Large Language Models and the Argument From the Poverty of the Stimulus

Nur Lan, Emmanuel Chemla, and Roni Katzir

January 2024

Abstract

How much of our linguistic knowledge is innate? According to much of theoretical linguistics, a fair amount. One of the best-known (and most contested) kinds of evidence for a large innate endowment is the so-called *argument from the poverty of the stimulus* (APS). In a nutshell, an APS obtains when human learners systematically make inductive leaps that are not warranted by the linguistic evidence. A weakness of the APS has been that it is very hard to assess what is warranted by the linguistic evidence. Current Artificial Neural Networks appear to offer a handle on this challenge, and a growing literature over the past few years has started to explore the potential implications of such models to questions of innateness. We focus here on Wilcox et al. (2023), who use several different networks to examine the available evidence as it pertains to wh-movement, including island constraints. They conclude that the (presumably linguistically-neutral) networks acquire an adequate knowledge of wh-movement, thus undermining an APS in this domain. We examine the evidence further, looking in particular at parasitic gaps and across-the-board movement, and argue that current networks do not, in fact, succeed in acquiring wh-movement and do not even provide a passable approximation of wh-movement. We also show that the performance of one of the models improves considerably when the training data are artificially enriched with instances of parasitic gaps and across-the-board movement. This finding suggests, albeit tentatively, that the failure of the networks when trained on natural, unenriched corpora is due to the insufficient richness of the linguistic input, thus supporting the APS.

1 Background: innateness and the argument from the poverty of the stimulus

One way in which linguists have argued that humans are born with nontrivial biases is through cases where speakers' linguistic knowledge goes beyond what seems warranted by the data they were exposed to. If humans systematically arrive at this knowledge given the data while linguistically-neutral learners do not, then humans are not linguistically neutral: they come to the task of language acquisition prepared. Reasoning of this kind is known as an *argument from the poverty of the stimulus* (APS), and since its introduction by Noam Chomsky over 50 years ago it has been central to the study of the human linguistic capacity.[1,2] Here we will focus on one APS, concerning wh-movement, but various other APSs have been discussed in the literature based on a range of empirical phenomena such as *one*-substitution (introduced in Baker 1978), subject-auxiliary inversion (introduced in Chomsky 1971), and plurals within compounds (introduced in Gordon 1985).

The APSs just mentioned (and others like them) have been taken to argue for nontrivial innate biases in humans. For example, the APS from subject-auxiliary inversion has been taken to support an innate bias for hierarchical transformations over linear ones. The APS from wh-movement that we discuss below will similarly support an intricate bias that a linguistically-neutral learner is not expected to have. The same holds for other APSs in the literature. In this, these APSs go beyond the early observation that children can produce and understand unboundedly many sentences after encountering only a finite number of sentences (Chomsky 1957, p. 15). While generalizing from a finite input to an infinite lan-

[1]The general considerations behind the APS are discussed already in chapter 1 of Chomsky 1965. Further considerations are discussed in Chomsky 1971, pp. 26–8, Chomsky 1975, pp. 30ff., and Chomsky, 1980, pp. 42ff., as well as in much subsequent work.

In addition to the APS, linguists have also identified other sources of evidence supporting the innateness of nontrivial linguistic knowledge. For example, there are arguments from the *richness* of the stimulus, where a pattern that is clearly represented in the input data and would be easily picked up by a linguistically-neutral learner is simply ignored by human learners. Evidence from typological asymmetries has also played a very important role in linguistic reasoning. A proper discussion of such sources of evidence falls outside the scope of the present paper, and in what follows we focus exclusively on the APS.

[2]Throughout the discussion we set aside the question of whether the knowledge under consideration is specific to linguistics (and, if so, how much of it is purely syntactic) or whether it is shared with other cognitive domains. Our sole focus is on whether a neutral learner would be justified in acquiring the relevant knowledge based on a given linguistic input.

guage is perhaps not entirely trivial, it is something that most learning algorithms do. And importantly, this ability does not imply any biases that a linguistically-neutral learner will not have.

While the APS has been central to linguistic reasoning, it has also generated much controversy. Contesting a given APS requires challenging either the knowledge attained by humans or the information available to the child learner. It is the latter that often comes under attack. The reason for this vulnerability is that it is extremely difficult to assess what information exactly is available to the child over the relevant time period (often years of exposure) and hard to tell what a general-purpose, linguistically-neutral learner would do with this kind of information. One can try to look for pieces of evidence that seem relevant for the knowledge at stake — e.g., as done for the case of subject-auxiliary inversion in English by Legate and Yang (2002) — but as noted by Lewis and Elman (2001), Perfors et al. (2011), and others, this methodology runs the risk of underestimating the available information: even if we fail to find the evidence we are looking for, a general-purpose learner might be able to take advantage of other sources of information. This methodology also risks *over*estimating the available information: even if we find several instances of the evidence we are after, a general-purpose learner might treat those instances as noise and fail to draw the inference that we intuitively expect it to. In the absence of an actual learner that can use the information that is available in an entire corpus it is just very hard to estimate whether the data support the knowledge under consideration.[3]

How then can we reason about the information available to the child and ask whether it suffices to support the acquisition of a given piece of knowledge by a linguistically-neutral learner? In an ideal world, one would (a) take a sufficiently powerful learner that can be seen to not be biased in favor of the relevant knowledge; (b) train this learner on a corpus that corresponds to the linguistic input that children receive; and (c) check whether the learner has indeed acquired the knowledge under consideration. In such an ideal world, one might perhaps be able

[3]See Pullum and Scholz (2002), Lidz et al. (2003), Foraker et al. (2009), Hsu and Chater (2010), Berwick et al. (2011), Perfors et al. (2011), and Pearl and Sprouse (2013), among others, for much relevant discussion.

In studies of analogous inductive leaps in other species, this worry regarding the input has been addressed by controlling the information available to the learners (see, e.g., Dyer and Dickinson 1994). To a certain extent this can be done with humans in experiments of artificial-grammar learning (see, e.g., Wilson 2006). But for the main APSs in the literature, which concern the normal course of child language acquisition, controlling the information available to the learner is not an option.

to work with a program induction algorithm for a general-purpose programming language such as Python, which is sufficiently powerful to represent the kinds of knowledge that linguists consider but can be seen as linguistically neutral. For example, nothing about a language like Python seems to favor programs that incorporate wh-movement of the kind that English has over programs that do not. One would still need to ensure that the learning algorithm itself does not bias the learner for or against linguistic patterns, but this can be done in various ways, such as by using a linguistically-neutral prior within a Bayesian learner. After training on a developmentally-realistic corpus, corresponding to a few years of human linguistic experience, the knowledge acquired by the algorithm can then be directly inspected at stage (c).

In the actual world, combining (a) through (c) is currently impossible. For many years, the combination of (a) and (b) was already a major barrier. General program induction algorithms of the kind just mentioned, for example, address (a) but fail on (b), since they are limited to very small training corpora. On the other end of the scale, n-gram models can easily be trained on very large corpora, thus addressing (b), but their representational capacity is much too limited to capture or even to adequately approximate linguistic knowledge such as wh-movement. Other models, such as probabilistic context-free grammars, fall between these two extremes but still typically struggle with the combination of (a) and (b) when it comes to patterns such as wh-movement.

The challenge of assessing the information available to the child has become less of an obstacle lately, with the advent of Large Language Models (LLMs). These models, which rely on modern architectures of artificial neural networks (ANNs), do not yet fully address any of (a) through (c) — a matter that has been discussed in recent literature and that we return to below — but they can be trained on very large corpora and are generally quite successful in acquiring sequential dependencies.[4] This has allowed a large and growing literature to use these models to ask questions relating to the learning of linguistic knowledge by LLMs,

[4]Long before the current models, earlier ANN architectures were used in debates of the APS, and in particular in attempts to argue against various versions of it (see Elman et al. 1996, Lewis and Elman 2001, and Reali and Christiansen 2005, among others, and see Berwick et al. 2011 for a critical analysis of some earlier attempts). Early ANNs, however, were limited in their capacities and generally trained on small corpora, and it is unclear whether they could be used to reason about whether a corpus that roughly corresponds to children's linguistic exposure supports the acquisition of complex grammatical knowledge. In this sense, these earlier models were not yet capable of addressing the combination of (a) and (b). The ability of current models to train on realistically large corpora is a helpful step towards using them constructively in debates about the APS.

often with specific reference to the APS. Of particular relevance to our purposes here is work starting with Linzen et al. (2016) and including Bernardy and Lappin (2017), Chowdhury and Zamparelli (2018), Gulordava et al. (2018), Kuncoro et al. (2018), Marvin and Linzen (2018), Wilcox et al. (2018, 2019, 2023), Bhattacharya and van Schijndel (2020), Chaves (2020), Warstadt et al. (2020), Huebner et al. (2021), Ozaki et al. (2022), and Yedetore et al. (2023), among others, that examines the preference of LLMs within minimal pairs. Here we focus on the application of LLMs to the domain of wh-movement, following Wilcox et al. (2018, 2019, 2023), Chowdhury and Zamparelli (2018), Bhattacharya and van Schijndel (2020), Chaves (2020), Warstadt et al. (2020), and Ozaki et al. (2022). In particular, we examine the claim by Wilcox et al. (2023; WFL) that current models debunk an APS in this domain: one that says that the input is insufficiently rich to allow a general-purpose learner to acquire wh-movement.[5]

The present paper extends WFL's probing of LLMs' knowledge of wh-movement, arriving at conclusions that are at odds with those of WFL. We start, in Section 2, with a brief overview of the general setup for the rest of the paper. Among other things we discuss how LLMs can be used as tools for assessing the information in a given corpus without assuming that these models are cognitively plausible in any way and without even asking whether these models have achieved an adequate knowledge of the pattern under consideration.[6] Rather, we treat these models as proxies for future learners and ask only whether these proxies provide a reasonable approximation of the target pattern. In Section 3 we discuss the success of LLMs in simple cases of wh-dependencies, as noted by WFL. In Section 4 we show that the scope of the LLMs' success is rather limited. In particular, LLMs fail to adequately approximate human knowledge of a much-studied family of cases, falling under the labels of parasitic gaps and across-the-board movement, in which certain additional gaps make an otherwise problematic gap inside an island acceptable. It is cases such as these that are typically taken by linguists to suggest an APS, and our findings show that the performance of current LLMs does not, in fact, debunk this APS. In Section 5, we ask whether the LLMs fail only due to their own limitations or whether their failure reflects also the insufficient

[5]See Pearl and Sprouse (2013) and Phillips (2013) for earlier discussion of APS in the context of acquiring islands.

[6]Our results do bear on the question of the cognitive plausibility of LLMs, however. In particular, since our results are negative they provide further evidence, if such was needed, that current LLMs are not cognitively plausible models of human linguistic cognition, *contra* Piantadosi (2023). See Katzir (2023), Kodner et al. (2023), Moro et al. (2023), and Rawski and Baumont (2023), among others, for additional discussion.

richness of their training data. We address this question by retraining one of the models on corpora that are clearly *not* impoverished with respect to the relevant patterns and showing that the performance of the model improves significantly on the enriched corpus. This, in turn, strengthens the APS, if also tentatively. Section 6 concludes.

2 The general setup

Simplifying considerably, a *gap*, such as the missing complement of 'with' in (1a) and (1c), appears if and only if it is preceded by an appropriate *filler*, such as the wh-phrase 'who' in (1a) and (1b). When there is both a filler and a gap (1a) or neither (1d) the result is good; when there is a filler and no gap (1b) or a gap and no filler (1c) the result is bad.[7]

(1) a. I know <u>who</u> you talked with __ yesterday. $_{(+filler,+gap)}$
b. * I know <u>who</u> you talked with Mary yesterday. $_{(+filler,-gap)}$
c. * I know <u>that</u> you talked with __ yesterday. $_{(-filler,+gap)}$
d. I know <u>that</u> you talked with Mary yesterday. $_{(-filler,-gap)}$

There is much further nuance to wh-movement, some of which we will briefly mention below. For now, let us consider how one might check if the input data are rich enough for a linguistically-neutral learner to acquire the knowledge of wh-movement. We mentioned earlier that in an ideal world, we could try to evaluate a given APS by (a) taking a sufficiently powerful learner that can be seen to not be biased in favor of the relevant knowledge; (b) training it on a developmentally-realistic corpus; and (c) checking whether the learner has indeed acquired the knowledge under consideration. We also mentioned that current LLMs do not quite handle any of (a)–(c). We will briefly review some of the shortcomings of LLMs with respect to each and then discuss how LLMs can still be helpful (if also inconclusive) in studying the APS.

2.1 Powerful and unbiased?

We do not know how powerful LLMs are. Representationally, ANNs are Turing-complete under idealized assumptions of infinite precision and computation time

[7]In order to make it possible to alternate the $\pm filler$ condition, and following WFL, we embed the relevant examples under 'I know': 'I know <u>who</u>...' ($+filler$) vs. 'I know <u>that</u>...' ($-filler$).

(Siegelmann and Sontag, 1991, 1995). Under realistic assumptions, however, the representational capacity of common ANN architectures are much more limited, as shown for example by Weiss et al. (2018) and Merrill et al. (2020) for recurrent neural networks, and by Hahn (2020) and Merrill et al. (2022) for the more recent transformer architecture. Moreover, even this limited representational capacity is often not attained in practice, and there is evidence suggesting that standard training methods prevent at least some models from acquiring key patterns (see El-Naggar et al. 2023 and Lan et al. 2022, 2023). Given these limitations we will avoid assuming that current models can learn the pattern of wh-movement that we focus on here.

The above might seem like a reason to avoid using ANNs for a study of the APS in the domain of wh-movement. As mentioned in the introduction, however, we will only rely here on the ability of ANNs to provide a reasonable approximation of wh-movement. If a given ANN can reach such an approximation from a sufficiently rich corpus, we can use it as a proxy for a good general-purpose learner, even if the ANN is not such a learner itself. We can then use the ANN to study the APS. If the model provides a reasonable approximation of wh-movement from a developmentally-realistic corpus, this suggests that a good general-purpose learner will learn the correct pattern from that corpus and that the APS in this domain does not hold. And if the model fails to reach such an approximation this suggests that a good general-purpose learner will not learn the correct pattern from that corpus and that the APS in this domain stands.

The use of ANNs as proxies still requires understanding how their biases relate to the approximations of the relevant linguistic patterns. Unfortunately, due to how poorly these models are understood, we cannot say with any certainty whether a given ANN is linguistically-neutral, and if not, whether its biases push it in the direction of a given linguistic pattern. Until more is known about these biases, and as correctly cautioned by Rawski and Heinz (2019), any claims about the neutrality of these models must be taken as tentative. Still, it strikes us as reasonable to assume that current LLMs are not particularly biased *against* the linguistic dependencies under consideration. This is especially so since these models have been developed over the past decades so as to succeed in capturing key patterns in linguistic sequences; therefore, if they do have linguistically-relevant biases after all, those are likelier to be in favor of the patterns under consideration rather than against them. Consequently, if the models fail to acquire an adequate approximation of the relevant dependencies, this failure can be taken to be informative. More directly, and as mentioned above, we will show in Section 5 that with richer training data, at least one model improves its approximation of the pattern of wh-movement, which will suggest that the failure of the model on its original training data is not due solely to its biases and other limitations but also to the lack of sufficient evidence in the data.

2.2 Training on developmentally-realistic corpora?

As discussed in detail by Warstadt and Bowman (2022), current models are not trained on developmentally-realistic corpora. Such a corpus would be the equivalent of the relevant input that a child receives over the first few years of their life. But the training data for current models are more informative than the input to the child in some ways and less informative in others. They are more informative, for example, in that they are orders of magnitude larger than what humans are exposed to in a whole lifetime. They are less informative in that they are purely textual and do not reflect prosody, environmental and social cues, and input from modalities other than speech, all of which are in principle available to children. See Warstadt and Bowman (2022) for further discussion.

