Learning Syntactic Structures from String Input

Ethan Gotlieb Wilcox[1], Jon Gauthier[2], Jennifer Hu[2]
Peng Qian[2], Roger Levy[2]
[1] Harvard University [2] Massachusetts Institute of Technology

Abstract

This chapter addresses a series of interrelated questions about the origin of syntactic structures: How do language learners generalize from the linguistic stimulus with which they are presented? To what extent does linguistic cognition recruit domain-general (i.e. not language-specific) processes and representations? And to what extent are rules and generalizations about linguistic structure separate from rules and generalizations about linguistic meaning? We address these questions by asking what syntactic generalizations can be acquired by a domain-general learner from string input alone. The learning algorithm we deploy is a neural-network based Language Model (Radford et al., 2019), which has been trained to provide probability distributions over strings of text. We assess its linguistic capabilities by treating it like a human subject in a psycholinguistics experiment, and inspect behavior in controlled, factorized tests that are designed to reveal the learning outcomes for one particular syntactic generalization. The tests presented in this chapter focus on a variety of syntactic phenomena in two broad categories: rules about the structure of the sentence and rules about the relationships between smaller lexical units, including scope and binding. Results indicate that our target model has learned many subtle syntactic generalizations, yet it still falls short of humanlike grammatical competence in some areas, notably for cases of parasitic gaps (e.g. "I know what you burned ___ after reading ___ yesterday"). We discuss the implications of these results under three interpretive frameworks, which view the model as (a) a counter-argument against claims of linguistic innateness, (b) a positive example of syntactic emergentism, and (c) a fully-articulated model of grammatical competence.

1 Introduction

How do we gain knowledge about the representations that underlie linguistic communication? For the last fifty years, researchers have relied on a combination of introspective judgements and experimental results. This dual approach has been effective, in part because much of the research has focused on adult human subjects,

who are capable of both describing internal judgements and producing experimentally-controlled behavior. However, recent advances in computing have produced a new generation of artificial linguistic agents who, for the first time, learn to produce output that could plausibly be called "humanlike" in its sophistication. Their performance has led researchers to ask whether they exhibit the same type of generalizations that underlie human linguistic communication. This chapter addresses both the epistemic question—how can we gain knowledge about what artificial agents have learned?—as well as several possible inferences that can be drawn from these results for linguistic theory. We focus on the learning outcomes of Artificial Neural Networks (ANNs), algorithms that can learn arbitrary relationships between input and output by representing information at one or more intermediary, or hidden, layers of representation. In particular, we investigate application of ANNs, known as language models (LMs): algorithms trained on large amounts of text that assign joint probability distributions over strings, which can be used to produce a probability distribution over a target word given its preceding context. In our case, all of our language models are trained on English. The ANNs we use are domain-general learning algorithms in the sense that they are general-purpose pattern recognizers, and have been used to successfully model data as disparate as photos of flowers and vehicles (Chen et al., 2020; Dosovitskiy et al., 2021) to the structure of proteins (Rives et al., 2021).

We will begin our discussion by presenting one answer to the empirical question—how can we gain knowledge about what artificial agents have learned? We introduce a paradigm for probing these models, dubbed the "psycholinguistics paradigm" because it treats ANNs like human subjects in a psycholinguistics experiment. In Section 3 and Section 4, we present case studies on specific syntactic phenomena in two different areas. While the ANNs discussed produce surprisingly human-like linguistic behavior, we argue that they do not in themselves constitute a theory of linguistic cognition — just as a human brain in itself does not constitute a theory of linguistic cognition. In order to make inferences about the knowledge underlying linguistic communication from ANN behavior, in Section 5, we discuss three interpretive frameworks, and discuss their implications for theories and future research.

2 Methods: Psycholinguistic Assessment of ANNs

Here we analyze Autoregressive Language Models (henceforth simply LMs) (Elman, 1990, 1991) that are trained to predict the next token x_i given a context of preceding tokens x_{-i}—$P(x_i|x_{-i})$. The theoretical and practical reasons to use LMs to address learnability questions in natural language syntax are compelling. Incremental word prediction plays a major role in human language processing (Hale, 2001; Levy, 2008a; Kuperberg and Jaeger, 2016), demonstrating that the autoregressive LM objective function corresponds to a highly optimized function for language in the human mind. If grammatical abstractions emerge from the application of general-purpose learning models trained on this objective function using human-scale learning data, it

places plausible lower bounds on the learnability of those abstractions. From an engineering standpoint, contemporary LMs have achieved impressive apparent fluency in the language they generate (Jozefowicz et al., 2016; Radford et al., 2019; Brown et al., 2020), raising for for the first time the prospect that sophisticated human-like grammatical abstractions could potentially be emergent in models trained without explicit grammatical supervision. However, these outputs—probability distributions of multi-thousand word vocabularies–are difficult to interpret. How, then, can we evaluate whether a model has "learned" a certain syntactic generalization? Two classes of approaches to this problem have emerged in recent years. The first is to focus on the internal representations learned by the model, attempting to find homomorphisms between these representations and the representations postulated by linguists (Giulianelli et al., 2018, and Hewitt and Manning, 2019; Hewitt et al., 2020 for LMs trained on the closely related Cloze objective function). Here, we take the alternate approach of behavioral testing, elucidating the implications of the models' word-by-word probability distributions for whether they are making human-like grammatical generalizations (building on previous work such as Linzen et al., 2016; Gulordava et al., 2018).

We can reveal what syntactic generalizations have been implicitly learned by an ANN by inspecting their conditional word probabilities values in controlled, factorized tests that are designed to reveal the learning outcomes for one particular syntactic rule, or generalization. These tests are inspired by the types deployed in psycholinguistic studies (Linzen et al., 2016; Futrell et al., 2018). Because of this, we refer to this methodology as the "Psycholinguistic Assessment Paradigm" for analysis of ANNs. To get an intuition about how these tests can reveal learned generalizations about the structure of a context C and what structures it can co-occur with, consider a comparison between two different contexts, C_1 and C_2, within which a critical word or phrase, w, can only occur with non-negligible probability if a latent structural property X holds of the context. For example, (1-a) is a famous garden-path sentence, where the context C_1 is structurally ambiguous between *man* serving as the head noun of the subject (premodified by the adjective *old*) versus as the main verb of the clause (in which case *The old* is a nounless subject); (1-b) is its unambiguous counterpart.

(1) a. $\overbrace{\text{The old man}}^{C_1}\ldots$ the walls. [AMBIGUOUS]

b. $\overbrace{\text{The old manned}}^{C_2}\ldots$ the walls. [UNAMBIGUOUS]

In this case, the garden-path disambiguating word *the* requires that the preceding word was the main verb of the clause; this is the latent property X that must hold of the context. We can rewrite $P(w|C)$ making explicit the marginalization over

whether X holds:

$$P(w|C) = P(w|X,C)P(X|C) + P(w|\neg X,C)P(\neg X|C) \tag{1}$$

and, since $P(w|\neg X,C) \approx 0$,

$$P(w|C) \approx P(w|X,C)P(X|C). \tag{2}$$

Furthermore, we specifically focus on cases where $P(w|X,C_1) \approx P(w|X,C_2)$—that is, if the latent property X holds, there is little difference betweeen the contexts of how likely w is. In Example (1), for example, if *man/manned* is a verb, then the only difference between the variants is whether the verb is in present or past tense; this difference is unlikely to affect whether a direct object comes next or what direct object is likely. In such cases, what pushes around the relative probability of w is $P(X|C_1)$ versus $P(X|C_2)$, namely how strongly the context predicts the latent property X. This means that the $P(w|C_1)$ versus $P(w|C_2)$ comparison, which we can easily compute for any trained LM, can be used to assess whether and how strongly the model's word predictions reflect the latent structural generalizations of interest.

