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

Nur Lan, Emmanuel Chemla, and Roni Katzir

September 2022

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. Wilcox et al. (2021) use such models to examine the available evidence as it pertains to wh-movement. 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 and show that the networks do not, in fact, succeed in acquiring wh-movement. More tentatively, our findings suggest that the failure of the networks is due to the insufficient richness of the linguistic input and not to inadequacies of the networks, thus supporting an APS, the first that is based on successful learners exposed to realistic amounts of linguistic input.

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 linguistic biases is through cases in which 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 (Chomsky, 1971, pp. 26–8, Chomsky, 1975, pp. 30ff., Chomsky, 1980, p. 34) it has been central to the investigation of the human linguistic capacity.[1]

[1]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

A well-known example of an APS concerns the reversal of the order between the subject and the auxiliary in yes/no questions in languages like English. Specifically, it has been claimed that when children are faced with the choice of which of two auxiliary verbs to place before the subject in a yes/no question, they systematically choose the highest auxiliary over the linearly first auxiliary. Consider (1), for example, which has two auxiliary verbs, *is* and *might*. In order to turn this sentence into a yes/no question speakers never place the linearly first but hierarchically lower *is* before the subject, as in (2a); instead, they systematically place the hierarchically higher but linearly second *might* before the subject, as in (2b).

(1) The cat that is running might jump

(2) a. * Is the cat that running might jump?

b. Might the cat that is running jump?

Significantly, it has been claimed by Chomsky (1971) and others that the input available to the child underdetermines this choice and does not contain enough information to justify the choice of higher over leftmost auxiliary. If true, more-or-less neutral learners would presumably have no strong preference between the two options, which in turn would suggest that children are not neutral learners of this kind: their innate endowment prepares them to acquire the highest-auxiliary generalization rather than the leftmost-auxiliary generalization.[2]

APSs of this kind 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 language is not trivial, it is something that many learning algorithms do. In particular, many linguistically-naive general-purpose learning approaches can handle this kind of generalization. And importantly such generalization does not imply any interesting linguistic biases such as the putative preference for hierarchical over linear transformations which the subject-auxiliary reversal data have been taken to support.

While 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, but this 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

---

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.

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]

The challenge of assessing the information available to the child has become less of an obstacle lately, with the advent of general-purpose artificial neural networks (ANNs) that can be trained on very large corpora. A remaining difficulty, however, has been assessing what these networks know and checking this knowledge against the kind of knowledge that features in APSs. In recent work, Wilcox et al. (2021; WFL) develop a new methodology (building on a paradigm from psycholinguistics) that probes the knowledge of various artificial neural networks and makes it possible to start asking whether the relevant networks can acquire human-like knowledge based on realistic corpora. WFL apply this methodology to the study of wh-movement and argue that ANNs not only learn the basic dependency between a wh filler and a subsequent gap but actually succeed in acquiring various island constraints that restrict wh-dependencies. This, WFL suggest, debunks an APS: one that says that the input is insufficiently rich to allow a general-purpose learner to acquire islands.[4] More broadly, WFL take their results to suggest that our innate biases might be linguistically neutral.

The present paper builds on WFL's methodology to probe ANNs' knowledge of wh-movement, arriving at conclusions that are at odds with those of WFL. We start, in Section 2, by reviewing WFL's methodology and its rationale. In Section 3 we proceed to show that the scope of the ANNs' success is rather limited: while, as WFL show, the ANNs have clearly learned something about wh-movement and perhaps about islands, this knowledge falls short of humans' knowledge in many important ways. In particular, ANNs fail to exhibit an adequate 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 the ANNs does not, in fact, debunk this APS. More tentatively, we provide evidence that the failure of the ANNs stems not from inadequacies of the ANNs themselves but rather from the insufficient richness of the linguistic input. If correct, this evidence constitutes the first APS in the literature in which otherwise successful general-purpose learners are trained on realistic amounts of linguistic data (and in some cases, amounts of data that are significantly greater than anything children are exposed to) and yet fail to acquire knowledge that humans have. Section 4 concludes.