The particular pattern that we discuss here can arguably be investigated on the basis of the information available in standard training corpora. Of course, this is not to say that the dependencies under consideration do not depend on such cues (a matter of ongoing discussion in the literature). But if, as WFL suggest and as our results further support, the basic pattern of wh-movement can be approximated based on text, there is no reason to think that the further approximation of parasitic gaps and across-the-board movement will crucially require extra-textual cues. This point will be reinforced by the evidence from retraining in Section 5.

As to the size and quality of the text in our training data, we use a range of corpora, reviewed immediately below, that span the spectrum from the very small (CHILDES) through mid-size (Wikipedia) to the very large (the training sets for GPT-2/3/j). We do so in an attempt to make up for the inadequacy of individual corpora to some extent, but we acknowledge that this is at best a partial remedy.

The models we use in our evaluation are the following, also summarized in Table 1: two models from Yedetore et al. (2023), an LSTM and a Transformer, that were trained on the CHILDES corpus of child-directed speech (MacWhinney, 2014);[8,9] an LSTM trained on English Wikipedia (Gulordava et al., 2018); a Transformer trained on English Wikipedia;[10] Open AI's GPT-2 (Radford et al., 2019); GPT-j (Wang and Komatsuzaki, 2021); and OpenAI's GPT-3 (Brown et al., 2020).[11] The two CHILDES-trained models, as mentioned, are taken from Yedetore et al. (2023). The LSTM trained on English Wikipedia and both GPT-2 and GPT-3 are used by WFL in their evaluation.[12]

In order to get a very rough sense of the number of years of linguistic experience that a given training corpus corresponds to we follow common practice (used also by WFL) based on Hart and Risley (1995)'s estimates about the number of words that American children typically hear during acquisition. According to these estimates, the models just mentioned were exposed to amounts of data ranging from ten months of linguistic experience (CHILDES LSTM and Transformer) through eight years of linguistic experience (Wikipedia LSTM and Transformer) to between 10 and 500 human lifetimes (GPT-2, GPT-3, and GPT-j); see Table 1. Indeed, WFL admit that the linguistic experience of some of the models is prob-

[8]The models in Yedetore et al. (2023) were trained on utterances of 52 children between the ages of six months to 12 years, from the North American English subset of the CHILDES corpus. The total training size amounts to 9.6 million words, which is considerably less than what children typically receive by the time they exhibit knowledge of the pattern under consideration here. Qualitatively, on the other hand, this training corpus is arguably more realistic than the much larger training corpora used for the remaining models.

[9]Out of ten models per architecture (LSTM/Transformer) trained in Yedetore et al. (2023) with different random seeds, we use the model with the best test perplexity.

[10]We added this Transformer since we wanted to evaluate the information in the English Wikipedia training corpus (the most realistic developmentally in terms of size of all the training corpora under consideration) using a more current architecture than the LSTM that WFL use. We used one of the large Transformer architectures used in Yedetore et al., 2023: 8 layers, hidden and embedding size 1600, and 16 attention heads, trained using the same training regime. Since the current task is limited to single sentences, we lowered the Transformer's context size to 30 (compared to 500 in Yedetore et al. 2023), closer to the average sentence size in the Wikipedia dataset (27.2).

[11]Model version 'text-davinci-003', the latest supported version not fine-tuned using reinforcement learning from human feedback (RLHF) for chat and other applications; however, the model is still trained with supervised fine-tuning, and it is proprietary. See `https://archive.today/2023.10.07-060351/https://platform.openai.com/docs/models/gpt-3-5` for OpenAI's documentation retrieved October 2023 (archived snapshot).

[12]WFL also use another LSTM, from Jozefowicz et al. (2016). We chose not to include that model in our evaluation since it is extremely slow to work with. For WFL's evaluation, which used a small number of sentences, this was not a problem, but our evaluation relied on a much larger number of sentences, making Jozefowicz et al. (2016)'s model impracticable.

Model	∼Tokens in training data	∼Human equivalent
CHILDES LSTM (Yedetore et al., 2023)	8.6 million	10 months
CHILDES Transformer (Yedetore et al., 2023)		
Wikipedia LSTM (Gulordava et al., 2018)	90 million	8 years
Wikipedia Transformer		
GPT-2 (Radford et al., 2019)	8 billion	730 years
GPT-3 (Brown et al., 2020)	114 billion	10,300 years
GPT-j (Wang and Komatsuzaki, 2021)	402 billion	36,540 years

Table 1: Training data size of the seven language models considered here, and the human linguistic experience equivalent to these data sizes; human equivalents follow WFL (based on Hart and Risley 1995) who assume a daily exposure to ∼30,000 words by children, or around 11 million words per year.

ably above and beyond that of children, and could thus weaken their argument against the APS in case of successful learning by the models. However, in the case of a negative result, as in the current work, a large training corpus strengthens the argument: if these models are exposed to amounts of data that go beyond what children are exposed to and still don't learn the constructions under consideration, this serves to strengthen the APS for these phenomena.

2.3 Inspecting LLM knowledge?

As mentioned, LLMs are very opaque. In particular, we cannot at present check whether they believe that a given continuation such as 'yesterday' or 'Mary' is grammatical following a given prefix such as 'I know who/that you talked about'. In fact, it is not clear whether current models even have a notion of grammaticality to begin with.

What LLMs do tell us is how *likely* they consider any given continuation. The problem is that grammaticality and probability are generally very different notions. And while the two are correlated — many ungrammatical continuations are also unlikely on any sensible notion of probability, and grammatical continuations are sometimes probable — this correlation is far from perfect (see Chomsky 1957, Berwick 2018, and Sprouse et al. 2018, among others, for relevant discussion). In particular, many grammatical continuations are highly unlikely; e.g., 'splat' is a grammatical but unlikely continuation of 'John would like to eat a freshly-made'. And in some cases an ungrammatical continuation can be likely; e.g., 'is' is a likely but ungrammatical continuation of 'The keys to the cabinet', an instance of so-called *agreement attraction* (see, e.g., Bock and Miller 1991 and Wagers et al. 2009).[13]

In the cases we are interested in here, however, probability and grammaticality are often quite well aligned, and — as in many other cases discussed in the literature mentioned earlier on evaluating LLMs on minimal pairs — it is easy to find examples such as (1) in which the grammatical continuation is significantly more probable than the ungrammatical one on any sensible notion of probability. So if we focus on such cases where grammaticality and probability are aligned, and if ANNs are sufficiently good learning models — at least, good enough to provide a crude approximation of the pattern under consideration — then we can use the probabilistic predictions of the resulting LLMs to evaluate the APS. If a given LLM systematically assigns a much higher probability to the grammatical continuation, this can be taken to suggest that the pattern of wh-movement is represented sufficiently well in the training data for the model to approximate it. While it remains unclear, as mentioned above, whether current ANNs themselves have a representation of grammaticality as distinct from probability or whether they can learn the true pattern, their success when trained on developmentally-realistic corpora would suggest that a good linguistically-neutral learner that does have such representational abilities might acquire the pattern.[14] Conversely, if

[13] Agreement attraction is a performance error. Speakers make such errors when distracted or in a hurry but less so when given more time. ANNs do not make this distinction: when they give a higher probability to an ungrammatical continuation their response reflects a faulty knowledge rather than a resource problem. This serves to further illustrate the inadequacy of ANNs as models of linguistic cognition but does not pose a problem for our use of these models as a tool for assessing the informativeness of the input data.

[14] Kodner and Gupta (2020) and Vázquez Martínez et al. (2023) note that success on current benchmarks of minimal pairs of the kind used below is no guarantee of human-like representations or learning. This observation is problematic for attempts to attribute to LLMs knowledge of actual linguistic patterns, but it does not affect our investigation below, which only uses LLMs as proxies for better learners and only relies on the ability of the LLMs to reach a reasonable approximation of the relevant pattern for simple cases. Even if the LLMs pass the test due to an approximation that is very different from the actual linguistic generalization, the success will still support the

the LLM does not systematically assign a much higher probability to the grammatical continuation, one potential explanation for this failure (though of course not the only one, and we discuss some potential alternatives below) is that the pattern of wh-movement is not sufficiently well represented in the input data to merit its approximation by the model. This, in turn, would suggest that a good linguistically-neutral learner will not acquire the pattern from the input data. In this way, LLMs — even if their representational inadequacies prevent them from providing more than a crude approximation of the pattern under consideration — can serve as useful proxies for future general-purpose learners and help us reason about the information available in the input data.

3 LLMs succeed in very simple cases of wh-movement

How rich is the input, then, when it comes to filler-gap dependencies of the wh kind? In very simple cases such as (1) above, the LLMs considered by WFL assign a higher probability to the grammatical continuation than to the ungrammatical one. Above we mentioned that success in cases such as those considered here, where probability and acceptability are aligned, should involve not just a higher probability to the grammatical continuation but a *much* higher one. However, in order to give the models a better chance of refuting the APS, we will adopt a very lenient criterion for success and only ask if the probability assigned to the grammatical continuation is higher than that assigned to the ungrammatical one, without taking into account how much higher it is. This will allow a network to be considered successful even if it prefers the grammatical continuation by the slightest of margins. This lenient condition for success will strengthen our conclusions from cases of failure, which we get to in the sections below: if a network fails even with this lenient condition of success, this failure can be taken seriously.

Here and below we will follow WFL (and the psycholinguistic literature that they build on) and illustrate using *surprisal* values, where the surprisal of x is $-\log P(x)$, which is simply the negative of the logarithmically-scaled probability of x.[15] The lower the probability the higher the surprisal; when the probability ap-

notion that the pattern is sufficiently well represented in the training data for a good model to acquire it.

[15]WFL's methodology includes looking not just at $+filler$ cases, as in (1a) and (1b), but also at the corresponding $-filler$ ones, as in (1c) and (1d). We will follow WFL in this in our discussion in sections 4.3 and 5 below, but for the present we will attempt to keep the presentation simple by

proaches 0 the surprisal tends to infinity, and as the probability approaches 1 the surprisal tends to 0. Since higher probability corresponds to lower surprisal, support for the model will come from its assigning lower surprisal to a grammatical continuation than to an ungrammatical one, which, as mentioned, is what WFL indeed find in simple cases.

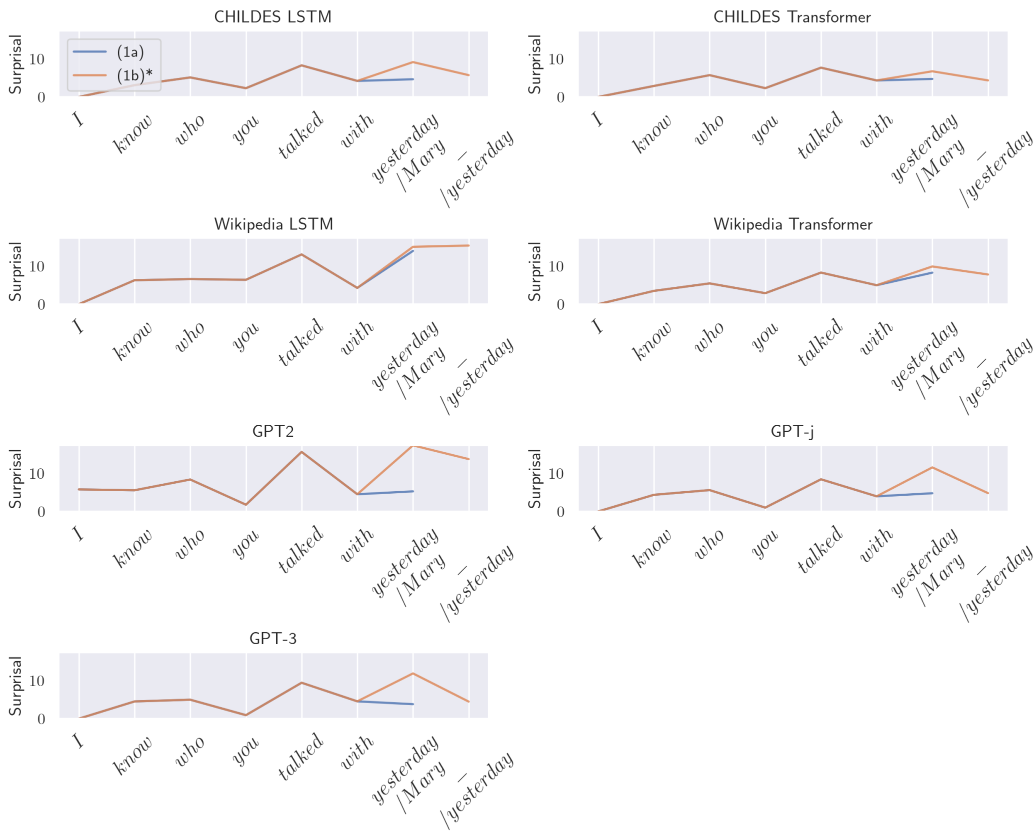

Figure 1: Raw surprisal values outputted by the LLMs for the grammatical (1a), in blue, and ungrammatical (1b), in orange. All models correctly output lower surprisal values for the grammatical continuation.

Figure 1 illustrates the preference of the models considered here for the grammatical continuation over the ungrammatical one in a very simple case by plotting surprisal values for sentences (1a) and (1b). All models assign a lower surprisal value (i.e., a higher probability) to the grammatical continuation 'yesterday' in the

considering only $+filler$ pairs.

gapped sentence than to ‘Mary’. This suggests that the input is sufficiently rich for a general-purpose learner to acquire from it an approximation of some basic aspects of wh-movement.

WFL further suggest that the LLMs go beyond the basic knowledge that fillers and gaps go hand in hand. Specifically, they claim that LLMs are aware of *islands* (Ross, 1967): configurations in which a gap is bad even if there is a filler upstream. We illustrate this with the following:

(2) * I know who [[the question whether __ jumped] surprised Mary yesterday].

While, as discussed above, a filler upstream generally increases the LLMs’ expectation of a gap downstream, this expectation should be reduced within the subject of the embedded clause in (2). This subject is an island to movement, and extraction from within it is unacceptable and presumably highly unlikely. Figure 2 shows that the models are indeed surprised by the gap in (2), suggesting that their training corpora are informative with respect to this aspect of wh-movement.[16]

WFL consider a range of similar cases and conclude that linguistically-neutral learners can acquire the intricacies of wh-movement from the input data. In other words, the input data are not impoverished after all with respect to wh-movement, and an APS in this domain falls apart.

4 LLMs fail on slightly more complex (but still simple) cases of wh-movement

We now turn to a well-studied nuance of islands: in various cases, an otherwise impossible gap inside an island is made possible by a separate gap elsewhere. For example, while (3a), with a subject-internal gap, is bad, its counterpart in (3b), which has an added gap in the direct-object position of the main clause, is good. This phenomenon is known as a *parasitic gap* (PG): the gap inside the subject island becomes acceptable parasitically, based on the direct-object gap.[17]

[16]The literature discusses various cases in which extraction from subjects (and other islands) is judged acceptable by speakers. Here and below we focus on relatively simple examples in which speaker judgments are clear, and our evaluation will concern the extent to which LLM preferences approximate these clear speaker judgments.

[17]Not all impossible gaps can be rescued in this way. For example, adding further gaps does little to improve (2) above.