The decomposition of Equation 2 offers us something further, too. In practice, we will be testing models using collections of sentence sets, or "items". Across items, the lexical particulars vary but the structural properties of the context are consistent; within an item, the lexical particulars are very tightly controlled. By taking the log of the inverse of the conditional word probability we get its *surprisal* (Hale, 2001; Levy, 2008a), converting a product into sums:

$$S(w|C) \equiv \log \frac{1}{P(w|C)} \approx \log \frac{1}{P(w|X,C)} + \log \frac{1}{P(X|C)}$$

Due to lexical differences, the first term $\log \frac{1}{P(w|X,C)}$ is likely to vary substantially across items, but not within an item. This means that we can use within-items repeated-measures analysis across multiple items to zero in on the presence and robustness of differences across contexts in the second term, $\log \frac{1}{P(X|C)}$, which is what we are really interested in: has the model acquired a generalization about what latent structures fit in what contexts? We use base-2 logs so that surprisals are measured in *bits*: one bit is associated with observing an event that occurs with probability 1/2, an event with probability 1 has a surprisal of 0 bits, and an event that is impossible has an infinite number of bits of surprisal. Surprisal also plays a substantial theoretical role in psycholinguistics: a word's surprisal contributes linearly to the amount of time it takes a native speaker to read it in naturalistic text (Hale, 2001; Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020, though see Brothers and Kuperberg, 2021). Using surprisal as a linking function allows us to make concrete predictions about model outputs based on human behavior, both in absolute terms (word x_i should be assigned approximate probability p based on human processing

times) and in *relative* terms (word x_i should be assigned a lower probability than word y_i).

Connecting with experimental psycholinguistics terminology, we can think of Example (1) as one "item" in two "conditions": UNAMBIGUOUS, in which the sentence up to "the walls", is unambiguously an NP+VP, and AMBIGUOUS, in which the first part is ambiguous between NP and NP+VP interpretations. Each item in the test contains a critical region, which is the underlined NP "the walls." If the model has learned that (1-a) is ambiguous between two interpretations, then it should assign the critical region a lower probability in this condition. Or, in terms of surprisal:

$$S(\text{the walls}|\text{The old manned}) < S(\text{the walls}|\text{The old man}). \tag{3}$$

However, the converse is not necessarily true: as this prediction only considers a single sentence pair, a given ANN model could produce output satisfying this criterion without learning the proper generalization. Thus, in order to abstract away from particular lexical items and sentential contexts, we assess models on *suites* of test items all of which exemplify the syntactic context and continuation of interest, and formulate our predictions over conditions instead of individual items. For example, we would generalize the single-item prediction in Equation (3) to

$$S(\text{NP}|\textsc{Unambiguous prefix}) < S(\text{NP}|\textsc{Ambiguous prefix}), \tag{4}$$

and when the word and prefix are sufficiently clear, we will simply write the condition name. Typically, test suites consist of 20-40 items in four conditions, with two predictions that specify an inequality between two conditions. If the model were to assign random probabilities to each word, it should make the correct guess $\sim 50\%$ of the time, for each prediction. All of our test suites come from `www.syntaxgym.org`, a website for hosting and sharing syntactic test suites (Gauthier et al., 2020). (We encourage readers of this article to explore SyntaxGym and create their own test suites to assess whatever aspect of syntax they wish.)

So far, we have been discussing ANN models only in the abstract. For the rest of this chapter we will look at the learning outcomes of GPT-2 (Radford et al., 2019), a large Transformer-based language model. Transformers (Vaswani et al., 2017) are a type of ANN architecture that has recently become dominant in machine learning research. While previous architectures encoded linear order through a mechanism called recurrence, in which the network sequentially consumed pieces of input (Elman, 1990; Hochreiter and Schmidhuber, 1997), Transformers encode linear order into their input directly and keep track of the relative importance of the relationship between each input piece at each layer of representation. The version of GPT-2 reported here was trained on a corpus of web-text that contained $\sim$8 billion tokens. Under the assumption that children are exposed to about 30,000 words per day ($\approx$ 11-million words per year) (Hart and Risley, 1995), this model has the linguistic experience comparable to about 8 human lifetimes. Results for other models, including models that implement recurrence, are hosted online at `syntaxgym.org`.

3 Gross Syntactic State: Complementizer Phrases

One of the key features of human language is that groups of adjacent lexical items form larger phrasal units, which are recursively nested within each other. The tests in this section are designed to assess whether the models have learned (1) whether groups of words form larger, cohesive phrases; and (2) whether they have learned to make generalizations about the phrase as a whole. The type of phrase we focus on here are Complementizer Phrases (CPs), which are the theoretical equivalent of sentences. In the base case, chunks of words that form CPs must constitute distinct utterances, which are orthographically distinguished in English with end-of-sentence punctuation. However, there are a variety of strategies that can be used to combine multiple CPs together in a single sentence, including conjunction, subordination and relativization. The tests deployed here leverage the latter two strategies to assess what models have learned about CPs and how they can be combined.

3.1 Subordination

Subordinating conjunctions, such as *when*, *as*, or *while*, set up temporal relationships between a *matrix* and *subordinate* CP. In order to test whether our model of interest has learned the proper generalizations about these words and the type of phrasal units they conjoin, test suites were created along the lines of (2). (Critical regions are underlined in this and future examples.) Two predictions should hold: First, if models have learned that presence of a subordinator necessitates the presence of two CPs, then end-of-sentence punctuation after the first phrase should be more surprising when a subordinator is present than when it is absent. That is, the period should have higher surprisal in the [+SUBORDINATOR, -MATRIX] condition than the [-SUBORDINATOR, -MATRIX] condition (Prediction 1). Second, if models have learned that two CPs cannot exist in the same sentence without conjunction then the second CP should have higher surprisal when no subordinator is present; i.e., in the [-SUBORDINATOR, +MATRIX] condition compared to the [+SUBORDINATOR, +MATRIX] condition (Prediction 2).

(2) Subordination

a. As the doctor studied the book, the nurse walked into the office. [+SUBORDINATOR, +MATRIX]
b. *As the doctor studied the book. [+SUBORDINATOR, -MATRIX]
c. *The doctor studied the book, the nurse walked into the office. [-SUBORDINATOR, +MATRIX]
d. The doctor studied the book. [-SUBORDINATOR, -MATRIX]

The results for this experiment are shown in Figure 1. We first focus on the left panel. Each sentence fed to the model is broken into non-overlapping regions, shown on the x-axis. The region of interest for our two predictions is the main clause,

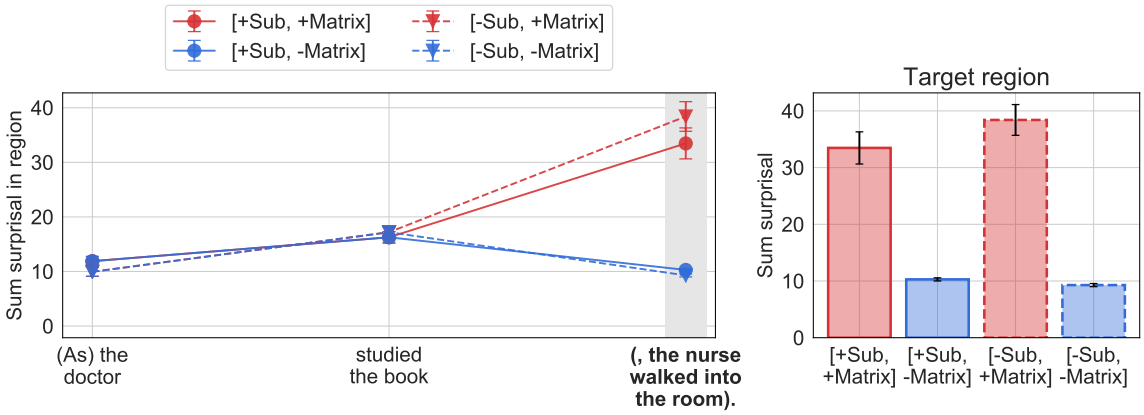

Figure 1: **Subordination.** Left: sum GPT-2-derived surprisal in each sentence region, averaged across items in test suite. Target region is highlighted with gray box and shown in boldface. Right: average sum surprisal in target region for each condition. GPT-2 output is consistent with the expectation of a matrix clause when a subordinator is present, and end-of-sentence punctuation when a subordinator is absent.

which is highlighted in bold. As a region may contain multiple tokens, the y-axis shows the summed surprisal across tokens for each region, averaged across all the items in a test suite. We now turn to the performance of GPT-2 with respect to the two predictions described above. For Prediction 1, GPT-2 achieves 95.7% accuracy, and for Prediction 2, it achieves 100% accuracy. This is reflected in the relative height of the bars in the right panel (corresponding to surprisal at the target region): Although it is not visually obvious from the scale, the [+SUBORDINATOR, -MATRIX] bar is higher than the [-SUBORDINATOR, -MATRIX] bar, as expected by Prediction 1; more visually evident is the fact that the [-SUBORDINATOR, +MATRIX] bar is higher than the [+SUBORDINATOR, +MATRIX] bar, as expected by Prediction 2. Overall, these results are consistent with the model having learned the relationship between subordiators and CPs.