[3]See Pullum and Scholz (2002), Legate and Yang (2002), Lidz et al. (2003), Foraker et al. (2009), and Hsu and Chater (2010), among others, for 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.

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

2 Methodological preliminaries

WFL, as just mentioned, investigate the extent to which a linguistically-neutral learner would succeed in acquiring various aspects of linguistic knowledge from a corpus that is broadly similar to the input available to the child learner. If a linguistically-neutral learner succeeds in this task with respect to a given piece of knowledge, this would undo an APS concerning this piece of knowledge. The knowledge that WFL focus on concerns wh-movement. Simplifying considerably, a *gap*, such as the missing complement of 'with' in (3a) and (3c), appears if and only if it is preceded by an appropriate *filler*, such as the wh-phrase 'who' in (3a) and (3b). When there is both a filler and a gap (3a) or neither (3d) the result is good; when there is a filler and no gap (3b) or a gap and no filler (3c) the result is bad.[5]

(3) 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. In an ideal world, one would (a) take a sufficiently powerful learner that can be seen to not be biased in favor of the knowledge of these dependencies; (b) train this learner on the relevant corpus; and (c) check whether the learner has indeed acquired the knowledge under consideration. In such an ideal world, one might perhaps be able to work with a Bayesian program induction algorithm for a general-purpose programming language such as LISP or Python using a description-length prior, where neither the programming language nor the learning algorithm can be taken to bias the learning in the direction of human-like knowledge of wh-movement, and where the knowledge acquired by the algorithm can 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, since many learning models could not be trained on realistic corpora (corresponding to years of human linguistic experience) in the first place, while those models that could be trained on such corpora — such as n-gram models — were insufficiently powerful to acquire or even represent sophisticated linguistic knowledge such as wh-movement dependencies.

More recently, ANNs have changed things considerably: these models are generally highly successful in acquiring sequential dependencies, they are arguably linguistically neutral, and they can be trained on very large corpora. Unfortunately, while ANNs offer a way past parts (a) and (b) they still pose a challenge with respect to part (c): it is all but impossible to inspect them and determine what they know. 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'.

[5]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>...' ($-filler$).

What ANNs do tell us is how *likely* they consider each such 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. 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*.[6]

In the cases we are interested in here, however, probability and grammaticality are often quite well aligned, and it is easy to find examples such as (3) in which the grammatical continuation is significantly more probable than the ungrammatical one. So if we focus on such cases where grammaticality and probability are aligned, and if ANNs are sufficiently good learning models, then we can use the probabilistic predictions of the ANNs to evaluate the APS. If the ANNs systematically prefer the grammatical continuation, this can be taken to suggest that the pattern of wh-movement is represented in the input data sufficiently well so as to allow a learner to pick it up. While it remains unclear whether ANNs themselves have a representation of grammaticality — the target of the APS — as distinct from probability, their success would suggest that a learner that does have such a representation might acquire the pattern. Conversely, if the ANNs do not systematically prefer the grammatical continuation, this would indicate that the pattern of wh-movement is not adequately represented in the input data.

The naive use of ANN preferences to evaluate the APS, as outlined above, assumes that ANNs are indeed sufficiently good learning models. This is an important qualification, and we will now discuss two possible concerns with ANNs and the extent to which they satisfy this qualification.

A first concern is that the ANNs might be biased against the relevant dependencies or possibly even too weak altogether to capture them. If that happens to be the case, then failure to systematically prefer the grammatical continuation will not teach us about the richness of the input data and will therefore be irrelevant for assessing the APS. This imperfection on the part of ANNs, however, is not very likely in the present case: ANNs have evolved over the past decades so as to succeed in capturing key patterns in linguistic sequences, so if anything it seems likelier that they are biased in favor of the relevant patterns (and therefore not as linguistically-neutral as advertised after all) rather than against them. Still, in order to reduce the risk of this problem it will be useful to check that the ANNs can, in fact, succeed when the input is clearly rich. We return to this matter in Section 3 and provide evidence that at least one ANN is indeed capable of achieving improved success when the training corpus is sufficiently rich.