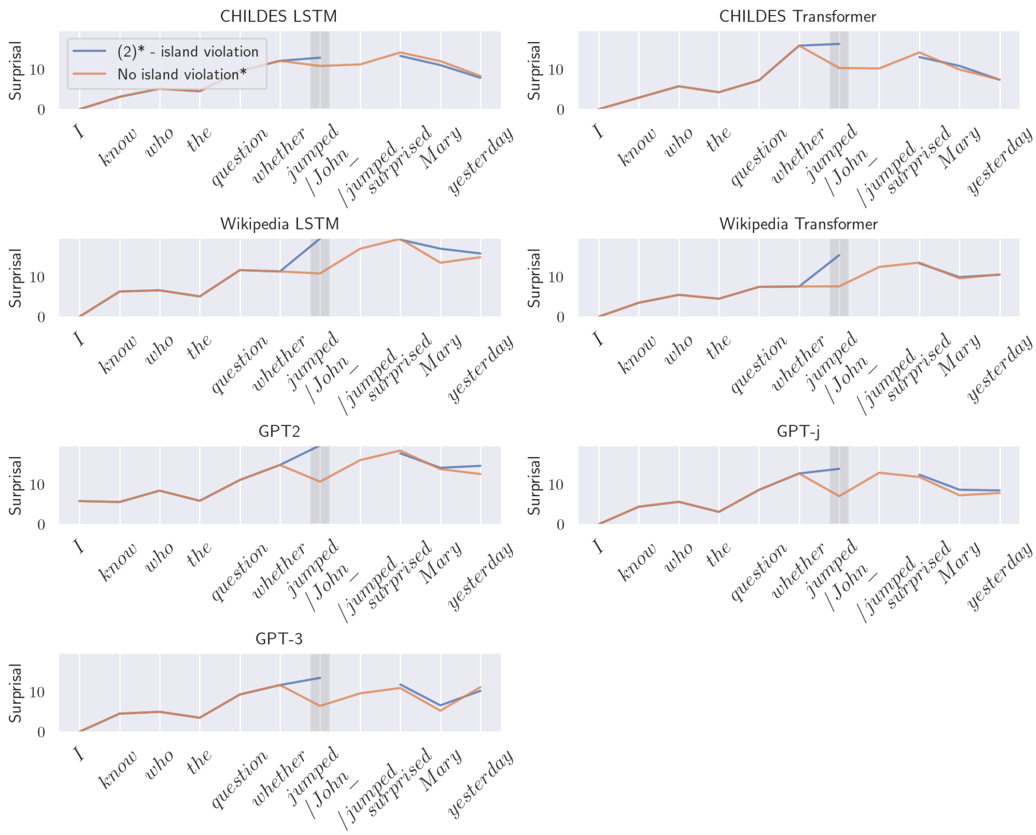

Figure 2: Raw surprisal values for the island violation sentence in (2), in blue, and a variant of the sentence with no island violation (we use 'John' instead of the island-internal gap), in orange. All models are correctly surprised to find a gap within the island. Note that since the variant with 'John' has no downstream gap that would correspond to the upstream filler, it is ungrammatical. For a grammatical version one could replace 'Mary' with a gap. This matter, however, is orthogonal to the surprisal at the island-internal gap, which is what this figure illustrates.

(3) a. * I know who [John's talking to __] is going to annoy you soon.
b. I know who [John's talking to __] is going to annoy __ soon.

Somewhat similarly, while (4a), with a gap inside a conjunct, is bad, its counterpart in (4b), where there is a gap in the other conjunct as well, is good. This phenomenon is known as *across-the-board movement* (ATB).[18]

(4) a. * I know who John [met __ recently] and [is going to annoy you soon].

b. I know who John [met __ recently] and [is going to annoy __ soon].

4.1 An initial failure

Do LLMs approximate the patterns of PG and ATB? Both Wilcox et al. (2018) and Chaves (2020) mention PG and ATB in passing, but we are not familiar with attempts in the literature to evaluate the success of LLMs in approximating these patterns. Figures 3-4 illustrate that all the LLMs that we are considering here fail on (3a) and (4a), even on our very lenient condition of success: they do not just fail to assign a much higher probability to the grammatical continuation over the ungrammatical one in this simple case; they actually prefer the *ungrammatical* continuation. This seems to indicate that the ANNs have failed to acquire a good approximation of the relevant constructions. This challenges WFL's claim that LLMs undo the APS in this domain: for LLMs to undo this APS, they would need to provide a passable approximation of PG and ATB, but their performance above does not suggest such an approximation.

If our entire empirical basis is the failure we just saw, however, our conclusions will remain weak. This is so for the following reason: while the behavior of a good linguistically-neutral learner on the examples above would indeed be informative about the APS, it is possible that current ANNs are simply not sufficiently good learners in this regard, and the inadequacies of the ANNs can in turn significantly limit our conclusions.

In the remainder of the present section we attempt to address the general concern about the adequacy of the ANNs, which we break down into two separate investigations. We first ask whether the failure that we just saw is an accident of the particular lexical choices that we used (Section 4.2). We then ask, building on WFL's methodology, whether the failure was due to a general preference for ungapped continuations that is so strong as to override a preference for the correct form (Section 4.3). Our investigations concern ways in which the models might have an approximation of PG and ATB that is obscured by weaknesses of

[18] We set aside the important question of what stands behind PGs and ATB and whether the two are related. See Ross (1967), Engdahl (1983), Haïk (1985), Williams (1990), Munn (1992), Postal (1993), Nissenbaum (2000), and Hornstein and Nunes (2002), among others, for discussion.

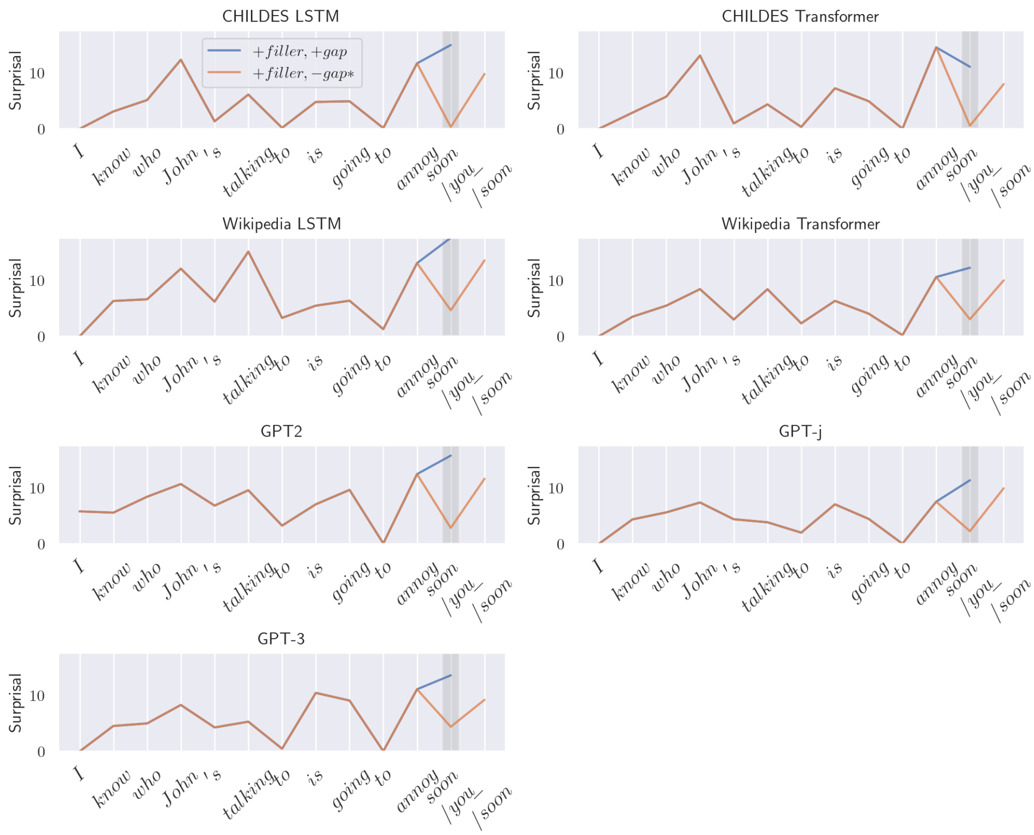

Figure 3: Raw surprisal values for the ungrammatical sentence (3a) which violates a subject island, in orange, and its grammatical variant (3b), in blue, where a parasitic gap makes it possible to escape the island. For measuring the model's expectation for a gap, surprisal is measured at the adverb 'soon', which indicates a gap. This is compared with surprisal at 'John' which plugs the gap at the same position. All networks wrongly assign a higher surprisal value to the grammatical continuation.

the models. By helping these models at test we aim to reveal this approximation if it exists, but in both sections we will fail to find evidence for it. This, in turn, will strengthen the challenge to WFL's claim: even with additional help at test, the models show no evidence that might undermine the APS.

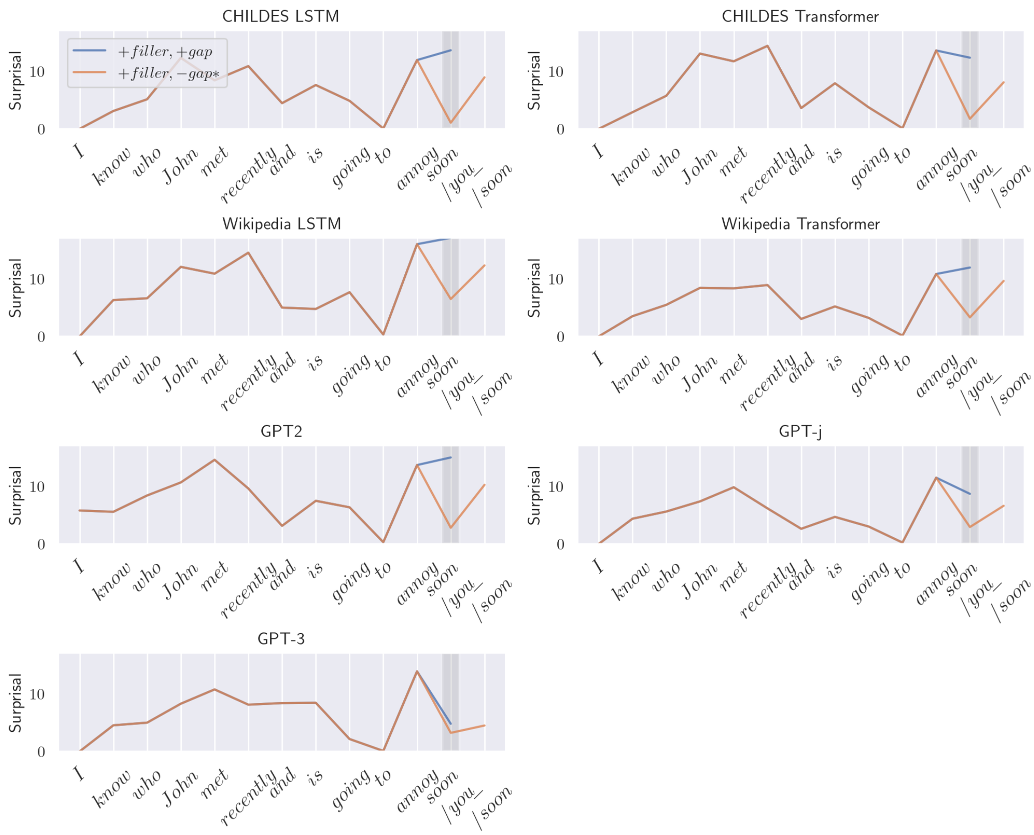

Figure 4: Raw surprisal values for the ungrammatical sentence (4a) which violates the coordinate structure constraint (orange), and its grammatical variant (4b) where ATB movement makes it possible to avoid the constraint (blue). All networks wrongly assign a higher surprisal value to the grammatical continuation ‘soon’ rather than to ‘John’.

4.2 Lexical accident?

Our illustration above of how the LLMs prefer the ungrammatical continuation over the grammatical one for ATB and PG used one pair of sentences for each of the two patterns. This raises the obvious worry that the failure of the LLMs reflects some accidents of the specific sentences that we used. This worry is lessened to some extent by the fact that we looked at a broad range of different models trained on different corpora: it seems unlikely that all these models and all these training corpora just happen to have the same blind spot when it comes to the specific sentences that we used above and that otherwise the models approximate the patterns well. Still, it is clearly useful to examine more systematically what happens when we vary the lexical choices for the two patterns.

In order to test the performance of the networks on PG and ATB sentences more broadly, we systematically varied the lexical choices in (3) and (4), repeated here.

(5) a. * I know who [John's talking to __] is going to annoy you soon.
b. I know who [John's talking to __] is going to annoy __ soon.

(6) a. * I know who John [met __ recently] and [is going to annoy you soon].
b. I know who John [met __ recently] and [is going to annoy __ soon].

We generated the sentences by template, using simple context-free grammars. Excerpts from these grammars and a sample of the generated sentences are given in Tables 2 and 3. The full grammars are given in Appendix A.[19] 8,064 sentence tuples were generated for PG and 6,624 for ATB. For a given model and a given pair of sentences, we looked at the surprisal of the model at the critical point on each member of the pair. For (5), for example, we checked whether after the shared prefix "I know who [John's talking to __] is going to annoy ..." surprisal was higher at the ungapped, ungrammatical continuation 'you' as in (5a) than in the gapped, grammatical continuation 'soon' as in (5b). If it was, and in line with our lenient condition for success that is satisfied by any kind of preference for the grammatical continuation regardless of its magnitude, this counted as a success. We will write $\Delta = Surprisal(\text{ungapped continuation}|\text{shared prefix}) - Surprisal(\text{gapped continuation}|\text{shared prefix})$, and we will write $\Delta_{+filler}$ to indicate that the shared prefix has an upstream filler. Using this notation, we can write the condition for success as $\Delta_{+filler} > 0$.

Figure 5 plots the results of examining $\Delta_{+filler}$ preferences for the PG and ATB datasets. In both cases, the best performance by a large margin is that of GPT-3, with 40.9% accuracy on the PG dataset and 71.6% accuracy on the ATB dataset. We are not sure to what extent these numbers can be taken to indicate an approximation of the relevant patterns. If it is a success then it is hardly a striking one. Nor is it particularly informative: recall that GPT-3 has been trained

[19] All experimental material, artificial grammars, and training and test data, as well as the source code, will be published as supplementary material once the paper can be de-anonymized. We will also be happy to share the material with referees anonymously during the review process.

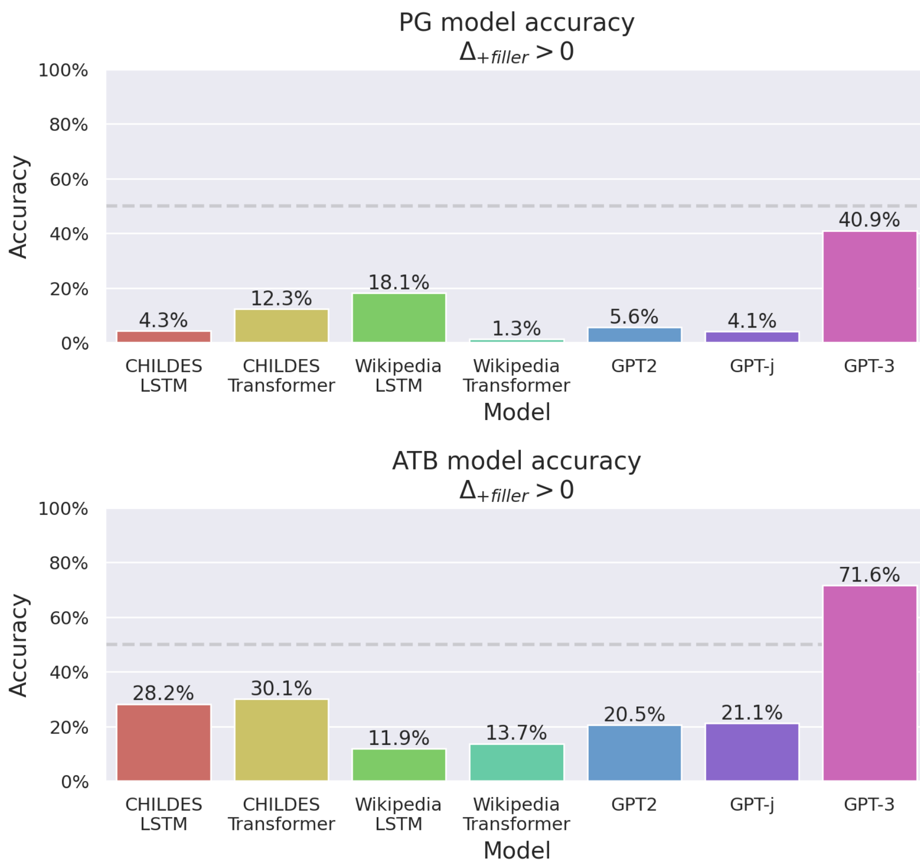

Figure 5: Model accuracy on the $\Delta{+}filler$ condition for the PG and ATB datasets. Accuracy is measured as the ratio of cases where the model assigns a higher probability to the grammatical sentence continuation.

on the equivalent of 10,000 years of linguistic experience (and is also further improved manually in various ways), so even if it approximates the relevant patterns, this does not indicate that a general-purpose learner would acquire the relevant knowledge from a developmentally-realistic corpus of just a few years of linguistic experience. Setting GPT-3 aside, the models perform very poorly, with the best performance on PG being Wikipedia LSTM's 18.1% accuracy and the best performance on ATB being CHILDES Transformer's 30.1% accuracy. In other words, the models do not just fail to prefer the grammatical continuation over the ungrammatical one, they positively prefer the ungrammatical continuation in the vast majority of the pairs. Helping the LLMs by testing them on a wide range of lexical choices, then, fails to reveal any evidence that the models have approximated the patterns of parasitic gaps and across-the-board movement.