3.2 Relative Clauses

One other way to conjoin CPs in a single sentence is via relativization, where a CP has been inserted medially into a host clause along the lines of (3), below. In this case, there is a dependency relationship between the top-level subject Noun Phrase and the verb of the embedded CP. Crucially, there are also dependencies that exist between the subject and verb at each CP layer, and because of the rules on relativization these follow a first-in-first out linear order—the first verb must have as its subject the last noun *(artist/painted)*, and the second verb must have the first noun as its subject *(painting/deteriorated)*.

(3) [$_{CP1}$ The painting$_{N1}$ [$_{CP2}$ that the artist$_{N2}$ with the long dark hair painted$_{V2}$] deteriorated$_{V1}$].

In order to test whether GPT-2 has learned the proper syntactic generalization about relativization, we exploit this last-in-first-out rule for subject/verb relationships in test suites following (4), below. Sentences are created with two NPs and two verbs such that one pairing of the two is semantically plausible and the other pairing is implausible. In addition, we use verb transitivity to add an additional element of plausibility; one reason why "deteriorated" is implausible as the first verb in the example above is because it is used intransitively in the vast majority of cases.[1] If the model has learned the last-in-first-out rule for relativized CPs, then it should be less surprised by the condition that forces the plausible pairing than the implausible one. Each sentence also contains a relative clause modifying the second NP, such that a model cannot succeed simply by assigning higher probability to the bigram "artist painted" than "artist deteriorated" without recourse to the rest of the sentence (cf. (3)). In (4), for example, it is implausible that a painting could paint and an artist could deteriorate, at least compared to the more plausible condition where the artist paints and the painting deteriorates. Therefore, we expect a model that has learned the correct relativization pattern to assign higher surprisal to "deteriorated painted" than "painted deteriorated", given the same prefix "The painting that the artist who lived long ago" (i.e., $S_{\text{IMPLAUSIBLE}} > S_{\text{PLAUSIBLE}}$).

(4) Relativized Complements (Center Embedding)

a. The painting that the artist who lived long ago painted deteriorated. [PLAUSIBLE]

b. The painting that the artist who lived long ago deteriorated painted. [IMPLAUSIBLE]

The results for this experiment are shown in Figure 2, again with regions on the x-axis and summed surprisal for each sentence region on the y-axis. At the critical regions, the blue line, which indicates surprisal in the IMPLAUSIBLE condition is higher than the red line, which indicates surprisal in the PLAUSIBLE condition. Looking at accuracy scores, GPT-2 achieves 85.7% accuracy on this test suite, indicating that it has learned the proper last-in-first-out rule for embedded clause completion.

3.3 Main Verb / Reduced Relative Clause Gardenpath Effects

Gardenpath sentences are sentences in which an initial, locally plausible, interpretation becomes globally implausible at a certain point, resulting in a variety of

[1] According to the data from Goldberg and Orwant (2013), it is used intransitively in over 98% of occurances (see `https://github.com/wilcoxeg/verb\_transitivity`)

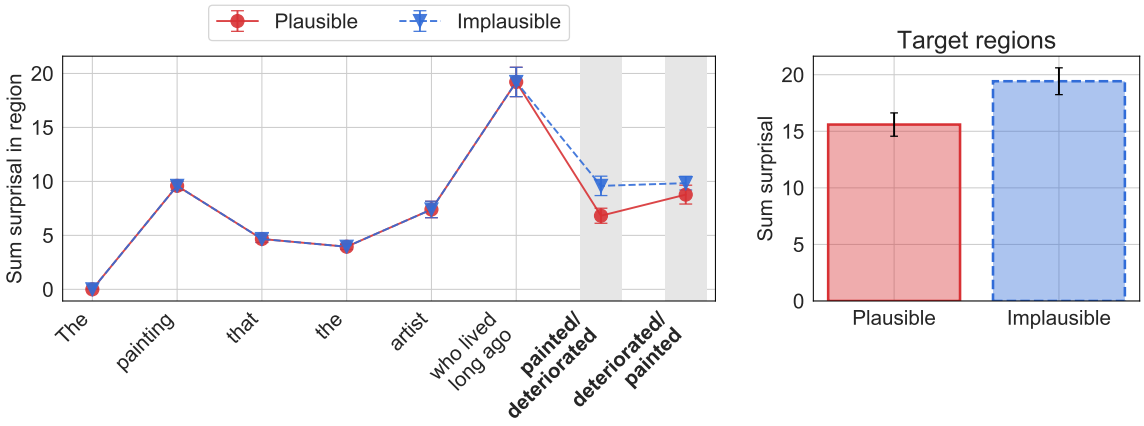

Figure 2: **Center embedding.** Left: sum GPT-2-derived surprisal in each sentence region, averaged across items in test suite. Right: average surprisal summed over target regions for each condition. GPT-2 output is consistent with the last-in-first-out rule for pairing nouns and verbs in relativized CPs.

well-studied processing effects, including regressive checking and slowdown in reading times (Pritchett, 1988; Ferreira and Henderson, 1991; Ferreira et al., 2001). At the outset, we introduced our methods using the famous gardenpath sentence "The old man the walls", however here we focus on a different type of gardenpath, called the Main-Verb/Reduced-Relative (MV/RR) gardenpath. For example, the string "The woman brought a sandwich from the kitchen..." is ambiguous between two interpretations: a Main-Verb Interpretation, in which the verb *brought* is a past-tensed verb associated with the matrix clause of the sentence, and the Reduced Relative Interpretation, in which *brought* is a passive participle, associated with a reduced relative clause. The two interpretations are given in (5), below.

(5) "The woman brought a sandwich..."

a. $[_{CP}\ [_{NP}$The woman$]\ [_{VP}$brought a sandwich$]\ ]$ (Main-Verb Interpretation)

b. $[_{NP}$ The woman $[_{CP}$ (who was) brought a sandwich$]\ ]$ (Reduced Relative Interpretation)

Crucially, the two different interpretations should lead to differing expectations for upcoming material. The main-verb interpretation should lead to expectations for end-of-sentence punctuation, adjuncts and other CP-modifiers, or coordinating conjunctions following *sandwich*. The reduced relative interpretation should lead to expectations for a verb, which is still required to complete the sentence. Because the statistics of English heavily favor the main verb interpretation (*brought* is rarely used as a passive particle), most people reading (5) parse the sentence along (5-a)—they are "led down the garden path"—and are surprised if they encounter a subsequent

verb. Note that, while the other tests in this chapter discuss model behavior between sets of items that differ in terms of their grammaticality, gardenpath sentences are not technically ungrammatical. What this test suite does have in common with the others, is that critical regions create a violation of structural expectations. These underlying structural expectations are the object of our interest—grammatical violations and gardenpath sentences are both useful for revealing them.

