistics. In *Oxford handbook of psycholinguistics*, ed. Pim Levelt and Alfonso Caramazza, 739–756. Oxford, UK: Oxford University Press.

Pickering, Martin, Stephen Barton, and Richard Shillcock. 1994. Unbounded dependencies, island constraints and processing complexity. In *Perspectives on sentence processing*, ed. Lyn Frazier Keith Rayner Charles Clifton Charles Clifton, Jr., 199–224. Hillsdale, NJ: Lawrence Erlbaum.

Pollard, Carl, and Ivan A. Sag. 1994a. *Head-driven phrase structure grammar*. Stanford, CA: Center for the Study of Language and Information.

Pollard, Carl, and Ivan A Sag. 1994b. *Head-driven phrase structure grammar*. University of Chicago Press.

Pullum, Geoffrey K, and Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The linguistic review* 18:9–50.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

Ravfogel, Shauli, Yoav Goldberg, and Tal Linzen. 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)*, ed. Jill Burstein, Christy Doran, and Thamar Solorio, 3532–3542. Minneapolis, Minnesota: Association for Computational Linguistics.

Richards, Norvin. 2001. *Movement in language: Interactions and architectures*. Oxford, UK: Oxford University Press.

Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* 118:e2016239118.

Rizzi, Luigi. 1982. *Issues in italian syntax*. Dordrecht: Floris.

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

Ross, John Robert. 1967. Constraints on variables in syntax. Doctoral Dissertation, MIT.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323:533–536.

Sabel, Joachim. 2002. A minimalist analysis of syntactic islands. *Linguistic Review* 19:271–315.

Saffran, Jenny R, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274:1926–1928.

Schlesewsky, Matthias, Gisbert Fanselow, Reinhold Kliegl, and Josef Krems. 2000. The subject preference in the processing of locally ambiguous wh-questions in german. In *German sentence processing*, ed. Lars Konieczny Barbara Hemforth, 65–93. Dordrecht: Springer.

Schmidhuber, Jürgen, F Gers, and Douglas Eck. 2002. Learning nonregular languages: A comparison of simple recurrent networks and lstm. *Neural Computation* 14:2039–2041.

Sekerina, Irina A. 2003. Scrambling and processing: Dependencies, complexity, and constraints. In *Word order and scrambling*, ed. Simin Karimi, 301–324. Malden, MA: Blackwell.

Shen, Yikang, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2018. Neural language modeling by jointly learning syntax and lexicon. In *6th International Conference on*

Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. Vancouver, Canada: OpenReview.net.

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. In *Philosophy, language, and artificial intelligence*, ed. Terry L. Rankin Jack Kulas, James H. Fetzer, 79–89. Dordrecht: Springer.

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

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

Sprouse, Jon, and Norbert Hornstein. 2013. *Experimental syntax and island effects*. Cambridge, UK: Cambridge University Press.

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

Stolcke, Andreas. 2002. Srilm — an extensible language modeling toolkit. In *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, ed. John H. L. Hansen and Bryan Pellom, 901–904. Denver, CO: International Speech Communication Association.

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

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

van Schijndel, Marten, and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, ed. Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 4704–4710. Brussels, Belgium: Association for Computational Linguistics.

Vig, Jesse, and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the second blackboxnlp workshop on analyzing and interpreting neural networks for nlp*, ed. Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, 63–76. Florence, Italy: Association for Computational Linguistics.

Weber, Ann, Anne Fernald, and Yatma Diop. 2017. When cultural norms discourage talking to babies: Effectiveness of a parenting program in rural senegal. *Child Development* 88:1513–1526.

Weiss, Gail, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision rnns for language recognition. In *Proceedings of the 56th annual meeting of the association for computational linguistics (short papers)*, ed. Iryna Gurevych and Yusuke Miyao, 740–745. Melbourne, Australia: Association for Computational Linguistics.