4.3 A preference for ungapped continuations?

Our second investigation, building on WFL's methodology, asks whether the networks have a local preference for or against gapped continuations that might make them succeed or fail for the wrong reasons.

Consider again (5) (="I know who [John's talking to __] is going to annoy *you/✓__ soon"). A sufficiently strong local preference about the critical area can affect a given ANN's success regardless of whether it has acquired any approximation of PG, or of wh-movement in general. It could be, for example, that the ANN assigns a higher probability to the grammatical continuation 'soon' than to the ungrammatical 'you' but that it does so because it ignores the filler ('who') altogether and simply prefers 'annoy soon' to 'annoy you'. Conversely, it is conceivable that the ANN has, in fact, acquired knowledge of wh-movement but that it incorrectly prefers 'you' to 'soon' because of similarly irrelevant reasons. For example, perhaps it has a strong preference for ungapped continuations in general, or perhaps it has such a preference in the present case because the lexical frequency of 'you' is very high.

To what extent might such local preferences affect the ANNs? We are not entirely sure. A good enough learner would presumably not get confused by such irrelevant factors, and the fact that all our models perform well on very simple filler-gap dependencies illustrated in Figures 1 and 2 is at least suggestive of their ability to overcome any such confusion when the training data are sufficiently rich. However, beyond this suggestive evidence it is hard to tell whether current ANNs are good enough learners in this sense, and it strikes us as reasonable to further investigate possible confusion by irrelevant factors that might override the preference for the correct pattern.

Following WFL, we will explore the possible effect of irrelevant factors of the kind just mentioned by looking at each LLM's preference for gapped over ungapped continuations and comparing this preference when there is an upstream filler and when there is no such filler. When an upstream filler is present, the model's preference for a gapped continuation (e.g., 'annoy soon') over the ungapped continuation ('annoy you') should be stronger than when an upstream filler is absent. In other words, we will be looking at whole paradigms of the shape we already saw in (1) and not just at those portions of the paradigm in which a filler is present. Such a paradigm is illustrated for PG in (7) and for ATB in (8). Underlined words indicate the ±filler alternations. Words in bold indicate the critical region that shows whether the continuation is gapped or not.

(7) PG

	$+gap$	$-gap$
$+filler$	I know who John's talking to is going to annoy **soon**.	*I know who John's talking to is going to annoy **you** soon.
$-filler$	*I know that John's talking to Mary is going to annoy **soon**.	I know that John's talking to Mary is going to annoy **you** soon.

(8) ATB

	$+gap$	$-gap$
$+filler$	I know who John met recently and is going to annoy **soon**.	*I know who John met recently and is going to annoy **you** soon.
$-filler$	*I know that John met Bob recently and is going to annoy **soon**.	I know that John met Bob recently and is going to annoy **you** soon.

Extending our lenient condition for success used above, we will now consider it a success for a given model on a particular paradigm if its preference for the gapped continuation (regardless of its absolute magnitude or even its sign) is higher in the presence of an upstream filler than in its absence. Above we introduced the notation $\Delta = Surprisal$(ungapped continuation|shared prefix)$-Surprisal$(gapped continuation|shared prefix) for the extent of the preference for the gapped continuation over the ungapped continuation, and we wrote $\Delta_{+filler}$ when the shared prefix had an upstream filler. We will now consider also the analogous $\Delta_{-filler}$, for the part of the paradigm where the shared prefix does not have an upstairs filler. And we will consider it a success for the model if $\Delta_{+filler} > \Delta_{-filler}$. This lenient condition of cross-paradigm success follows the logic of difference-in-differences and is very much in line with WFL's evaluation.[20]

In order to test the models across a large number of paradigms, with many different lexical choices, we used the same grammars mentioned in the previous

[20]Of course, this new criterion still allows for various irrelevant factors to affect success. For example, a model could become successful simply by deciding that 'who' corresponds to a high probability for 'soon' and a low probability for 'you' anywhere in the sentence and that 'that' corresponds to the opposite. We set aside such worries here.

section. In our earlier discussion we used the $+filler$ pairs generated by the grammar. In the present section we use also the corresponding $-filler$ pairs, and from each paradigm of $+filler$ and $-filler$ pairs we compute $\Delta_{\pm filler}$ values. Excerpts from the grammars are provided in Table 2 (for PGs) and Table 3 (for ATB).

PG Grammar
$S \rightarrow \langle PREAMBLE \rangle\ \langle \pm F \rangle\ \langle \pm G \rangle$
$\langle PREAMBLE \rangle \rightarrow$ *I know*
$\langle +F \rangle \rightarrow$ *<u>who</u>* $\langle NAME1 \rangle \langle GEN \rangle\ \langle NP \rangle$
$\langle -F \rangle \rightarrow$ *<u>that</u>* $\langle NAME1 \rangle \langle GEN \rangle\ \langle NP \rangle\ \langle \underline{NAME2} \rangle$
$\langle +G \rangle \rightarrow \langle CONN \rangle\ \langle V \rangle$ ⟨**ADV**⟩
$\langle -G \rangle \rightarrow \langle CONN \rangle\ \langle V \rangle$ ⟨**OBJ**⟩ $\langle ADV \rangle$
$\langle GEN \rangle \rightarrow$ *'s*
$\langle NP \rangle \rightarrow \langle NP\_SIMPLE \rangle \mid \langle NP\_COMPLEX \rangle$
$\langle NP\_SIMPLE \rangle \rightarrow \langle GERUND \rangle$
$\langle NP\_COMPLEX \rangle \rightarrow \langle N\_EMBEDDED \rangle$ 'to' $\langle V\_EMBEDDED \rangle$
$\langle CONN \rangle \rightarrow$ 'is about to' | 'is likely to' | 'is going to' | 'is expected to'
$\langle V \rangle \rightarrow$ 'bother' | 'annoy' | 'disturb'
$\langle OBJ \rangle \rightarrow$ 'you' | 'us' | 'Kim'
$\langle GERUND \rangle \rightarrow$ 'talking to' | 'dancing with' | 'playing with'
$\langle N\_EMBEDDED \rangle \rightarrow$ 'decision' | 'intent' | 'effort' | 'attempt' | 'failure'
$\langle V\_EMBEDDED \rangle \rightarrow$ 'talk to' | 'call' | 'meet' | 'dance with' | 'play with'
$\langle ADV \rangle \rightarrow$ 'soon' | 'eventually'
...
⇒ I know <u>who</u> John's talking to is going to annoy **soon**. $_{(+filler,+gap)}$
⇒ * I know <u>who</u> John's talking to is going to annoy **you** soon. $_{(+filler,-gap)}$
⇒ * I know <u>that</u> John's talking to <u>Mary</u> is going to annoy **soon**. $_{(-filler,+gap)}$
⇒ I know <u>that</u> John's talking to <u>Mary</u> is going to annoy **you** soon. $_{(-filler,-gap)}$

Table 2: Excerpt from the context-free grammar used to generate PG sentences for the experiments in Section 4.3, and sample sentences generated from it. Underlined words alternate according to the $\pm filler$ condition; words in bold mark the position where the $\pm gap$ condition becomes evident and surprisal is measured.

Figure 6 plots the LLMs' performance for the cross-paradigm (difference-in-

ATB Grammar

$S \rightarrow \langle PREAMBLE \rangle \; \langle \pm F \rangle \; \langle CONN \rangle \; \langle \pm G \rangle$
$\langle PREAMBLE \rangle \rightarrow$ *I know*
$\langle +F \rangle \rightarrow$ <u>*who*</u> $\langle NAME1 \rangle \; \langle VP1 \rangle \; \langle ADV1 \rangle$
$\langle -F \rangle \rightarrow$ <u>*that*</u> $\langle NAME1 \rangle \; \langle VP1 \rangle \; \langle \underline{NAME2} \rangle \; \langle ADV1 \rangle$
$\langle +G \rangle \rightarrow \langle CONN \rangle \; \langle VP2 \rangle$ ⟨**ADV2**⟩
$\langle -G \rangle \rightarrow \langle CONN \rangle \; \langle VP2 \rangle$ ⟨**OBJ**⟩ $\langle ADV2 \rangle$
$\langle CONN \rangle \rightarrow$ ‘and is going to’
$\langle ADV1 \rangle \rightarrow$ ‘recently’ | ‘lately’
$\langle ADV2 \rangle \rightarrow$ ‘soon’ | ‘today’ | ‘now’
$\langle CONN \rangle \rightarrow$ ‘and is going to’
$\langle VP1 \rangle \rightarrow \langle VP1\_SIMPLE \rangle \mid \langle VP1\_COMPLEX \rangle$
$\langle VP1\_SIMPLE \rangle \rightarrow$ ‘met’ | ‘saw’
$\langle VP2 \rangle \rightarrow \langle VP2\_COMPLEX \rangle \mid \langle VP2\_SIMPLE \rangle$
$\langle VP2\_SIMPLE \rangle \rightarrow$ ‘hug’ | ‘slap’ | ‘kiss’
$\langle OBJ \rangle \rightarrow$ ‘you’ | ‘us’ | ‘Kim’
...
⇒ I know <u>who</u> John met recently and is going to hug **soon**. $(+filler, +gap)$
⇒ *I know <u>who</u> John met recently and is going to hug **you** soon. $(+filler, -gap)$
⇒ *I know <u>that</u> John met <u>Bob</u> recently and is going to hug **soon**. $(-filler, +gap)$
⇒ I know <u>that</u> John met <u>Bob</u> recently and is going to hug **you** soon. $(-filler, -gap)$

Table 3: Excerpt from the context-free grammar used to generate ATB sentences for the experiments in Section 4.3, and sample sentences generated from it. Underlined words alternate according to the $\pm filler$ condition; words in bold mark the position where the $\pm gap$ condition becomes evident and surprisal is measured.

differences) condition. All models except CHILDES LSTM have higher scores for the present measure of $\Delta_{+filler} > \Delta_{-filler}$ than they did for the earlier measure of $\Delta_{+filler} > 0$ (Figure 5), and this holds for both PG and ATB. However, we can see that only GPT-j and GPT-3 obtain scores that are convincingly high. But GPT-j is trained on the equivalent of 500 lifetimes of human linguistic exposure, and GPT-3 is trained on the equivalent of 100 lifetimes and fine-tuned further on downstream language tasks. Even GPT-2, trained on the equivalent of ten lifetimes — and thus two orders of magnitude at least above what a child hears by the time they have knowledge of PG and ATB — only reaches modest scores, below 80%. And the smaller models obtain much lower scores. This includes the English Wikipedia LSTM and Transformer, whose training corpus corresponds to about 8 years of linguistic exposure, arguably the most realistic developmentally in terms of size of all the models.

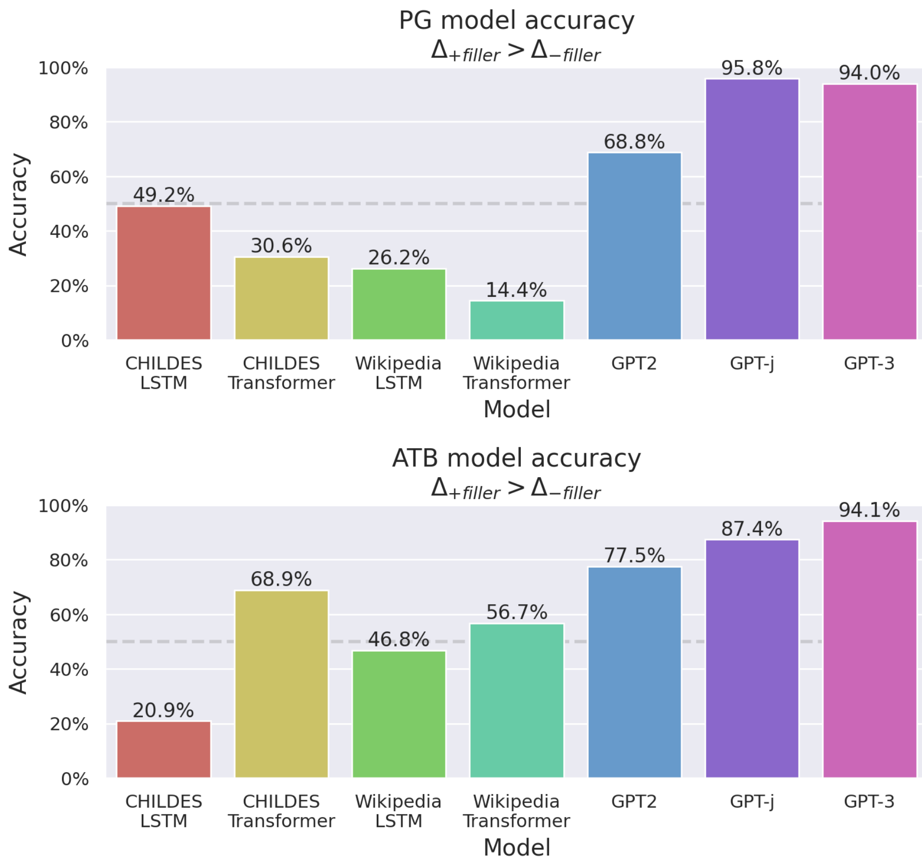

Figure 6: Model accuracy on the difference-in-differences condition for the PG and ATB datasets. Accuracy is measured as the ratio of cases where $\Delta{+}filler >$ $\Delta{-}filler$, i.e., when the model shows a relative higher preference for a gap when the gap follows a filler than when it does not.

The gradual improvement of LLM scores as the corpora become very large suggests that current models are in principle capable of improving their approximation of the pattern of wh-movement, but also that this improvement requires much more information than what is present in a corpus that corresponds to anything a child might encounter. We return to the potential of richer training data to improve an LLMs approximation of the patterns under consideration in the next section. In the meantime we conclude that even with considerable help at test, the performance of the LLMs provides no evidence against the APS.

5 A general inability to acquire a suitable preference?

Recall WFL's contention that LLMs show that a linguistically-neutral learner can acquire knowledge of wh-movement from a realistic corpus. In the face of our results from the previous section, WFL's claim needs to be abandoned: current LLMs provide no basis for such a conclusion. Of course, this is not the same as saying that LLMs provide evidence *for* the APS: the failure of the LLMs might be due entirely to their own limitations and not be informative about the richness of the training corpora. In the present section, however, we will go one step further and provide tentative evidence that the failure of the LLMs is due also to the insufficient richness of the training corpora and not just to weaknesses of the ANNs. We do so by helping one of our models at training: we retrain the Wikipedia Transformer model on an enriched corpus that includes multiple instances of PG and ATB. As we show, the performance of the model improves significantly on the enriched corpus, suggesting that the failure on the original corpus reflects the poverty of that corpus.