While garden-path effects have traditionally been used to gain insight into how the human processor manages uncertainty during online interpretation, they rely on assumptions that the processor is forming specific types of representations over which to have uncertainty in the first place. While previous investigation of gardenpath effects in ANNs has demonstrated how they can be used as models of psycholinguistic processing (Futrell et al., 2018; Van Schijndel and Linzen, 2018, 2021), we use them here to build up empirical support for the hypothesis that ANNs have learned abstract phrasal units including complex NPs and CPs. We do so by creating test suites following (6), where items can come in one of four conditions: The main verb can either be AMBIGUOUS or UNAMBIGUOUS between being a main verb and a reduced relative; for example, *brought* is ambiguous but *given* is not. Additionally, sentences can be UNREDUCED in which case they explicitly introduce relative clauses with the relativizer and passivizer *who was*, or else they can be REDUCED, in which case the relative clause is not signaled overtly. Note that the only condition which is a true 'gardenpath' is the [REDUCED, AMBIGUOUS] condition; even though *brought* is ambiguous by itself, when it is preceded by a passive verb (as in *was brought*) it can only be analyzed as introducing a relative clause. Therefore, if models had correctly learned that both passive participles and overt relativization and passivization introduce relative clauses that modify the subject NP, they should find the main verb *tripped* more surprising in this condition than in any of the others. In other terms, we expect an interaction between REDUCTION and AMBIGUITY that results in a superadditive amount of surprisal.

(6) MV/RR Gardenpath Effects:

a. The woman brought a sandwich from the kitchen tripped on the carpet. [REDUCED, AMBIGUOUS]
b. The woman who was brought a sandwich from the kitchen tripped on the carpet. [UNREDUCED, AMBIGUOUS]
c. The woman given a sandwich from the kitchen tripped on the carpet. [REDUCED, UNAMBIGUOUS]
d. The woman who was given a sandwich from the kitchen tripped on the carpet. [UNREDUCED, UNAMBIGUOUS]

The results for this experiment are shown in Figure 3. The top plot shows the region-by-region surprisal for GPT-2, with regions on the x-axis and summed surprisal on the y-axis. We can see that, as per our prediction, the [REDUCED, AMBIGUOUS] condition induces higher surprisal than any of the other conditions in the critical

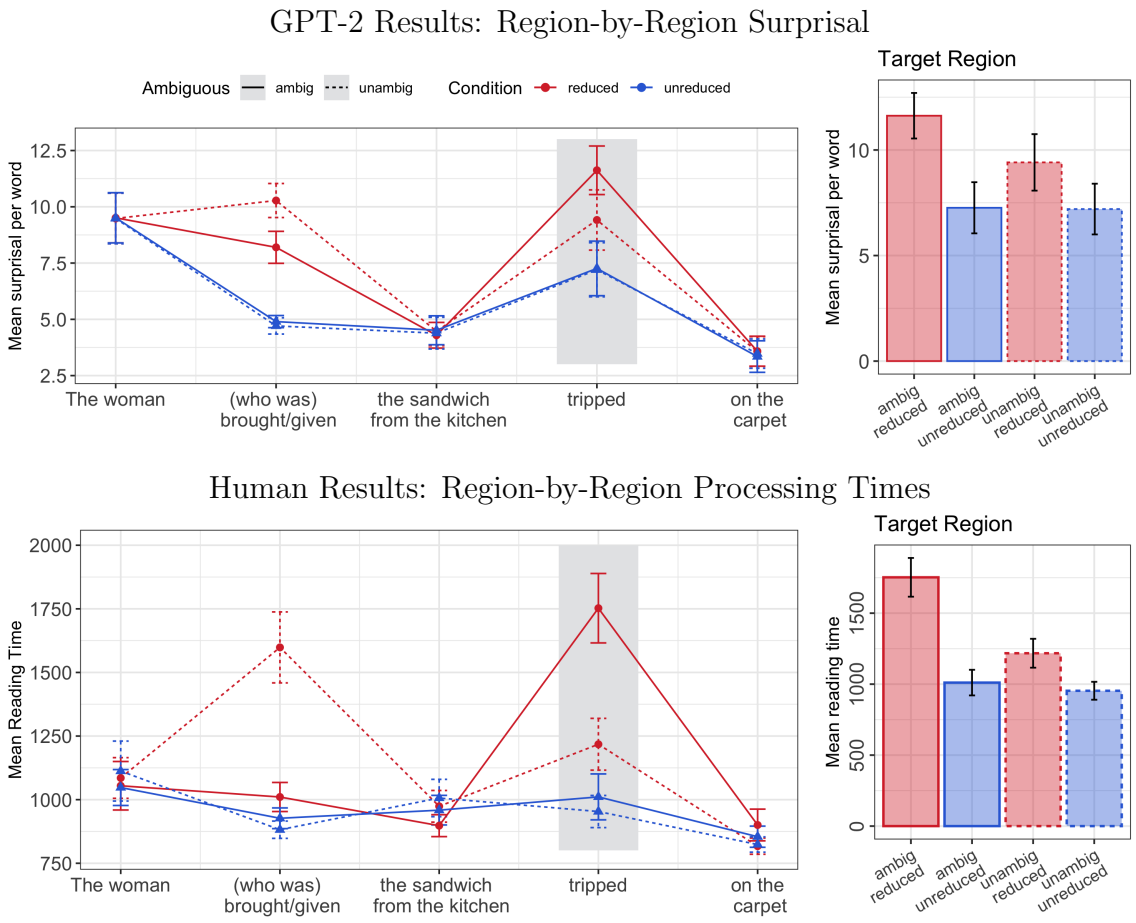

Figure 3: **Main Verb / Reduced-Relative Clause Gardenpath Effects.** Top: mean sum surprisal in each sentence region for GPT-2. Bottom: mean reading time (ms) in each sentence region for human subjects. Both GPT-2 and humans show increased processing difficulty when the sentence is ambiguous between two structural interpretations.

region (the main verb). To compare model-derived surprisal to word-by-word processing times in human subjects, we show reading time data from Vani et al. (2021) on the y-axis of the bottom plot. Processing times are collected using the Maze task (Forster et al., 2009; Boyce et al., 2020), an experimental paradigm in which participants read through a sentence one word at a time. For each word, they must select between a correct and plausible continuation and an implausible distractor word. The time that it takes to make this selection has been shown to be a good measure of incremental processing difficulty, with reduced spillover effects, which are a common feature of other incremental processing measures, such as self-paced-reading (Boyce et al., 2020). By framing the model's incremental predictions as surprisal values, we can see how its behavior is strikingly similar to that of human subjects in two key ways. First, we find increased RTs and surprisal values for passive participles (i.e., for *given*) in the [REDUCED, UNAMBIGUOUS] condition. And more importantly, we see hugely elevated RTs and surprisal values in the critical region for the [REDUCED, AMBIGUOUS] condition. Overall, these results compatible with the hypothesis that models, like humans, are deriving different structural interpretations of the different conditions, which drive expectation for upcoming material.

One extremely revealing and previously unreported pattern shows up in these data. Even though the verb *given* unambiguously introduces a relative clause, we find that GPT-2 is still slightly 'gardenpathed' by it. That is, we find elevated surprisal at *tripped* in the critical region for the [REDUCED, UNAMBIGUOUS] condition (dotted red line). This same 'gardenpath ghost' effect shows up in human reading times, too, where we find elevated RTs for 'tripped' in the same condition. Traces of this effect have been observed in previous studies (F. Ferreira, P.C.), however the increased sensitivity of the Maze task means Vani et al. (2021) is the first to report the effect as significant. For humans, this effect could be explained by a noisy channel process (Shannon, 1956), which has been hypothesized to be at-play in a number of language processing phenomena (Levy, 2008b). The string "The woman given..." may be so unlikely that participants re-interpret it as the active past-tense "The woman gave...", which leads to a garden-path effect, with similar re-interpretations for other items in the test suite. However, it is unclear whether and how models could implement such reinterpretation. Thus, these data present an interesting and open question that demands further investigation.

4 Dependencies, Scope and Binding

In this section, we take a look at the syntactic generalizations learned by GPT-2 that have to do with the co-variation between individual lexical items and phrasal units that are "smaller" than CPs, particularly Noun Phrases.