White, Jennifer C, and Ryan Cotterell. 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 and the 11th International Joint Conference on Natural Language Processing*, ed. Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 454–463. Online: Association for Computational Linguistics.

Wilcox, Ethan, Roger Levy, and Richard Futrell. 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the second blackboxnlp workshop on analyzing and interpreting neural networks for nlp*, ed. Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, 181–190. Florence, Italy: Association for Computational Linguistics.

Wilcox, Ethan, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? In *Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp*, ed. Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, 211–221. Brussels, Belgium: Association for Computational Linguistics.

Wilcox, Ethan, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019b. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of naacl-hlt 2019*, ed. Sudipta Kar, Farah Nadeem, Laura Burdick, Greg Durrett, and Na-Rae Han. Minneapolis, Minnesota: Association for Computational Linguistics.

Wilcox, Ethan, Pranali Vani, and Roger P. Levy. 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 and the 11th International Joint Conference on Natural Language Processing, ed. Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 939–952. Online: Association for Computational Linguistics.

Wilcox, Ethan Gotlieb, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 2020 Meeting of the Cognitive Science Society*, ed. Yang Xu Blair C. Armstrong Stephanie Denison, Michael Mack, 1707–1713. Online: Cognitive Science Society.

Yang, Yuan, and Steven T. Piantadosi. 2022. One model for the learning of language. *Proceedings of the National Academy of Sciences* 119:e2021865119.

Yoshida, Masaya. 2006. Constraints and mechanisms in long-distance dependency formation. Doctoral Dissertation, University of Maryland, College Park.

Ethan Gotlieb Wilcox

Department of Linguistics

Harvard University

wilcoxeg@g.harvard.edu

Richard Futrell

Department of Language Science

University of California, Irvine

rfutrell@uci.edu

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

Roger Levy

Department of Brain and Cognitive Science

Massachusetts Institute of Technology

rplevy@mit.edu

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:

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability.
Linguistic Inquiry. Advance publication. https://doi.org/10.1162/ling_a_00491

$$\frac{P(w^+|C_{\text{what}})}{P(w^+|C_{\text{that}})} = \frac{\sum_X P(w|X, C_{\text{what}})P(w^+|C_{\text{what}})}{\sum_X P(w^+|X, C_{\text{that}})P(X|C_{\text{that}})} \tag{5}$$

$$= \frac{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{what}}) + P(w^+|x^-, C_{\text{what}})P(x^-|C_{\text{what}})}{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{that}}) + P(w^+|x^-, C_{\text{that}})P(x^-|C_{\text{that}})}. \tag{6}$$

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:

$$\frac{P(w^+|C_{\text{what}})}{P(w^+|C_{\text{that}})} \approx \frac{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{what}})}{P(w^+|x^+, C_{\text{what}})P(x^+|C_{\text{that}})}. \tag{7}$$

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:

$$\frac{P(w^+|C_{\text{what}})}{P(w^+|C_{\text{that}})} \approx \frac{P(x^+|C_{\text{what}})}{P(x^+|C_{\text{that}})}. \tag{8}$$

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. As described in Section 3.2, one can indeed argue that for truly human-like linguistic generalization in contexts where omitting the gapped phrase is ungrammatical without a filler, we should have $P(x^+|C_{\text{that}}) \approx 0$ so the *larger* the value of this ratio, the more human-like the model's behavior. If we take the negative log, we get a difference in surprisals ranging between $(-\infty, \infty)$ whose value should be *smaller* (closer to $-\infty$) for more human-like generalization:

$$-\log P(w^{+}|C_{\text{what}}) + \log P(w^{+}|C_{\text{that}}). \tag{9}$$

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

$$-\log P(w^{-}|C_{\text{what}}) + \log P(w^{-}|C_{\text{that}}) \tag{10}$$

would be indicative of human-like sensitivity to filler–gap structural relationships.

These differences of log inverse-probabilities in Equations (9) and (10) 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), Egnlish, 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