The additional instances for the enriched training corpus are generated by template, using a variant of the CFGs that we used in sections 4.2 and 4.3. To increase the probability that an improved performance by the model will reflect generalization rather than memorization, the structure of the additional instances is different from those of the test sentences from Section 4.3. Example training and test sentences are given in Table 4. The full CFGs used in creating the additional instances are given in Appendix C.

From each CFG of each phenomenon (PG/ATB), we sampled 100 sentences for the two grammatical conditions ($+filler, +gap$ and $-filler, -gap$), totaling 200 extra sentences. These sentences were added to the original English Wikipedia dataset, and the model was trained using the same regime as in Yedetore et al. (2023) (itself based on the training regime in Gulordava et al., 2018). The model was trained for 48 hours or until reaching the early-stop condition from

PG	**Training Examples** • I know who John's attitude towards upset yesterday. • I know who John's friendship with will annoy soon. • I know who John's praising of amused lately. **Test Examples** • I know who John's talking to is about to bother soon. • I know who John's playing with is going to annoy eventually. • I know who John's failure to dance with is going to disturb soon.
ATB	**Training Examples** • I know who John saw yesterday and kissed today. • I know who John helped recently and married today. • I know who John hugged often and will insult soon. **Test Examples** • I know who John met recently and is going to complain to Patricia about soon. • I know who John said that Mary saw lately and is going to be glad to hug now. • I know who John asked Mary about lately and is going to claim that Patricia will hug today.

Table 4: Example training and test sentences for the retraining task in Section 5. A sample of simplified training PG and ATB sentences are added to the model's original training data (English Wikipedia), and the model is then tested on the full battery of sentences from Section 4.3. The full CFGs for the training and test datasets are given in Appendices A and C.

Yedetore et al. (2023) which stops the training if the validation loss does not improve for more than two consecutive epochs. Due to the long training times of the model, the results reported here are for one random seed with no hyper-parameters search. Since the goal of this experiment was to demonstrate the model's ability to improve significantly given more data, this was sufficient.

The model's performance on the training and test set, before and after retraining, is visualized in Figure 7.

For both ATB and PG the performance of the model improves significantly. For ATB, the raw $\Delta_{+filler} > 0$ accuracy score improves from 13.7% to 35.8%, and the difference-in-differences $\Delta_{+filler} > \Delta_{-filler}$ score improves from 56.7% to 97.2%. This is a dramatic improvement over the performance of the model on

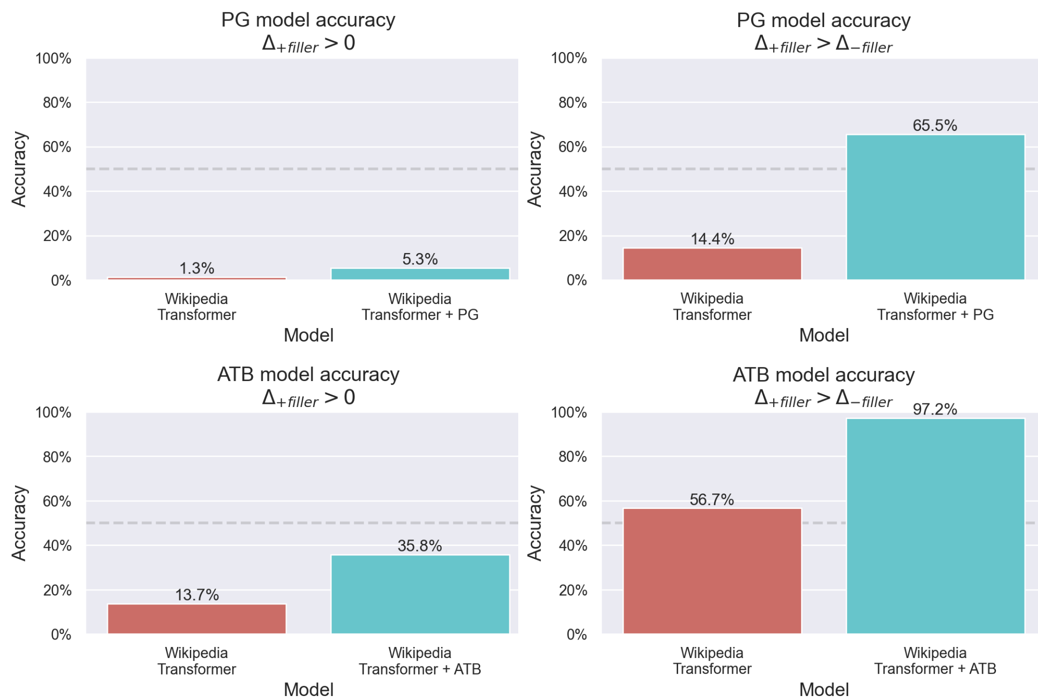

Figure 7: Accuracy for the retrained Transformer model, when trained on the original Wikipedia vs. when trained on the same dataset with extra PG and ATB sentences. The left figures plot accuracy for the $+filler$ condition, and the right figures plot accuracy for the difference-in-difference condition.

its original corpus and is higher than the performance of other architectures when trained on a much larger corpus. For PG, the raw score improves modestly, from 1.3% to 5.3%, while the difference-in-differences score improves from 14.4% to 65.5%. The raw score for PG certainly doesn't inspire confidence that the model has acquired the dependency. Recall, however, that this is not what we were after here. Our question was whether the model is so weak that its poor performance when trained on the original corpus reflects its inability to do better. The retraining results show that the model can do considerably better once the corpus is sufficiently rich.

Caution is required in interpreting this result. Like all current LLMs, our model is opaque, and we are limited in the conclusions that we can draw from it. In particular, while we observe that it has improved when retrained on a corpus that is enriched in a certain way, it is possible that there are other kinds of evidence for the patterns under consideration that a good general-purpose learner would be able to make use of and that our model cannot. What we found is consistent with such evidence being in the original corpus. Our use of retraining data that were structurally different from the test data was aimed at lessening this worry, since improvement suggests an ability to generalize and not just memorize. This, in turn, increases the plausibility that the model would have been able to generalize from other kinds of relevant examples if the original corpus had been sufficiently rich. But the opacity of the model prevents us from saying more, and our results here must be qualified accordingly.

6 Conclusion

The APS has been central to linguists' reasoning about innateness for a long time. It has always been difficult, however, to estimate just how much information a linguistically-neutral learner might hope to extract from a realistic input. Modern ANNs promise to change this, and their linguistic knowledge and learning have been the topic of research of a growing literature. We focused here on WFL, who use LLMs to argue that the stimulus is rich enough when it comes to wh-movement and that this dismantles the APS in this domain. We showed that this conclusion is premature: by looking at parasitic gaps and across-the-board movement we showed that several ANNs fail to reach a passable approximation of the pattern of wh-movement.

Is it possible that some future linguistically-neutral learner will succeed where the models that we have examined have failed? Of course. As we mentioned, current models are too opaque and too poorly understood (and current training corpora are too unrealistic developmentally) to definitively settle the question of whether the APS for wh-movement holds. We note, however, that the architectures we have considered are generally successful in approximating many other aspects of linguistic data and that we evaluated the models using an extremely lenient criterion for success. And some of the models have been provided with very generous amounts of linguistic input, in some cases several orders of magnitude beyond what children receive. Given that none of the ANNs reached an adequate approximation of the pattern for the relatively simple examples that we have considered — and given that at least one network did seem capable of improving its approximation when retrained on a clearly rich corpus — we find it likelier that the stimulus is simply too poor to warrant the acquisition of the relevant aspects of knowledge from a corpus that is even remotely realistic developmentally by a linguistically-neutral learner. And if that turns out to be the case, adult speakers' knowledge of these aspects would mean that children are innately endowed in ways that are not linguistically neutral.

References

Baker, C. L. 1978. *Introduction to generative transformational syntax*. Englewood Cliffs, New Jersey: Prentice-Hall.

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

Berwick, Robert C. 2018. Revolutionary new ideas appear infrequently. In *Syntactic Structures after 60 years: The impact of the Chomskyan revolution in linguistics*, ed. Norbert Hornstein, Howard Lasnik, Pritty Patel-Grosz, and Charles Yang, volume 60, 177–194. Walter de Gruyter.

Berwick, Robert C., Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science* 35:1207–1242.

Bhattacharya, Debasmita, and Marten van Schijndel. 2020. Filler-gaps that neural networks fail to generalize. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, ed. Raquel Fernández and Tal Linzen, 486–495. Online: Association for Computational Linguistics.

Bock, Kathryn, and Carol A Miller. 1991. Broken agreement. *Cognitive Psychology* 23:45–93.

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

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

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

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

Chomsky, Noam. 1971. *Problems of knowledge and freedom: The Russell lectures*. Pantheon Books.

Chomsky, Noam. 1975. *Current issues in linguistic theory*. The Hague: Mouton.

Chomsky, Noam. 1980. *Rules and representations*. New York: Columbia University Press.

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. Emily M. Bender, Leon Derczynski, and Pierre Isabelle, 133–144. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Dyer, Fred C., and Jeffrey A. Dickinson. 1994. Development of sun compensation by honeybees: How partially experienced bees estimate the sun's course. *Proceedings of the National Academy of Sciences* 91:4471–4474.

El-Naggar, Nadine, Andrew Ryzhikov, Laure Daviaud, Pranava Madhyastha, and Tillman Weyde. 2023. Formal and empirical studies of counting behaviour in relu rnns. In *Proceedings of 16th edition of the International Conference on Grammatical Inference*, ed. François Coste, Faissal Ouardi, and Guillaume Rabusseau, volume 217 of *Proceedings of Machine Learning Research*, 199–222. PMLR.

Elman, Jeffrey L., Elizabeth Bates, M. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. 1996. *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: The MIT Press.

Engdahl, E. 1983. Parasitic gaps. *Linguistics and Philosophy* 6:5–34.

Foraker, Stephani, Terry Regier, Naveen Khetarpal, Amy Perfors, and Joshua Tenenbaum. 2009. Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science* 33:287–300.

Gordon, Peter. 1985. Level-ordering in lexical development. *Cognition* 21:73–93.

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

Proceedings of NAACL 2018, 1195–1205.

Hahn, Michael. 2020. Theoretical Limitations of Self-Attention in Neural Sequence Models. *Transactions of the Association for Computational Linguistics* 8:156–171.

Haïk, Isabelle. 1985. The syntax of operators. Doctoral Dissertation, MIT.

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

Hornstein, Norbert, and Jairo Nunes. 2002. On asymmetries between parasitic gap and across-the-board constructions. *Syntax* 5.

Hsu, Anne S., and Nick Chater. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science* 34:972–1016.

Huebner, Philip A., Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, ed. Arianna Bisazza and Omri Abend, 624–646. Online: Association for Computational Linguistics.

Jozefowicz, Rafal, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs]* .

Katzir, Roni. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics* 17.

Kodner, Jordan, and Nitish Gupta. 2020. Overestimation of syntactic representation in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1757–1762. Online: Association for Computational Linguistics.

Kodner, Jordan, Sarah Payne, and Jeffrey Heinz. 2023. Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023). ArXiv preprint arXiv:2308.03228.

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

Lan, Nur, Emmanuel Chemla, and Roni Katzir. 2023. Benchmarking neural network generalization for grammar induction. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, 131–140. Gothenburg, Sweden: Association for Computational Linguistics.

Lan, Nur, Michal Geyer, Emmanuel Chemla, and Roni Katzir. 2022. Minimum description length recurrent neural networks. *Transactions of the Association for Computational Linguistics* 10:785–799.

Legate, Julie Anne, and Charles Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review* 19.

Lewis, John D, and Jeffery L Elman. 2001. A connectionist investigtion of linguistic arguments from the poverty of the stimulus: Learning the unlearnable. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

Lidz, Jeffrey, Sandra Waxman, and Jennifer Freedman. 2003. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition* 89:B65–B73.

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.

MacWhinney, Brian. 2014. *The Childes Project: Tools for Analyzing Talk, Volume II: The Database*. New York: Psychology Press, 3 edition.

Marvin, Rebecca, and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031* .

Merrill, William, Ashish Sabharwal, and Noah A. Smith. 2022. Saturated Transformers are Constant-Depth Threshold Circuits. *Transactions of the Association for Computational Linguistics* 10:843–856.

Merrill, William, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A Smith, and Eran Yahav. 2020. A formal hierarchy of rnn architectures. *arXiv preprint arXiv:2004.08500* .

Moro, Andrea, Matteo Greco, and Stefano F. Cappa. 2023. Large languages, impossible languages and human brains. *Cortex* 167:82–85.

Munn, Alan. 1992. A null operator analysis of ATB gaps. *The Linguistic Review* 9:1–26.

Nissenbaum, Jon. 2000. Investigations of covert phrase movement. Doctoral Dissertation, MIT, Cambridge, Mass.

Ozaki, Satoru, Dan Yurovsky, and Lori Levin. 2022. How well do LSTM language models learn filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics* 5:76–88.

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

Perfors, Amy, Joshua Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118:306–338.

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 University Press.

Piantadosi, Steven T. 2023. Modern language models refute Chomsky's approach to language. Ms., available at https://ling.auf.net/lingbuzz/007180, March 2023.

Postal, Paul M. 1993. Parasitic gaps and the across-the-board phenomenon. *Linguistic Inquiry* 24:735–754.

Pullum, Geoffrey, and Barbara Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19:9–50.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1:9.

Rawski, Jon, and J Baumont. 2023. Modern language models refute nothing. *Lingbuzz Preprint* .

Rawski, Jonathan, and Jeffrey Heinz. 2019. No free lunch in linguistics or machine learning: Response to pater. *Language* 95:e125–e135.

Reali, Florencia, and Morten Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science* 29:1007–1028.

Ross, John R. 1967. Constraints on variables in syntax. Doctoral Dissertation, MIT, Cambridge, MA.

Siegelmann, H. T., and E. D. Sontag. 1995. On the computational power of neural nets. *Journal of Computer and System Sciences* 50:132–150.

Siegelmann, Hava T., and Eduardo D. Sontag. 1991. Turing computability with neural nets. *Applied Mathematics Letters* 4:77–80.

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

Vázquez Martínez, Héctor Javier, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. *arXiv preprint arXiv:2310.20093* .

Wagers, M.W., E.F. Lau, and C. Phillips. 2009. Agreement attraction in comprehension: representations and processes. *Journal of Memory and Language* 61:206–237.

Wang, Ben, and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Warstadt, Alex, and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural*

language, 17–60. CRC Press.
Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8:377–392.
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 (Volume 2: Short Papers)*, 740–745. Melbourne, Australia: Association for Computational Linguistics.
Wilcox, Ethan, Roger Levy, and Richard Futrell. 2019. What syntactic structures block dependencies in rnn language models? *arXiv preprint arXiv:1905.10431* .
Wilcox, Ethan, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies?gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221.
Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry* 1–44.
Williams, Edwin. 1990. The ATB-theory of parasitic gaps. *The Linguistic Review* 6:265–279.
Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:945–982.
Yedetore, Aditya, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech.