4.1 Filler–Gap Dependencies

Filler–gap dependencies are dependencies between a *filler* (a wh-word such as *who* or *what*) and a gap, which is an empty syntactic position (denoted with underscores in examples below). The relationship is a true bidirectional dependency, in that the presence of a filler necessitates the presence a downstream gap, and gaps are not licensed in the absence of an upstream filler. We assess whether GPT-2 has learned this dependency by creating a test suite with four conditions, outlined in (7). If the model has learned that gaps must be licensed by a filler, then after encountering a transitive verb the model should expect an NP. Thus, skipping over the NP gap and proceeding directly to an indirect object (*to the patient*) should be less surprising when a filler is present compared to when it is absent. That is, at the indirect object, $S_{[+\text{FILLER, }+\text{GAP}]} < S_{[\text{-FILLER, }+\text{GAP}]}$ (Prediction 1). Conversely, if the model has learned that fillers set up expectations for gaps that must be discharged at likely gap sites, then the direct object of a verb *the new medicine* should be more likely in the absence of a filler than in its presence. That is, at the direct object, $S_{[\text{-FILLER, -GAP}]} < S_{[+\text{FILLER, -GAP}]}$ (Prediction 2). For both of these predictions, in order to appropriately set up expectations for either objects or gaps in object position, we used obligatorily transitive verbs, such as *administer* in our example below.[2]

(7) Filler–Gap Dependencies

a. I remember what the nurse should administer ___ to the patient without further delay. [+FILLER, +GAP]

b. *I remember what the nurse should administer the new medicine to the patient without further delay. [+FILLER, -GAP]

c. *I remember that the nurse should administer ___ to the patient without further delay. [-FILLER, +GAP]

d. I remember that the nurse should administer the new medicine to the patient without further delay. [-FILLER, -GAP]

The results for this experiment are shown in Figure 4, with each sentence region on the x-axis and summed surprisal for all the words in that region on the y-axis. As expected from Prediction 1, we find that the post-gap preposition is higher surprisal in the [-Filler, +Gap] condition than in the [+Filler, +Gap] condition. The model has an accuracy score of 95.8% for this prediction. Turning towards Prediction 2, GPT-2 exhibits higher surprisal in the filled argument position when a filler is present (i.e, in [+Filler, -Gap]) compared to when it is absent ([-Filler, -Gap]), and the model is 100% accurate for this prediction. These high scores suggest that the model has learned the correct bidirectional dependency between fillers and gaps. See `www.syntaxgym.org` for suites that test gaps in subject and indirect object position, for which we find similar, high accuracy performance.

[2]These tests are adapted from Wilcox et al. (2018); for a fuller discussion of the paradigm see Wilcox et al. (2021).

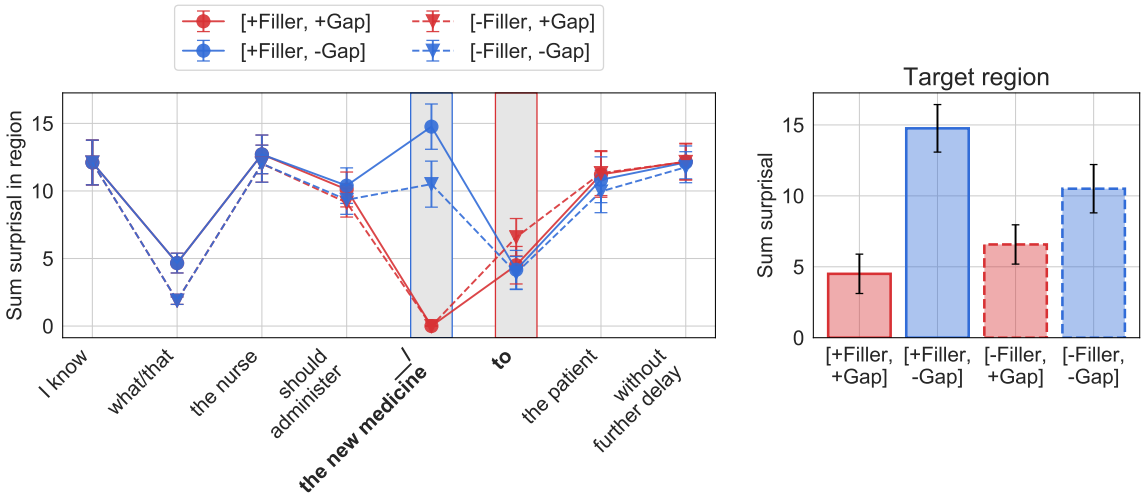

Figure 4: **Filler–Gap Dependencies (Object Gap).** Left: sum GPT-2-derived surprisal in each sentence region, averaged across items in test suite. Surprisal values at the object position in the +Gap conditions are shown as 0 for visualization purposes. Note that the target region is different for +/-Gap conditions (as indicated by red and blue outlines, respectively). Right: average surprisal summed over target region for each condition. GPT-2 output is consistent with the bidirectional dependency between fillers and gaps in object position.

4.1.1 Island constraints

Filler–gap dependencies have played an outsized role in generative linguistics, in part because they are one of the few dependencies that is claimed to be truly unbounded. Gaps can be licensed by fillers that are an arbitrary (and possibly infinite!) number of nodes away in a phrase structure. But while the dependency is potentially unbounded in length, it is not entirely unconstrained. For example, gaps are licensed in subject, object, or indirect object position in embedded clauses, but gaps are not licensed when they appear in an adjunct phrase attached to the clause, as demonstrated in (8).

(8) * I know who the nurse admitted the patient after inspecting ___ yesterday night.

The structural constraints on filler–gap dependency are known as island constraints (Ross, 1967). Island constraints have been one of the core empirical patterns cited as evidence for Nativist approaches to language (Phillips, 2013)—children within the same language community learn island constraints without any direct supervision, and the same types of structural configurations give rise to island constraints in unrelated languages. A handful of recent studies (Wilcox et al., 2018, 2019, 2021) have assessed whether or not ANN models can learn island constraints using the

following methodology: If the model has learned that fillers do not license gaps in islands, then the presence of an upstream filler should not impact the surprisal of a gap inside island structures. That is, in (8), *yesterday* should not be any more or less likely in a minimal-pair counterpart without a filler. Wilcox et al. (2021) find that GPT-2 successfully attenuates its expectations for gaps inside island constructions for eight of the most-studied islands, including adjuncts, suggesting that it has learned the proper grammatical generalizations for the distribution of the filler–gap dependency in English.

Below, we build upon this work and provide one example of the learning limits for GPT-2. While it is true that gaps are not licensed in adjunct clauses in English, there is one exception to this rule: They may be if the adjunct directly follows a clause which hosts an gap, as in (9-d). In this case, the gap inside the adjunct clause is said to be 'parasitic' on the embedded clause gap, and this construction is referred to as the *parasitic gap* construction (Ross, 1967; Engdahl, 1983). To test whether models have learned such adverb phrases as the legitimate loci of gaps, we created sentences following (9), where the HostGap factor indicates variation in the host gap, not the 'parasitic' gap. Two types of sentences were created, one like (9) below where the adverb phrase included a gerund verb, and one in which the verb was tensed. If models have learned the parasitic gap construction, we would expect no difference in surprisal based on the presence of a filler in the -HostGap condition (which is effectively an island). However, we would expect surprisal to be affected by the presence of absence of a filler in the +HostGap condition, where the presence of a wh-word renders the whole sentence grammatical.[3]

(9) Parasitic Gaps

a. *I know that the nurse admitted the patient after she inspected __ yesterday. [-Filler, -HostGap]

b. *I know who the nurse admitted the patient after she inspected __ yesterday. [+Filler, -HostGap]

c. *I know that the nurse admitted __ after she inspected __ yesterday. [-Filler, +HostGap]

d. I know who the nurse admitted __ after she inspected __ yesterday. [+Filler, +HostGap]