A Appendix: context-free grammars

A.1 PG

$\langle S \rangle \rightarrow \langle S\_FG \rangle$
$\langle S\_FG \rangle \rightarrow \langle PREAMBLE \rangle\ \langle F \rangle\ \langle G \rangle$
$\langle S\_XG \rangle \rightarrow \langle UNGRAMMATICAL \rangle\ \langle PREAMBLE \rangle\ \langle XF \rangle\ \langle G \rangle$
$\langle S\_FX \rangle \rightarrow \langle UNGRAMMATICAL \rangle\ \langle PREAMBLE \rangle\ \langle F \rangle\ \langle XG \rangle$
$\langle S\_XX \rangle \rightarrow \langle PREAMBLE \rangle\ \langle XF \rangle\ \langle XG \rangle$

$\langle UNGRAMMATICAL \rangle$ $\rightarrow$ '*'
$\langle GEN \rangle$ $\rightarrow$ ''s'
$\langle OBJ \rangle$ $\rightarrow$ 'you' | 'us' | 'Kim'
$\langle NAME1 \rangle$ $\rightarrow$ 'Bob' | 'John'
$\langle NAME2 \rangle$ $\rightarrow$ 'Mary' | 'Jennifer'
$\langle NAME3 \rangle$ $\rightarrow$ 'James' | 'Michael'
$\langle NAME4 \rangle$ $\rightarrow$ 'Patricia' | 'Linda'
$\langle PREAMBLE \rangle$ $\rightarrow$ 'I know'
$\langle F \rangle$ $\rightarrow$ 'who' $\langle NAME1 \rangle$ $\langle GEN \rangle$ $\langle NP \rangle$
$\langle XF \rangle$ $\rightarrow$ 'that' $\langle NAME1 \rangle$ $\langle GEN \rangle$ $\langle NP \rangle$ $\langle NAME2 \rangle$
$\langle G \rangle$ $\rightarrow$ $\langle CONN \rangle$ $\langle V \rangle$ $\langle ADV2 \rangle$
$\langle XG \rangle$ $\rightarrow$ $\langle CONN \rangle$ $\langle V \rangle$ $\langle OBJ \rangle$ $\langle ADV2 \rangle$
$\langle NP \rangle$ $\rightarrow$ $\langle NP\_SIMPLE \rangle$ | $\langle NP\_COMPLEX \rangle$
$\langle NP\_SIMPLE \rangle$ $\rightarrow$ $\langle GERUND \rangle$
$\langle NP\_COMPLEX \rangle$ $\rightarrow$ $\langle N\_EMBEDDED \rangle$ 'to' $\langle V\_EMBEDDED \rangle$
$\langle CONN \rangle$ $\rightarrow$ 'is about to' | 'is likely to' | 'is going to' | 'is expected to'
$\langle V \rangle$ $\rightarrow$ 'bother' | 'annoy' | 'disturb'
$\langle GERUND \rangle$ $\rightarrow$ 'talking to' | 'dancing with' | 'playing with'
$\langle N\_EMBEDDED \rangle$ $\rightarrow$ 'decision' | 'intent' | 'effort' | 'attempt' | 'failure'
$\langle V\_EMBEDDED \rangle$ $\rightarrow$ 'talk to' | 'call' | 'meet' | 'dance with' | 'play with'
$\langle ADV1 \rangle$ $\rightarrow$ 'recently' | 'earlier'
$\langle ADV2 \rangle$ $\rightarrow$ 'soon' | 'eventually'

A.2 ATB

$\langle S \rangle$ $\rightarrow$ $\langle S\_FG \rangle$
$\langle S\_FG \rangle$ $\rightarrow$ $\langle PREAMBLE \rangle$ $\langle F \rangle$ $\langle G \rangle$
$\langle S\_XG \rangle$ $\rightarrow$ $\langle UNGRAMMATICAL \rangle$ $\langle PREAMBLE \rangle$ $\langle XF \rangle$ $\langle G \rangle$
$\langle S\_FX \rangle$ $\rightarrow$ $\langle UNGRAMMATICAL \rangle$ $\langle PREAMBLE \rangle$ $\langle F \rangle$ $\langle XG \rangle$
$\langle S\_XX \rangle$ $\rightarrow$ $\langle PREAMBLE \rangle$ $\langle XF \rangle$ $\langle XG \rangle$
$\langle UNGRAMMATICAL \rangle$ $\rightarrow$ '*'
$\langle GEN \rangle$ $\rightarrow$ ''s'
$\langle OBJ \rangle$ $\rightarrow$ 'you' | 'us' | 'Kim'
$\langle NAME1 \rangle$ $\rightarrow$ 'John'
$\langle NAME2 \rangle$ $\rightarrow$ 'Mary'
$\langle NAME3 \rangle$ $\rightarrow$ 'Bob'
$\langle NAME4 \rangle$ $\rightarrow$ 'Patricia'

$\langle PREAMBLE \rangle$ → 'I know'
$\langle F \rangle$ → 'who' $\langle NAME1 \rangle$ $\langle VP1 \rangle$ $\langle ADV1 \rangle$
$\langle XF \rangle$ → 'that' $\langle NAME1 \rangle$ $\langle VP1 \rangle$ $\langle NAME3 \rangle$ $\langle ADV1 \rangle$
$\langle G \rangle$ → $\langle CONN \rangle$ $\langle VP2 \rangle$ $\langle ADV2 \rangle$
$\langle XG \rangle$ → $\langle CONN \rangle$ $\langle VP2 \rangle$ $\langle OBJ \rangle$ $\langle ADV2 \rangle$
$\langle ADV1 \rangle$ → 'recently' | 'lately'
$\langle ADV2 \rangle$ → 'soon' | 'today' | 'now'
$\langle CONN \rangle$ → 'and is going to'
$\langle VP1 \rangle$ → $\langle VP1\_SIMPLE \rangle$ | $\langle VP1\_COMPLEX \rangle$
$\langle VP1\_SIMPLE \rangle$ → 'met' | 'saw'
$\langle VP1\_COMPLEX \rangle$ → $\langle VP1\_ABOUT \rangle$ | $\langle VP1\_TO \rangle$ | $\langle VP1\_ADJ \rangle$ | $\langle VP1\_EMBEDDED \rangle$
$\langle VP1\_ABOUT \rangle$ → $\langle V\_ABOUT\_PAST \rangle$ $\langle NAME2 \rangle$ 'about'
$\langle V\_ABOUT\_PAST \rangle$ → 'asked' | 'told'
$\langle VP1\_TO \rangle$ → $\langle V\_TO\_PAST \rangle$ $\langle NAME2 \rangle$ 'to' $\langle V\_TRANS\_INF\_TO \rangle$
$\langle V\_TO\_PAST \rangle$ → 'wanted' | 'asked'
$\langle V\_TRANS\_INF\_TO \rangle$ → 'call' | 'invite'
$\langle VP1\_ADJ \rangle$ → 'was' $\langle ADJ1 \rangle$ 'to' $\langle V\_TRANS\_INF\_ADJ \rangle$
$\langle ADJ1 \rangle$ → 'eager' | 'happy'
$\langle V\_TRANS\_INF\_ADJ \rangle$ → 'meet' | 'see'
$\langle VP1\_EMBEDDED \rangle$ → $\langle V\_EMBEDDING\_PAST \rangle$ 'that' $\langle NAME2 \rangle$ $\langle V\_TRANS\_PAST\_EM$
$\langle V\_EMBEDDING\_PAST \rangle$ → 'said' | 'insisted'
$\langle V\_TRANS\_PAST\_EMBEDDED \rangle$ → 'met' | 'saw'
$\langle VP2 \rangle$ → $\langle VP2\_COMPLEX \rangle$ | $\langle VP2\_SIMPLE \rangle$
$\langle VP2\_SIMPLE \rangle$ → 'hug' | 'slap' | 'kiss'
$\langle VP2\_COMPLEX \rangle$ → $\langle VP2\_ABOUT \rangle$ | $\langle VP2\_TO \rangle$ | $\langle VP2\_ADJ \rangle$ | $\langle VP2\_EMBEDDED \rangle$
$\langle VP2\_ABOUT \rangle$ → $\langle V\_ABOUT\_FUTURE \rangle$ 'to' $\langle NAME4 \rangle$ 'about'
$\langle V\_ABOUT\_FUTURE \rangle$ → 'complain' | 'write'
$\langle VP2\_TO \rangle$ → $\langle V\_TO\_FUTURE \rangle$ $\langle NAME4 \rangle$ 'to' $\langle V\_TRANS\_INF\_TO\_FUTURE \rangle$
$\langle V\_TO\_FUTURE \rangle$ → 'encourage' | 'beg'
$\langle V\_TRANS\_INF\_TO\_FUTURE \rangle$ → 'hug' | 'slap' | 'kiss'
$\langle VP2\_ADJ \rangle$ → 'be' $\langle ADJ2 \rangle$ 'to' $\langle V\_TRANS\_INF\_ADJ2 \rangle$
$\langle ADJ2 \rangle$ → 'afraid' | 'glad'
$\langle V\_TRANS\_INF\_ADJ2 \rangle$ → 'hug' | 'slap' | 'kiss'
$\langle VP2\_EMBEDDED \rangle$ → $\langle V\_EMBEDDING\_FUTURE \rangle$ 'that' $\langle NAME4 \rangle$ 'will' $\langle V\_TRANS\_$
$\langle V\_EMBEDDING\_FUTURE \rangle$ → 'claim' | 'predict'
$\langle V\_TRANS\_FUTURE \rangle$ → 'hug' | 'slap' | 'kiss'

B Appendix: model failures

Worst 5 four-tuples of sentences per phenomenon (PG, ATB), per model (CHILDES LSTM/Transformer, Wikipedia LSTM/Transformer, GPT-2, GPT-j, GPT-3). Surprisal values are given in parentheses at the relevant position.

B.1 PG – CHILDES LSTM

(9)

	$+gap$	$-gap$
$+filler$	I know who Bob's effort to call is going to bother **eventually (14.75)**	*I know who Bob's effort to call is going to bother **you (0.64)** eventually
$-filler$	*I know that Bob's effort to call Mary is going to bother **eventually (12.74)**	I know that Bob's effort to call Mary is going to bother **you (1.70)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -3.06$		

(10)

	$+gap$	$-gap$
$+filler$	I know who Bob's effort to call is going to annoy **eventually (15.82)**	*I know who Bob's effort to call is going to annoy **you (0.39)** eventually
$-filler$	*I know that Bob's effort to call Mary is going to annoy **eventually (13.65)**	I know that Bob's effort to call Mary is going to annoy **you (1.23)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -3.01$		

(11)

	$+gap$	$-gap$
$+filler$	I know who Bob's effort to call is going to bother **soon (13.28)**	*I know who Bob's effort to call is going to bother **you (0.64)** soon
$-filler$	*I know that Bob's effort to call Mary is going to bother **soon (11.53)**	I know that Bob's effort to call Mary is going to bother **you (1.70)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.80$		

(12)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John's dancing with is going to disturb **eventually (15.03)**	*I know <u>who</u> John's dancing with is going to disturb **us (1.82)** eventually
$-filler$	*I know <u>that</u> John's dancing with Mary is going to disturb **eventually (13.69)**	I know <u>that</u> John's dancing with Mary is going to disturb **us (3.27)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -2.80$		

(13)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Bob's failure to meet is going to disturb **eventually (13.66)**	*I know <u>who</u> Bob's failure to meet is going to disturb **us (1.63)** eventually
$-filler$	*I know <u>that</u> Bob's failure to meet Mary is going to disturb **eventually (12.41)**	I know <u>that</u> Bob's failure to meet Mary is going to disturb **us (3.17)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -2.79$		

B.2 PG – CHILDES Transformer

(14)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Bob's attempt to play with is going to bother **soon (12.45)**	*I know <u>who</u> Bob's attempt to play with is going to bother **us (1.55)** soon
$-filler$	*I know <u>that</u> Bob's attempt to play with Jennifer is going to bother **soon (10.69)**	I know <u>that</u> Bob's attempt to play with Jennifer is going to bother **us (3.20)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.41$		

(15)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Bob's effort to play with is going to bother **soon (12.27)**	*I know <u>who</u> Bob's effort to play with is going to bother **us (1.02)** soon
$-filler$	*I know <u>that</u> Bob's effort to play with Jennifer is going to bother **soon (10.31)**	I know <u>that</u> Bob's effort to play with Jennifer is going to bother **us (2.32)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.27$		

(16)

	$+gap$	$-gap$
$+filler$	I know who Bob's attempt to play with is about to bother **soon (12.78)**	*I know who Bob's attempt to play with is about to bother **us (1.63)** soon
$-filler$	*I know that Bob's attempt to play with Jennifer is about to bother **soon (10.91)**	I know that Bob's attempt to play with Jennifer is about to bother **us (2.89)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.13$		

(17)

	$+gap$	$-gap$
$+filler$	I know who Bob's attempt to play with is expected to bother **soon (13.39)**	*I know who Bob's attempt to play with is expected to bother **us (2.02)** soon
$-filler$	*I know that Bob's attempt to play with Jennifer is expected to bother **soon (11.92)**	I know that Bob's attempt to play with Jennifer is expected to bother **us (3.66)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.10$		

(18)

	$+gap$	$-gap$
$+filler$	I know who Bob's failure to play with is going to bother **soon (12.32)**	*I know who Bob's failure to play with is going to bother **us (1.80)** soon
$-filler$	*I know that Bob's failure to play with Jennifer is going to bother **soon (10.63)**	I know that Bob's failure to play with Jennifer is going to bother **us (3.14)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.03$		

B.3 PG – Wikipedia LSTM

(19)

	$+gap$	$-gap$
$+filler$	I know who John's intent to call is going to bother **eventually (16.59)**	*I know who John's intent to call is going to bother **us (4.90)** eventually
$-filler$	*I know that John's intent to call Jennifer is going to bother **eventually (13.47)**	I know that John's intent to call Jennifer is going to bother **us (7.20)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -5.42$		

(20)

	$+gap$	$-gap$
$+filler$	I know who John's failure to call is going to bother **eventually (16.07)**	*I know who John's failure to call is going to bother **us (4.93)** eventually
$-filler$	*I know that John's failure to call Jennifer is going to bother **eventually (13.77)**	I know that John's failure to call Jennifer is going to bother **us (7.89)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -5.26$		

(21)

	$+gap$	$-gap$
$+filler$	I know who John's failure to call is going to bother **soon (15.47)**	*I know who John's failure to call is going to bother **us (4.93)** soon
$-filler$	*I know that John's failure to call Jennifer is going to bother **soon (13.68)**	I know that John's failure to call Jennifer is going to bother **us (7.89)** soon
$\Delta_{-filler} - \Delta_{+filler} = -4.74$		

(22)

	$+gap$	$-gap$
$+filler$	I know who John's attempt to talk to is going to bother **eventually (15.99)**	*I know who John's attempt to talk to is going to bother **us (4.42)** eventually
$-filler$	*I know that John's attempt to talk to Jennifer is going to bother **eventually (14.78)**	I know that John's attempt to talk to Jennifer is going to bother **us (7.65)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -4.44$		

(23)

	$+gap$	$-gap$
$+filler$	I know who John's intent to talk to is going to bother **eventually (16.62)**	*I know who John's intent to talk to is going to bother **you (5.39)** eventually
$-filler$	*I know that John's intent to talk to Jennifer is going to bother **eventually (14.89)**	I know that John's intent to talk to Jennifer is going to bother **you (8.06)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -4.41$		

B.4 PG – Wikipedia Transformer

		$+gap$	$-gap$
(24)	$+filler$	I know <u>who</u> Bob's attempt to talk to is about to bother **soon (11.09)**	*I know <u>who</u> Bob's attempt to talk to is about to bother **you (3.73)** soon
	$-filler$	*I know <u>that</u> Bob's attempt to talk to Jennifer is about to bother **soon (7.94)**	I know <u>that</u> Bob's attempt to talk to Jennifer is about to bother **you (10.53)** soon
	$\Delta_{-filler} - \Delta_{+filler} = -9.95$		

		$+gap$	$-gap$
(25)	$+filler$	I know <u>who</u> Bob's attempt to talk to is expected to bother **soon (10.01)**	*I know <u>who</u> Bob's attempt to talk to is expected to bother **you (3.49)** soon
	$-filler$	*I know <u>that</u> Bob's attempt to talk to Jennifer is expected to bother **soon (7.23)**	I know <u>that</u> Bob's attempt to talk to Jennifer is expected to bother **you (10.49)** soon
	$\Delta_{-filler} - \Delta_{+filler} = -9.78$		

		$+gap$	$-gap$
(26)	$+filler$	I know <u>who</u> Bob's attempt to talk to is expected to bother **eventually (10.76)**	*I know <u>who</u> Bob's attempt to talk to is expected to bother **you (3.49)** eventually
	$-filler$	*I know <u>that</u> Bob's attempt to talk to Jennifer is expected to bother **eventually (8.07)**	I know <u>that</u> Bob's attempt to talk to Jennifer is expected to bother **you (10.49)** eventually
	$\Delta_{-filler} - \Delta_{+filler} = -9.70$		

(27)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John's intent to talk to is expected to bother **eventually (10.98)**	*I know <u>who</u> John's intent to talk to is expected to bother **us (4.03)** eventually
$-filler$	*I know <u>that</u> John's intent to talk to Jennifer is expected to bother **eventually (8.32)**	I know <u>that</u> John's intent to talk to Jennifer is expected to bother **us (11.06)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -9.69$		

(28)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Bob's decision to talk to is about to bother **soon (10.66)**	*I know <u>who</u> Bob's decision to talk to is about to bother **you (3.70)** soon
$-filler$	*I know <u>that</u> Bob's decision to talk to Jennifer is about to bother **soon (7.83)**	I know <u>that</u> Bob's decision to talk to Jennifer is about to bother **you (10.50)** soon
$\Delta_{-filler} - \Delta_{+filler} = -9.63$		

B.5 PG – GPT-2

(29)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Bob's talking to is going to annoy **eventually (16.51)**	*I know <u>who</u> Bob's talking to is going to annoy **you (2.77)** eventually
$-filler$	*I know <u>that</u> Bob's talking to Jennifer is going to annoy **eventually (16.17)**	I know <u>that</u> Bob's talking to Jennifer is going to annoy **you (4.70)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -2.26$		

(30)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Bob's decision to meet is about to bother **eventually (18.76)**	*I know <u>who</u> Bob's decision to meet is about to bother **you (2.17)** eventually
$-filler$	*I know <u>that</u> Bob's decision to meet Jennifer is about to bother **eventually (18.42)**	I know <u>that</u> Bob's decision to meet Jennifer is about to bother **you (4.04)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -2.20$		

(31)

	$+gap$	$-gap$
$+filler$	I know who John's decision to call is likely to disturb **soon (14.28)**	*I know who John's decision to call is likely to disturb **you (2.73)** soon
$-filler$	*I know that John's decision to call Mary is likely to disturb **soon (15.19)**	I know that John's decision to call Mary is likely to disturb **you (5.84)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.19$		

(32)

	$+gap$	$-gap$
$+filler$	I know who John's talking to is likely to disturb **eventually (16.82)**	*I know who John's talking to is likely to disturb **you (1.81)** eventually
$-filler$	*I know that John's talking to Jennifer is likely to disturb **eventually (16.53)**	I know that John's talking to Jennifer is likely to disturb **you (3.68)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -2.16$		

(33)

	$+gap$	$-gap$
$+filler$	I know who Bob's decision to call is likely to disturb **soon (14.31)**	*I know who Bob's decision to call is likely to disturb **you (2.70)** soon
$-filler$	*I know that Bob's decision to call Mary is likely to disturb **soon (15.37)**	I know that Bob's decision to call Mary is likely to disturb **you (5.91)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.15$		

B.6 PG – GPT-j

(34)

	$+gap$	$-gap$
$+filler$	I know who John's attempt to talk to is about to annoy **soon (11.48)**	*I know who John's attempt to talk to is about to annoy **you (1.62)** soon
$-filler$	*I know that John's attempt to talk to Mary is about to annoy **soon (12.86)**	I know that John's attempt to talk to Mary is about to annoy **you (4.70)** soon
$\Delta_{-filler} - \Delta_{+filler} = -1.70$		

(35)

	$+gap$	$-gap$
$+filler$	I know who John's failure to dance with is about to annoy **eventually (13.09)**	*I know who John's failure to dance with is about to annoy **you (1.53)** eventually
$-filler$	*I know that John's failure to dance with Mary is about to annoy **eventually (13.60)**	I know that John's failure to dance with Mary is about to annoy **you (3.69)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -1.66$		

(36)

	$+gap$	$-gap$
$+filler$	I know who John's attempt to talk to is about to annoy **eventually (12.23)**	*I know who John's attempt to talk to is about to annoy **you (1.62)** eventually
$-filler$	*I know that John's attempt to talk to Mary is about to annoy **eventually (13.71)**	I know that John's attempt to talk to Mary is about to annoy **you (4.70)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -1.59$		

(37)

	$+gap$	$-gap$
$+filler$	I know who John's decision to dance with is about to annoy **eventually (13.70)**	*I know who John's decision to dance with is about to annoy **you (1.26)** eventually
$-filler$	*I know that John's decision to dance with Mary is about to annoy **eventually (13.85)**	I know that John's decision to dance with Mary is about to annoy **you (2.95)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -1.54$		

(38)

	$+gap$	$-gap$
$+filler$	I know who John's decision to dance with is about to annoy **soon (12.74)**	*I know who John's decision to dance with is about to annoy **you (1.26)** soon
$-filler$	*I know that John's decision to dance with Mary is about to annoy **soon (12.92)**	I know that John's decision to dance with Mary is about to annoy **you (2.95)** soon
$\Delta_{-filler} - \Delta_{+filler} = -1.51$		

B.7 PG – GPT-3

(39)

	$+gap$	$-gap$
$+filler$	I know who Bob's talking to is likely to annoy **soon (14.53)**	*I know who Bob's talking to is likely to annoy **you (3.39)** soon
$-filler$	*I know that Bob's talking to Mary is likely to annoy **soon (13.53)**	I know that Bob's talking to Mary is likely to annoy **you (7.20)** soon
$\Delta_{-filler} - \Delta_{+filler} = -4.81$		

(40)

	$+gap$	$-gap$
$+filler$	I know who John's talking to is likely to annoy **soon (14.06)**	*I know who John's talking to is likely to annoy **you (3.55)** soon
$-filler$	*I know that John's talking to Mary is likely to annoy **soon (13.59)**	I know that John's talking to Mary is likely to annoy **you (7.60)** soon
$\Delta_{-filler} - \Delta_{+filler} = -4.52$		

(41)

	$+gap$	$-gap$
$+filler$	I know who Bob's playing with is likely to annoy **soon (13.36)**	*I know who Bob's playing with is likely to annoy **you (2.75)** soon
$-filler$	*I know that Bob's playing with Mary is likely to annoy **soon (13.46)**	I know that Bob's playing with Mary is likely to annoy **you (6.94)** soon
$\Delta_{-filler} - \Delta_{+filler} = -4.09$		

(42)

	$+gap$	$-gap$
$+filler$	I know who John's talking to is expected to annoy **soon (12.97)**	*I know who John's talking to is expected to annoy **you (3.05)** soon
$-filler$	*I know that John's talking to Mary is expected to annoy **soon (12.61)**	I know that John's talking to Mary is expected to annoy **you (6.56)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.87$		

(43)

	$+gap$	$-gap$
$+filler$	I know who Bob's talking to is likely to annoy **eventually (15.21)**	*I know who Bob's talking to is likely to annoy **you (3.39)** eventually
$-filler$	*I know that Bob's talking to Mary is likely to annoy **eventually (15.18)**	I know that Bob's talking to Mary is likely to annoy **you (7.20)** eventually
$\Delta_{-filler} - \Delta_{+filler} = -3.84$		

B.8 ATB – CHILDES LSTM

(44)

	$+gap$	$-gap$
$+filler$	I know who John told Mary about lately and is going to encourage Patricia to slap **now (9.84)**	*I know who John told Mary about lately and is going to encourage Patricia to slap **you (1.88)** now
$-filler$	*I know that John told Mary about Bob lately and is going to encourage Patricia to slap **now (9.24)**	I know that John told Mary about Bob lately and is going to encourage Patricia to slap **you (3.35)** now
$\Delta_{-filler} - \Delta_{+filler} = -2.08$		

(45)

	$+gap$	$-gap$
$+filler$	I know who John insisted that Mary met recently and is going to be afraid to kiss **now (8.53)**	*I know who John insisted that Mary met recently and is going to be afraid to kiss **Kim (10.20)** now
$-filler$	*I know that John insisted that Mary met Bob recently and is going to be afraid to kiss **now (8.46)**	I know that John insisted that Mary met Bob recently and is going to be afraid to kiss **Kim (12.17)** now
$\Delta_{-filler} - \Delta_{+filler} = -2.04$		

(46)

	$+gap$	$-gap$
$+filler$	I know who John insisted that Mary met recently and is going to be afraid to kiss **today (8.56)**	*I know who John insisted that Mary met recently and is going to be afraid to kiss **Kim (10.20)** today
$-filler$	*I know that John insisted that Mary met Bob recently and is going to be afraid to kiss **today (8.53)**	I know that John insisted that Mary met Bob recently and is going to be afraid to kiss **Kim (12.17)** today
$\Delta_{-filler} - \Delta_{+filler} = -2.00$		

(47)

	$+gap$	$-gap$
$+filler$	I know who John insisted that Mary met recently and is going to be afraid to slap **now (9.34)**	*I know who John insisted that Mary met recently and is going to be afraid to slap **Kim (9.90)** now
$-filler$	*I know that John insisted that Mary met Bob recently and is going to be afraid to slap **now (9.29)**	I know that John insisted that Mary met Bob recently and is going to be afraid to slap **Kim (11.85)** now
$\Delta_{-filler} - \Delta_{+filler} = -2.00$		

(48)

	$+gap$	$-gap$
$+filler$	I know who John wanted Mary to invite recently and is going to be afraid to slap **now (9.41)**	*I know who John wanted Mary to invite recently and is going to be afraid to slap **Kim (10.37)** now
$-filler$	*I know that John wanted Mary to invite Bob recently and is going to be afraid to slap **now (9.22)**	I know that John wanted Mary to invite Bob recently and is going to be afraid to slap **Kim (12.16)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.98$		

B.9 ATB – CHILDES Transformer

(49)

	$+gap$	$-gap$
$+filler$	I know who John asked Mary about lately and is going to be afraid to slap **now (10.21)**	*I know who John asked Mary about lately and is going to be afraid to slap **us (3.41)** now
$-filler$	*I know that John asked Mary about Bob lately and is going to be afraid to slap **now (9.42)**	I know that John asked Mary about Bob lately and is going to be afraid to slap **us (4.14)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.51$		

(50)

	$+gap$	$-gap$
$+filler$	I know who John said that Mary saw recently and is going to slap **now (10.72)**	*I know who John said that Mary saw recently and is going to slap **us (3.14)** now
$-filler$	*I know that John said that Mary saw Bob recently and is going to slap **now (10.18)**	I know that John said that Mary saw Bob recently and is going to slap **us (4.00)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.40$		

(51)

	$+gap$	$-gap$
$+filler$	I know who John said that Mary saw lately and is going to be afraid to slap **now (9.88)**	*I know who John said that Mary saw lately and is going to be afraid to slap **us (3.24)** now
$-filler$	*I know that John said that Mary saw Bob lately and is going to be afraid to slap **now (9.45)**	I know that John said that Mary saw Bob lately and is going to be afraid to slap **us (4.21)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.40$		

(52)

	$+gap$	$-gap$
$+filler$	I know who John asked Mary to call lately and is going to be afraid to slap **now (10.08)**	*I know who John asked Mary to call lately and is going to be afraid to slap **us (3.59)** now
$-filler$	*I know that John asked Mary to call Bob lately and is going to be afraid to slap **now (9.83)**	I know that John asked Mary to call Bob lately and is going to be afraid to slap **us (4.72)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.38$		

(53)

	$+gap$	$-gap$
$+filler$	I know who John said that Mary saw recently and is going to be afraid to slap **now (10.12)**	*I know who John said that Mary saw recently and is going to be afraid to slap **us (3.78)** now
$-filler$	*I know that John said that Mary saw Bob recently and is going to be afraid to slap **now (9.98)**	I know that John said that Mary saw Bob recently and is going to be afraid to slap **us (5.01)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.37$		

B.10 ATB – Wikipedia LSTM

(54)

	$+gap$	$-gap$
$+filler$	I know who John asked Mary about recently and is going to kiss **soon (14.80)**	*I know who John asked Mary about recently and is going to kiss **us (8.39)** soon
$-filler$	*I know that John asked Mary about Bob recently and is going to kiss **soon (14.40)**	I know that John asked Mary about Bob recently and is going to kiss **us (11.63)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.64$		

(55)

	$+gap$	$-gap$
$+filler$	I know who John met recently and is going to beg Patricia to kiss **soon (15.45)**	*I know who John met recently and is going to beg Patricia to kiss **us (10.27)** soon
$-filler$	*I know that John met Bob recently and is going to beg Patricia to kiss **soon (15.46)**	I know that John met Bob recently and is going to beg Patricia to kiss **us (13.61)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.33$		

(56)

	$+gap$	$-gap$
$+filler$	I know who John asked Mary about recently and is going to be glad to kiss **soon (14.47)**	*I know who John asked Mary about recently and is going to be glad to kiss **us (9.19)** soon
$-filler$	*I know that John asked Mary about Bob recently and is going to be glad to kiss **soon (14.54)**	I know that John asked Mary about Bob recently and is going to be glad to kiss **us (12.59)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.33$		

(57)

	$+gap$	$-gap$
$+filler$	I know who John was happy to meet recently and is going to kiss **soon (13.61)**	*I know who John was happy to meet recently and is going to kiss **you (7.32)** soon
$-filler$	*I know that John was happy to meet Bob recently and is going to kiss **soon (13.36)**	I know that John was happy to meet Bob recently and is going to kiss **you (10.36)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.29$		

(58)

	$+gap$	$-gap$
$+filler$	I know who John told Mary about recently and is going to beg Patricia to kiss **soon (15.49)**	*I know who John told Mary about recently and is going to beg Patricia to kiss **us (11.17)** soon
$-filler$	*I know that John told Mary about Bob recently and is going to beg Patricia to kiss **soon (15.26)**	I know that John told Mary about Bob recently and is going to beg Patricia to kiss **us (14.20)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.26$		

B.11 ATB – Wikipedia Transformer

(59)

	$+gap$	$-gap$
$+filler$	I know who John saw lately and is going to write to Patricia about **now (8.22)**	*I know who John saw lately and is going to write to Patricia about **us (7.18)** now
$-filler$	*I know that John saw Bob lately and is going to write to Patricia about **now (8.21)**	I know that John saw Bob lately and is going to write to Patricia about **us (8.38)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.21$		

(60)

	$+gap$	$-gap$
$+filler$	I know who John saw recently and is going to write to Patricia about **now (8.01)**	*I know who John saw recently and is going to write to Patricia about **us (7.28)** now
$-filler$	*I know that John saw Bob recently and is going to write to Patricia about **now (8.11)**	I know that John saw Bob recently and is going to write to Patricia about **us (8.53)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.15$		

(61)

	$+gap$	$-gap$
$+filler$	I know who John met lately and is going to write to Patricia about **now (8.26)**	*I know who John met lately and is going to write to Patricia about **us (7.40)** now
$-filler$	*I know that John met Bob lately and is going to write to Patricia about **now (8.35)**	I know that John met Bob lately and is going to write to Patricia about **us (8.58)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.09$		

(62)

	$+gap$	$-gap$
$+filler$	I know who John was happy to see lately and is going to write to Patricia about **now (8.46)**	*I know who John was happy to see lately and is going to write to Patricia about **us (7.52)** now
$-filler$	*I know that John was happy to see Bob lately and is going to write to Patricia about **now (8.45)**	I know that John was happy to see Bob lately and is going to write to Patricia about **us (8.54)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.04$		

(63)

	$+gap$	$-gap$
$+filler$	I know who John was happy to meet lately and is going to write to Patricia about **now (8.51)**	*I know who John was happy to meet lately and is going to write to Patricia about **us (7.68)** now
$-filler$	*I know that John was happy to meet Bob lately and is going to write to Patricia about **now (8.57)**	I know that John was happy to meet Bob lately and is going to write to Patricia about **us (8.72)** now
$\Delta_{-filler} - \Delta_{+filler} = -0.98$		

B.12 ATB – GPT-2

(64)