The results of this experiment are shown in Figure 5, with mean by-region surprisal on the y-axis. While there is no difference in surprisal based on the presence of an upstream filler in the -HostGap condition, as expected, we also find no effect of *filler* in the +HostGap, which should be the case if the model has learned the

[3]We do not include the -Gap conditions, as we predict the expectations for gaps set up by wh-complementizers to be fully discharged at the first gap site. These expectations correspond to the contrast in grammaticality between the filled-gap version of our basic tests (**I know what you read the paper.*) compared to the filled-gap version of a parasitic gap sentence (*I know what you read ___ before shredding the paper.*)

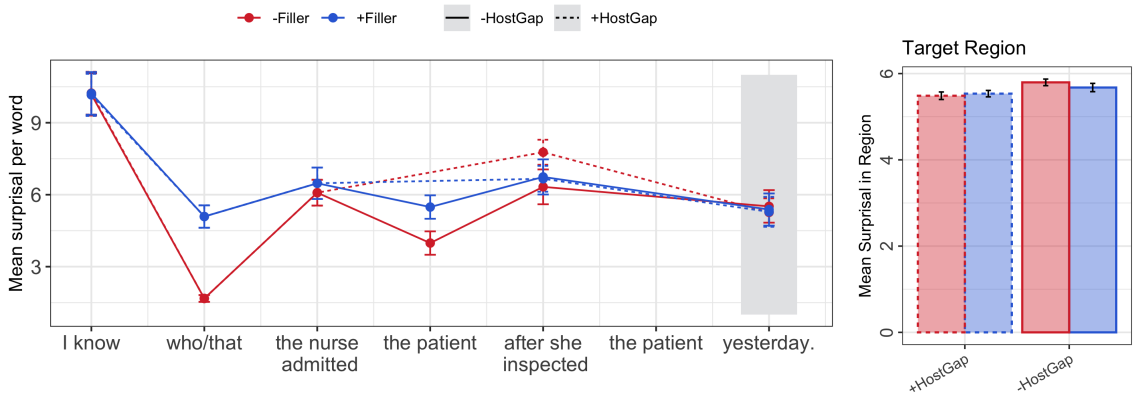

Figure 5: **Parasitic Gaps.** If GPT-2 had learned parasitic gaps, we would expect a difference in surprisal between +/-FILLER in the +HOSTGAP condition, but not in the -HOSTGAP condition. We see no difference between conditions, indicating that GPT-2 has not learned the proper generalization.

proper generalization.

4.2 Reflexive Anaphora

Anaphoric pronouns form dependencies with referring Noun Phrases (or r-expressions), with which they co-refer. For example, in (10), the anaphoric pronoun *herself* refers to the antecedent *The author with the senators*.

(10) The author with the senators hurt herself.

Unlike fillers and gaps, anaphoric dependencies are not bidirectional: although an anaphoric pronoun requires an r-expression, r-expressions may be used freely without subsequent pronouns. However, like fillers and gaps, the relationship is subject to a number of restrictions, including that the r-expression and pronoun must match in gender and number features and, structurally, that the r-expression must $c-$command the anaphor (Carnie, 2021). To test whether the models have learned the dependency, as well as its structural restrictions, we create test suites following (11). Each sentence contains an anaphoric pronoun, a head-noun that $c-$commands the pronoun, and a modifying prepositional phrase with a secondary noun that does not $c-$command the pronoun. There are four conditions: The head-noun can either be SINGULAR or PLURAL, and it can either MATCH or MISMATCH with the anaphoric pronoun in terms of number. We run two suites of tests, one with feminine anaphoric pronouns like (11), and one with masculine anaphoric pronouns.[4] For these sentences, the non-$c-$commanding NPs always mismatch with the head noun, meaning that they are all

[4]Note that while anaphors must agree in gender with the r-expression, we do not test this generalization.

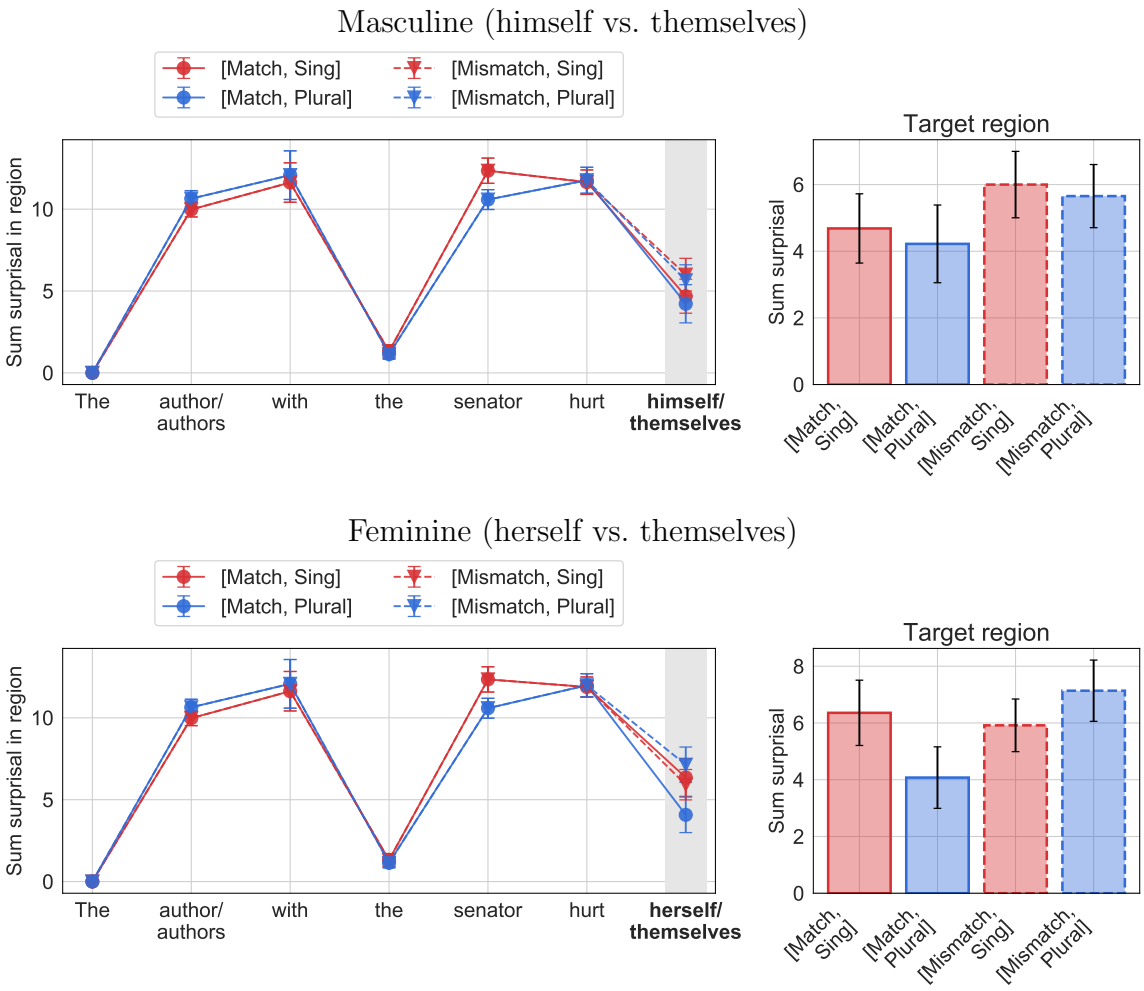

Figure 6: **Reflexive Anaphora (Number Agreement).** Top: masculine reflexive pronoun. Bottom: feminine reflexive pronoun. GPT-2 output is broadly consistent with the expected agreement pattern, except in the case where it must correctly predict a singular feminine pronoun ("herself").

"distractors" that could possibly trick the model into expecting the wrong number on the pronoun. If the model has learned the correct generalizations, then it should exhibit lower surprisal in the MATCH condition, both when the head NP is plural (Prediction 1) and singular (Prediction 2).