	$+gap$	$-gap$
$+filler$	I know who John said that Mary saw lately and is going to be glad to kiss **soon (14.91)**	*I know who John said that Mary saw lately and is going to be glad to kiss **you (4.06)** soon
$-filler$	*I know that John said that Mary saw Bob lately and is going to be glad to kiss **soon (16.39)**	I know that John said that Mary saw Bob lately and is going to be glad to kiss **you (7.08)** soon
$\Delta_{-filler} - \Delta_{+filler} = -1.53$		

(65)

	$+gap$	$-gap$
$+filler$	I know who John said that Mary met lately and is going to be glad to kiss **soon (14.46)**	*I know who John said that Mary met lately and is going to be glad to kiss **you (4.11)** soon
$-filler$	*I know that John said that Mary met Bob lately and is going to be glad to kiss **soon (15.85)**	I know that John said that Mary met Bob lately and is going to be glad to kiss **you (6.93)** soon
$\Delta_{-filler} - \Delta_{+filler} = -1.43$		

(66)

	$+gap$	$-gap$
$+filler$	I know who John said that Mary met lately and is going to be glad to kiss **now (12.15)**	*I know who John said that Mary met lately and is going to be glad to kiss **you (4.11)** now
$-filler$	*I know that John said that Mary met Bob lately and is going to be glad to kiss **now (13.67)**	I know that John said that Mary met Bob lately and is going to be glad to kiss **you (6.93)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.29$		

(67)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John said that Mary met lately and is going to kiss **soon (13.26)**	*I know <u>who</u> John said that Mary met lately and is going to kiss **you (5.31)** soon
$-filler$	*I know <u>that</u> John said that Mary met Bob lately and is going to kiss **soon (15.08)**	I know <u>that</u> John said that Mary met Bob lately and is going to kiss **you (8.30)** soon
$\Delta_{-filler} - \Delta_{+filler} = -1.18$		

(68)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John was eager to see lately and is going to predict that Patricia will kiss **soon (11.50)**	*I know <u>who</u> John was eager to see lately and is going to predict that Patricia will kiss **Kim (10.84)** soon
$-filler$	*I know <u>that</u> John was eager to see Bob lately and is going to predict that Patricia will kiss **soon (11.45)**	I know <u>that</u> John was eager to see Bob lately and is going to predict that Patricia will kiss **Kim (11.95)** soon
$\Delta_{-filler} - \Delta_{+filler} = -1.16$		

B.13 ATB – GPT-j

(69)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John told Mary about lately and is going to hug **now (8.12)**	*I know <u>who</u> John told Mary about lately and is going to hug **us (6.59)** now
$-filler$	*I know <u>that</u> John told Mary about Bob lately and is going to hug **now (8.88)**	I know <u>that</u> John told Mary about Bob lately and is going to hug **us (9.42)** now
$\Delta_{-filler} - \Delta_{+filler} = -2.07$		

(70)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John told Mary about lately and is going to hug **today (8.69)**	*I know <u>who</u> John told Mary about lately and is going to hug **us (6.59)** today
$-filler$	*I know <u>that</u> John told Mary about Bob lately and is going to hug **today (9.54)**	I know <u>that</u> John told Mary about Bob lately and is going to hug **us (9.42)** today
$\Delta_{-filler} - \Delta_{+filler} = -1.98$		

(71)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John said that Mary met recently and is going to be afraid to hug **now (7.54)**	*I know <u>who</u> John said that Mary met recently and is going to be afraid to hug **us (4.59)** now
$-filler$	*I know <u>that</u> John said that Mary met Bob recently and is going to be afraid to hug **now (9.06)**	I know <u>that</u> John said that Mary met Bob recently and is going to be afraid to hug **us (8.07)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.96$		

(72)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John said that Mary met recently and is going to be afraid to slap **now (10.08)**	*I know <u>who</u> John said that Mary met recently and is going to be afraid to slap **us (4.22)** now
$-filler$	*I know <u>that</u> John said that Mary met Bob recently and is going to be afraid to slap **now (10.79)**	I know <u>that</u> John said that Mary met Bob recently and is going to be afraid to slap **us (6.81)** now
$\Delta_{-filler} - \Delta_{+filler} = -1.88$		

(73)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John told Mary about recently and is going to be afraid to hug **today (10.32)**	*I know <u>who</u> John told Mary about recently and is going to be afraid to hug **us (6.16)** today
$-filler$	*I know <u>that</u> John told Mary about Bob recently and is going to be afraid to hug **today (11.78)**	I know <u>that</u> John told Mary about Bob recently and is going to be afraid to hug **us (9.46)** today
$\Delta_{-filler} - \Delta_{+filler} = -1.84$		

B.14 ATB – GPT-3

(74)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John told Mary about lately and is going to be afraid to hug **now (8.21)**	*I know <u>who</u> John told Mary about lately and is going to be afraid to hug **you (5.06)** now
$-filler$	*I know <u>that</u> John told Mary about Bob lately and is going to be afraid to hug **now (8.39)**	I know <u>that</u> John told Mary about Bob lately and is going to be afraid to hug **you (8.09)** now
$\Delta_{-filler} - \Delta_{+filler} = -2.85$		

(75)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> John told Mary about lately and is going to slap **now (9.43)**	*I know <u>who</u> John told Mary about lately and is going to slap **you (4.70)** now
$-filler$	*I know <u>that</u> John told Mary about Bob lately and is going to slap **now (10.25)**	I know <u>that</u> John told Mary about Bob lately and is going to slap **you (8.27)** now
$\Delta_{-filler} - \Delta_{+filler} = -2.75$		

(76)

	$+gap$	$-gap$
$+filler$	I know who John told Mary about lately and is going to slap **soon (7.22)**	*I know who John told Mary about lately and is going to slap **you (4.70)** soon
$-filler$	*I know that John told Mary about Bob lately and is going to slap **soon (8.33)**	I know that John told Mary about Bob lately and is going to slap **you (8.27)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.47$		

(77)

	$+gap$	$-gap$
$+filler$	I know who John told Mary about lately and is going to be glad to slap **soon (9.32)**	*I know who John told Mary about lately and is going to be glad to slap **you (3.35)** soon
$-filler$	*I know that John told Mary about Bob lately and is going to be glad to slap **soon (9.52)**	I know that John told Mary about Bob lately and is going to be glad to slap **you (5.82)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.27$		

(78)

	$+gap$	$-gap$
$+filler$	I know who John told Mary about lately and is going to be glad to slap **today (10.41)**	*I know who John told Mary about lately and is going to be glad to slap **you (3.35)** today
$-filler$	*I know that John told Mary about Bob lately and is going to be glad to slap **today (10.64)**	I know that John told Mary about Bob lately and is going to be glad to slap **you (5.82)** today
$\Delta_{-filler} - \Delta_{+filler} = -2.24$		

C Appendix: context-free grammars for retraining task (Section 5)

C.1 PG – retraining corpus

$\langle S \rangle \rightarrow \langle S\_FG \rangle$
$\langle S\_FG \rangle \rightarrow \langle PREAMBLE \rangle\ \langle F \rangle\ \langle G \rangle$
$\langle S\_XG \rangle \rightarrow \langle UNGRAMMATICAL \rangle\ \langle PREAMBLE \rangle\ \langle XF \rangle\ \langle G \rangle$
$\langle S\_FX \rangle \rightarrow \langle UNGRAMMATICAL \rangle\ \langle PREAMBLE \rangle\ \langle F \rangle\ \langle XG \rangle$
$\langle S\_XX \rangle \rightarrow \langle PREAMBLE \rangle\ \langle XF \rangle\ \langle XG \rangle$
$\langle GEN \rangle \rightarrow$ ‘’s’
$\langle OBJ \rangle \rightarrow$ ‘you’ | ‘us’ | ‘Kim’
$\langle NAME1 \rangle \rightarrow$ ‘Bob’ | ‘John’
$\langle NAME2 \rangle \rightarrow$ ‘Mary’ | ‘Jennifer’
$\langle NAME3 \rangle \rightarrow$ ‘James’ | ‘Michael’
$\langle NAME4 \rangle \rightarrow$ ‘Patricia’ | ‘Linda’
$\langle PREAMBLE \rangle \rightarrow$ ‘I know’
$\langle F \rangle \rightarrow$ ‘who’ $\langle NAME1 \rangle\ \langle GEN \rangle\ \langle SUBJ \rangle$
$\langle XF \rangle \rightarrow$ ‘that’ $\langle NAME1 \rangle\ \langle GEN \rangle\ \langle SUBJ \rangle\ \langle NAME2 \rangle$
$\langle G \rangle \rightarrow \langle G\_PAST \rangle \mid \langle G\_FUTURE \rangle$
$\langle XG \rangle \rightarrow \langle XG\_PAST \rangle \mid \langle XG\_FUTURE \rangle$
$\langle G\_PAST \rangle \rightarrow \langle VP\_PAST \rangle\ \langle ADV\_PAST \rangle$
$\langle G\_FUTURE \rangle \rightarrow \langle VP\_FUTURE \rangle\ \langle ADV\_FUTURE \rangle$
$\langle XG\_PAST \rangle \rightarrow \langle VP\_PAST \rangle\ \langle XG\_OBJ \rangle\ \langle ADV\_PAST \rangle$
$\langle XG\_FUTURE \rangle \rightarrow \langle VP\_FUTURE \rangle\ \langle XG\_OBJ \rangle\ \langle ADV\_FUTURE \rangle$
$\langle VP\_PAST \rangle \rightarrow$ ‘upset’ | ‘distracted’ | ‘worried’ | ‘annoyed’ | ‘amused’ | ‘delighted’
$\langle VP\_FUTURE \rangle \rightarrow$ ‘will’ $\langle V\_FUTURE \rangle$
$\langle V\_FUTURE \rangle \rightarrow$ ‘upset’ | ‘distract’ | ‘worry’ | ‘annoy’ | ‘amuse’ | ‘delight’
$\langle XG\_OBJ \rangle \rightarrow \langle NAME4 \rangle \mid \langle OBJ \rangle$
$\langle SUBJ \rangle \rightarrow$ ‘attitude towards’ | ‘friendship with’ | ‘praising of’ | ‘fight with’ | ‘kissing with’ | ‘asking abo
$\langle ADV\_PAST \rangle \rightarrow$ ‘yesterday’ | ‘recently’ | ‘often’ | ‘constantly’ | ‘today’ | ‘lately’ | ‘earlier’
$\langle ADV\_FUTURE \rangle \rightarrow$ ‘today’ | ‘soon’ | ‘tomorrow’ | ‘now’ | ‘quickly’

C.2 ATB – retraining corpus

$\langle S \rangle \rightarrow \langle S\_FG \rangle$

$\langle S\_FG \rangle \rightarrow \langle PREAMBLE \rangle\ \langle F \rangle\ \langle G \rangle$

$\langle S\_XG \rangle \rightarrow \langle UNGRAMMATICAL \rangle\ \langle PREAMBLE \rangle\ \langle XF \rangle\ \langle G \rangle$

$\langle S\_FX \rangle \rightarrow \langle UNGRAMMATICAL \rangle\ \langle PREAMBLE \rangle\ \langle F \rangle\ \langle XG \rangle$

$\langle S\_XX \rangle \rightarrow \langle PREAMBLE \rangle\ \langle XF \rangle\ \langle XG \rangle$

$\langle GEN \rangle \rightarrow$ ‘’s’

$\langle OBJ \rangle \rightarrow$ ‘you’ | ‘us’ | ‘Kim’

$\langle NAME1 \rangle \rightarrow$ ‘John’

$\langle NAME2 \rangle \rightarrow$ ‘Mary’

$\langle NAME3 \rangle \rightarrow$ ‘Bob’

$\langle NAME4 \rangle \rightarrow$ ‘Patricia’

$\langle PREAMBLE \rangle \rightarrow$ ‘I know’

$\langle F \rangle \rightarrow$ ‘who’ $\langle SUBJ\_F \rangle$

$\langle XF \rangle \rightarrow$ ‘that’ $\langle SUBJ\_XF \rangle$

$\langle CONN \rangle \rightarrow$ ‘and’

$\langle SUBJ\_F \rangle \rightarrow \langle ONE\_SUBJ\_F \rangle \mid \langle TWO\_SUBJ\_F \rangle$

$\langle SUBJ\_XF \rangle \rightarrow \langle ONE\_SUBJ\_XF \rangle \mid \langle TWO\_SUBJ\_XF \rangle$

$\langle ONE\_SUBJ\_F \rangle \rightarrow \langle NAME1 \rangle\ \langle V1 \rangle\ \langle ADV\_PAST\_1 \rangle\ \langle CONN \rangle$

$\langle TWO\_SUBJ\_F \rangle \rightarrow \langle NAME1 \rangle\ \langle V1 \rangle\ \langle ADV\_PAST\_1 \rangle\ \langle CONN \rangle\ \langle NAME3 \rangle$

$\langle ONE\_SUBJ\_XF \rangle \rightarrow \langle NAME1 \rangle\ \langle V1 \rangle\ \langle NAME2 \rangle\ \langle ADV\_PAST\_1 \rangle\ \langle CONN \rangle$

$\langle TWO\_SUBJ\_XF \rangle \rightarrow \langle NAME1 \rangle\ \langle V1 \rangle\ \langle NAME2 \rangle\ \langle ADV\_PAST\_1 \rangle\ \langle CONN \rangle\ \langle NAME3 \rangle$

$\langle G \rangle \rightarrow \langle G\_PAST \rangle \mid \langle G\_FUTURE \rangle$

$\langle XG \rangle \rightarrow \langle XG\_PAST \rangle \mid \langle XG\_FUTURE \rangle$

$\langle XG\_PAST \rangle \rightarrow \langle VP2\_PAST \rangle\ \langle XG\_OBJ \rangle\ \langle ADV\_PAST\_2 \rangle$

$\langle XG\_FUTURE \rangle \rightarrow \langle VP2\_FUTURE \rangle\ \langle XG\_OBJ \rangle\ \langle ADV\_FUTURE \rangle$

$\langle G\_PAST \rangle \rightarrow \langle VP2\_PAST \rangle\ \langle ADV\_PAST\_2 \rangle$

$\langle G\_FUTURE \rangle \rightarrow \langle VP2\_FUTURE \rangle\ \langle ADV\_FUTURE \rangle$

$\langle XG\_OBJ \rangle \rightarrow \langle NAME4 \rangle \mid \langle OBJ \rangle$

$\langle V1 \rangle \rightarrow$ ‘saw’ | ‘hugged’ | ‘helped’ | ‘met’ | ‘pushed’ | ‘praised’ | ‘chased’ | ‘hired’ | ‘invited’ | ‘promoted’

$\langle VP2 \rangle \rightarrow \langle VP2\_PAST \rangle \mid \langle VP2\_FUTURE \rangle$

$\langle VP2\_PAST \rangle \rightarrow \langle V2\_PAST \rangle$

$\langle V2\_PAST \rangle \rightarrow$ ‘kissed’ | ‘slapped’ | ‘insulted’ | ‘annoyed’ | ‘hurt’ | ‘mocked’ | ‘teased’ | ‘supported’ | ‘m

$\langle VP2\_FUTURE \rangle \rightarrow$ ‘will’ $\langle V2\_FUTURE \rangle$

$\langle V2\_FUTURE \rangle \rightarrow$ ‘kiss’ | ‘slap’ | ‘insult’ | ‘annoy’ | ‘hurt’ | ‘mock’ | ‘tease’ | ‘support’ | ‘marry’

$\langle ADV\_PAST\_1 \rangle \rightarrow$ ‘yesterday’ | ‘recently’ | ‘often’ | ‘constantly’

$\langle ADV\_PAST\_2 \rangle \rightarrow$ ‘today’ | ‘lately’ | ‘earlier’ | ‘regularly’ | ‘repeatedly’

$\langle ADV\_FUTURE \rangle$ $\rightarrow$ 'today' | 'soon' | 'tomorrow' | 'now' | 'quickly'