(11) Reflexive Anaphora: Number Agreement

a. The author with the senators hurt herself. [MATCH, SING]
b. The author with the senators hurt themselves. [MISMATCH, SING]
c. The authors with the senator hurt themselves. [MATCH, PLURAL]
d. The authors with the senator hurt herself. [MISMATCH, PLURAL]

The results for this experiment are shown in Figure 6. Beginning with masculine anaphoric pronouns (top), we find that GPT-2 is 89% accurate for both Prediction 1 (plural agreement) and Prediction 2 (singular agreement). Turning to the feminine anaphoric pronoun (bottom), for Prediction 1 GPT-2 is 100% accurate; however for Prediction 2, where the model must correctly predict a singular pronoun, the model is only 21% accurate. This indicates that the model is distracted by the number mismatching NP in the prepositional clause, but only in cases where the pronoun is feminine.

4.3 Negative Polarity Items

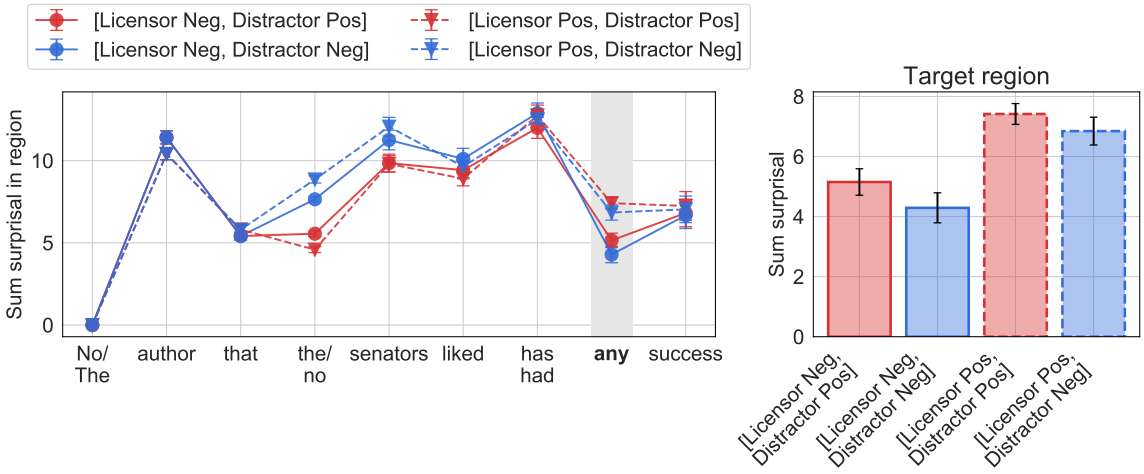

Figure 7: **Negative Polarity Item Licensing.** Left: sum GPT-2-derived surprisal in each sentence region, averaged across items in test suite. Right: average surprisal summed over target region for each condition. GPT-2 output is consistent with the structural relationships licensing negative polarity items.

Negative Polarity Items (NPIs) are a class of polarity-sensitive items, which must appear in downward-entailing environments, or contexts that license inferences from

supersets to subsets (Chierchia, 2013). One way to construct these environments is with negative quantifiers, such as the English determiner *no*. In order for the NPI to be in the scope of the negative quantifier, it must be $c-$commanded by it, thus we can set up suites to determine whether or not GPT-2 has learned the relevant structural relationship that are similar to the tests used for anaphoric pronouns, schematized in (12). Sentences have an NPI, *any*, and a subject NP that contains either a negative determiner or a positive determiner (*the*). Each subject NP is modified with a relative clause whose subject likewise contains either a negative or positive determiner. If models have learned the correct rules for NPI licensing, then surprisal at the NPI should be lower for conditions where the subject NP is headed with *no* than for conditions where it is headed with *the*. We formulate this in terms of three predictions: at the NPI, $S_{[\text{NEG, NEG}]} < S_{[\text{POS, NEG}]}$ (Prediction 1), $S_{[\text{NEG, POS}]} < S_{[\text{POS, NEG}]}$ (Prediction 2), and $S_{[\text{NEG, POS}]} < S_{[\text{POS, POS}]}$ (Prediction 3).

(12) Negative Polarity Item Licensing

a. No author that the senators liked has had any success. [NEG, POS]
b. No author that no senators liked has had any success. [NEG, NEG]
c. *The author that the senators liked has had any success. [POS, POS]
d. *The author that no senators liked has had any success. [POS, NEG]

The results for this experiment are shown in Figure 7, with sentence regions on the x-axis and surprisal on the y-axis. We find that the model is 100% accurate for Prediction 1, and 97% accurate for Predictions 2 and Prediction 3, suggesting that it has learned the proper generalizations for NPI licensing in English.

5 Discussion

The previous sections have introduced several assessments of GPT-2's syntactic capacity. First, we tested the model's ability to represent abstract categories like CPs and NPs, as well as the relationships between them. We next evaluated its ability to learn rules about lexical dependencies, including negative polarity items, anaphoric pronouns and even empty syntactic positions (gaps). We found that, in the vast majority of cases, GPT-2 successfully learned the relevant grammatical generalizations: the model achieves above 90% accuracy for predictions on suites testing for Subordination, Relative Clause completion, Filler–Gap dependencies, and NPI Licensing. There were two notable cases, however, where the model did not make human-like predictions. These were for predictions about the feminine anaphoric pronoun, *herself*, as well as for cases of *parasitic gaps*, when gaps that appear in typically-ungrammatical positions can exist 'parasitically' on the presence of a previous host gap.

What do these results—many positive, some negative—mean for our theoretical understanding of human linguistic cognition? The type of answer that we will explore here is that they can bear on a series of questions about the "origin" of linguistic

structures, associated with the generative tradition in linguistics. These questions include: What learning biases to children bring to the task of language acquisition? How do language learners generalize from the linguistic stimulus with which they are presented? To what extent does linguistic cognition recruit domain general (i.e. not language-specific) processes and representations? And to what extent are rules and generalizations about linguistic structure separate from rules and generalizations about linguistic meaning? Below, we introduce three frameworks for answering these interrelated questions and present the role that the learning outcomes of ANNs can play for each. Finally, we discuss how our results might motivate future development for each approach.

Before moving on to the three interpretive frameworks, we want to return to an issue mentioned during the introduction: the issue of data scale and how it interacts with syntactic generalization. Because we do not know, precisely, what syntactic structures are present in GPT-2's ∼8 billion training tokens, it may be premature to say that our tests capture genuine syntactic generalization as opposed to, say, mere recapitulation of its training data. In order to interrogate this concern, we want to distinguish between two forms of generalization—compositional generalization and what we will call lexico-categorical generalization. Let's consider the case of Negative Polarity Items, discussed in Section 4.3. We argued that what the experiment in this section showed was that GPT-2 had learned that NPIs require negative licensors and also that NPIs must exist in a certain structural relationship *vis a vis* their licensor, using target sentences of the form "*No* + NP + Relative Clause (RC) + *any...*" Now, we can imagine the model learning the structural relationship in two possible ways. First, it could be the case that the model was never exposed to sentences that were structurally similar to our target sentence during training. Rather, the model saw training sentences like (a) "No + NP + *any...*" and (b) "No + NP + RC...". If this were the case, in order to produce the observed behavior, the model would have to make generalizations about a novel frame by composing (a) and (b) together. This would be an example of compositional generalization. On the other hand, it may be the case that that the model did have experience with sentences of the form "No + NP + RC + *any...*". In this case, in order to produce the correct behavior, the model had to recognize our target items as similar, and flexibly extend the generalizations it had made during training to sentences that are realized with different items, but share the same structural properties. This would be an example of lexico-categorical generalization.

Because the experiments we have presented do not control the types of syntactic structures available to the model during training, many of the conclusions we can draw are about lexico-categorical generalization, rather than compositional generalization (although our tests certainly don't rule out the latter). But we don't want to downplay the importance of lexico-categorical generalization. In order to flexibly extend grammatical knowledge, models must learn that words fall into abstract categories and must learn to formulate distributional rules at the level of the category

rather than at the level of the word. Both of these behaviors are key features of human linguistic cognition, and they are being learned, here, by a domain-general artificial learning algorithm. Below, we turn back to our main discussion by introducing three interpretive frameworks for understanding the role of GPT-2's learning outcomes.

5.1 Syntactic Nativism

This framework posits that language acquisition is powered by innately given and language-specific (rather than domain-general) processes and representations. A primary task of language acquisition is to derive the syntactic generalizations underlying grammatical adult speech using these inbuilt processes and representations. Traditionally, syntactic nativism is motivated by a type of logical argument called The Argument from the Poverty of the Stimulus (APS; Chomsky, 1975; Pullum and Scholz, 2002; Legate and Yang, 2002; Clark and Lappin, 2010). We can present the APS with the following four premises: (i) For a given linguistic generalization (or rule), it is either learned via a language-specific learning process or via a domain-general learning process. (ii) The linguistic data a child is exposed to are compatible with a large number of possible generalizations $L_0 \dots L_n$ and yet (iii) the child ends up learning the generalizations associated with their actual language, L_0. (iv) There are no domain-general learning processes that favor L_0 over $L_1 \dots L_n$. Therefore, the generalization could not have been learned via a domain-general learning process.

Positive results demonstrating the human-like syntactic behavior of ANNs complicate premise (iv). If an algorithm that is sufficiently domain general can be found that *does* learn the correct generalizations, given an appropriate approximation of a child's linguistic input, then premise (iv) must be restricted to a smaller set of phenomena. Note, however, that the previous reasoning hinges on three premises: (a) The learning model must be sufficiently domain general; (b) it must have learned the correct generalization; and (c) its training data must be a "reasonable enough" approximation of the linguistic stimulus of a human child. As for (a), although ANNs — like any learning algorithm — do possess bias, their architecture is designed to handle arbitrary relationships between arbitrary pieces of input, and they have been successfully deployed in a number of domains. Much of the introduction presented an argument that the psycholinguistic assessment paradigm successfully solves (b). This leaves us with (c): Is GPT-2's training data (or the training data of any ANN) a close-enough approximation of the human linguistic stimulus for these models to play a valid role in winnowing down the scope of the APS? There are a number of dimensions along which this question could be asked, including genre, size, and modality, all of which reveal large differences between a human child and GPT-2. In terms of genre, children learn with the assistance of child-directed speech (Rowe, 2008), while GPT-2 was trained on an adult-directed corpus of web text; in terms of size, GPT-2 was trained on about eight lifetimes of human linguistic experience; and in terms of modality, GPT-2 was trained on pre-segmented text, whereas children must induce

the relevant representations from the speech stream.

Although these differences are a serious consideration, we believe that they do not rule out the learning outcomes of ANNs as a novel piece of empirical evidence that bears on the APS. As far as genre, if the purpose of child-directed speech is (in part) to assist the child in making the relevant syntactic generalizations, then adult-directed language may pose a more difficult dataset from which to make the necessary syntactic generalizations. Additionally, there is evidence that speech genre does not play an important role in language learning, including a number of cultures that do not practice child-directed speech (Weber et al., 2017; Cristia et al., 2019). Turning towards dataset size, we have focused on GPT-2 because it is one well-studied model that represents the state of the art in machine learning. However, many of the tests presented in this chapter have been conducted on models trained with much smaller datasets (Wilcox et al., 2021), even including a version of GPT-2 trained on the equivalent of the linguistic experience of a four-year old (Hu et al., 2020). In addition, some recurrence-based models, such as the one presented in Gulordava et al. (2018) which was trained on the linguistic experience of an 8-year-old, have shown remarkable success at learning hierarchical-sensitive dependencies. Finally, given the rapid advances in machine learning and recent focus on low-resource and computationally efficient natural language processing, we believe that the amount of data required to learn crucial syntactic generalizations from string input alone will decrease in the future.

5.2 Syntactic Emergentism (Non-Syntactic Nativism)

This framework posits that human language learners are endowed with certain (potentially domain-specific) structural biases, but these biases are semantic or logical in nature. Syntactic generalizations such as part of speech, hierarchical relationships and locality constraints are an emergent property, either supervening on patterns in the semantic structures that underlie utterances, or else resulting from fixed constraints in the mapping between these structures and linear-order forms.

Under the syntactic emergentist framework, ANNs provide a positive example of the emergence of syntactic behavior from a suitably domain-general model, paired with a proxy for the task of human communication. Compare this to the role for ANNs under more nativist approaches, where they are useful only so far as they can provide counter-argument against the APS but not as constructive hypotheses in their own right about how language is learned and implemented. Under this framework, what should be learned from cases where models are not able to learn the relevant facts about syntax? One constructive approach, we suggest, is that these phenomena may point to cases where the relevant structures are acquired either through non-syntactic learning pressures or processing constraints.

5.3 Full Emergentism

This framework treats human linguistic knowledge as an emergent phenomenon which arises from the interaction of multiple simple units of an interconnected network (McClelland et al., 1986; Elman, 1990, 1991). The task of language acquisition is to derive syntactic generalizations which accurately replicate the observed distribution of human language. Under this view, ANNs are concrete algorithmic hypotheses — a candidate solution for representing and processing syntactic structures. This position has been summarized recently by Baroni (2021), who suggests that it would be fruitful to take ANNs seriously as "algorithmic linguistic theories making predictions about utterance acceptability." One attractive feature of this approach is that, as far as they are instantiations of a particular theory, ANNs make clear and explicit predictions about unseen data. Thus, even for researchers who may be averse to emergent approaches to language, ANNs can provide useful baselines against which to clarify and compare algebraic models of linguistic phenomena. Here, however, instead of focusing on unseen data and untested structural phenomena, we have instead focused on providing a sanity check. Does GPT-2 produce reasonable predictions for well-studied syntactic phenomena?

Our results demonstrate that a wide range of syntactic phenomena can emerge from the simple task of predicting the next word, providing strong empirical support for the emergentist perspective. However, model failure to capture some aspects of syntax suggest that in addition to generating novel predictions about untested syntactic structures, researchers may want to continue developing and testing new models on the well-studied grammatical phenomena discussed in this chapter.

References

Baroni, M. (2021). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv preprint arXiv:2106.08694*.

Boyce, V., Futrell, R., and Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Brothers, T. and Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

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

Carnie, A. (2021). *Syntax: A generative introduction*. John Wiley & Sons.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. (2020). Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. OUP Oxford.

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

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

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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

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

Engdahl, E. (1983). Parasitic gaps. *Linguistics and philosophy*, pages 5–34.

Ferreira, F., Christianson, K., and Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of psycholinguistic research*, 30(1):3–20.

Ferreira, F. and Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6):725–745.

Forster, K. I., Guerrera, C., and Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, 41(1):163–171.

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

Gauthier, J., Hu, J., Wilcox, E., Qian, P., and Levy, R. (2020). Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

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

Goldberg, Y. and Orwant, J. (2013). A dataset of syntactic-ngrams over time from a very large corpus of english books.

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

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

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

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

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

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

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

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

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

Kuperberg, G. R. and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.

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

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

Levy, R. (2008b). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 234–243.

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

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1986). *Parallel distributed processing*, volume 2. MIT press Cambridge, MA.

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

Pritchett, B. L. (1988). Garden path phenomena and the grammatical basis of language processing. *Language*, pages 539–576.

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

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

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., 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(15).

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

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of child language*, 35(1):185–205.

Shannon, C. (1956). The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2(3):8–19.

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

Van Schijndel, M. and Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In *CogSci*.

Van Schijndel, M. and Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

Vani, P., Wilcox, E., and Levy, R. (2021). I(nterpolated) maze: High-sensitivity measurement of ungrammatical input processing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Wilcox, E., Futrell, R., and Levy, R. (2021). Using computational models to test syntactic learnability. *Lingbuzz Preprint: lingbuzz/006327*.

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

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

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