The second concern is that the ANN might be letting irrelevant factors obscure

[6] 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. However, since our goal is not to study ANNs for their own sake but rather to use them to evaluate the informativeness of the input data, the inadequacy of ANNs as cognitive models is not inherently problematic.

the extent to which it has acquired the pattern of wh-movement. Using (3) as an example, it could be that the ANN correctly assigns a higher probability to 'yesterday' (3a) than to 'Mary' (3b) but that it does so for the wrong reasons (perhaps because the lexical frequency of 'yesterday' just happens to be higher than that of 'Mary'). The network's success, then, should not count as evidence for its grasp of wh-movement (and against the APS). Conversely, it could be that the ANN has, in fact, acquired a good knowledge of wh-movement but that it incorrectly prefers 'Mary' to 'yesterday' because of similarly irrelevant reasons (perhaps the lexical frequency of 'Mary' is very high). The network's failure, then, should not count as evidence against its knowledge of wh-movement (and for the APS). Again, a good enough learner would not get confused by such irrelevant factors as lexical frequency, but we have no reason to think that ANNs are indeed good enough in this sense.

This second worry is serious, but following WFL we can mitigate it to some extent by looking at a given ANN's preferences across full paradigms such as (3) and not just at those portions of the paradigm in which a filler is present (which would have sufficed if the learning model could have been trusted to be good enough and to not be confused by irrelevant factors such as lexical frequencies). Specifically, using the logic of difference-in-differences, we can compare the ANN's relative probabilities for 'yesterday' (gap) and 'Mary' (no gap) when there is a filler upstream with the relative probabilities when such a filler is absent. We will consider the network successful on a given example if its preference for a gap is higher when the gap follows a filler than when it does not. This alleviates the worry about an independent preference for one of the target words over the other leading to spurious successes or failures.

Concretely, and following WFL and other works, we can implement this broader comparison across the full paradigm as follows. We consider the *surprisal* of the continuation given the prefix, $S(\text{continuation}|\text{prefix}) = -\lg P(\text{continuation}|\text{prefix})$; that is, the negative of the logarithmically-scaled conditional probability of the continuation given the prefix. The lower the probability the higher the surprisal; when the probability approaches 0 the surprisal tends to infinity, and as the probability approaches 1 the surprisal approaches 0.

We can now take the preference for the non-gap over the gap in the presence of the filler 'who' in (3) to be: $\Delta_{+filler} = S(\text{'yesterday'}|\text{'I know }\underline{\text{who}}\text{ you talked with'}) - S(\text{'Mary'}|\text{'I know }\underline{\text{who}}\text{ you talked with'})$. Analogously: $\Delta_{-filler} = S(\text{'yesterday'}|\text{'I know }\underline{\text{that}}\text{ you talked with'}) - S(\text{'Mary'}|\text{'I know }\underline{\text{that}}\text{ you talked with'})$. The modified criterion for ANN success, then, can be implemented as whether $\Delta_{+filler} < \Delta_{-filler}$, or, equivalently, whether $\Delta_{-filler} - \Delta_{+filler} > 0$.

The above, then, explains why and how the probabilistic preferences of ANNs can serve to probe the APS. On the conditions discussed above, systematic success in terms of probabilities strongly suggests that the input is sufficiently rich to support the acquisition of the relevant linguistic knowledge by a linguistically-neutral learner. And the lack of systematic success is similarly suggestive of an input that is too impoverished to support the acquisition of the relevant knowledge by a linguistically-neutral learner. If we consider a given piece of linguistic knowledge that human speakers have, and if we show that ANNs fail to show systematic success when trained on a corpus that is at least as rich as what is available to the child learner, then we have an APS.

In the next section we build on WFL and apply the methodological considerations

Model	Tokens in training data	Human equivalent
GRNN	90 million	8-year-old
JRNN	1 billion	80-year-old
GPT-2	$\sim$8 billion	10 lifetimes
GPT-3	$\sim$114 billion	100 lifetimes

Table 1: Training data size of the four language models considered here, and the human linguistic experience equivalent to these data sizes; based on estimates by Wilcox et al. (2021) who assume a daily exposure to $\sim$30,000 words by human children.

above to examine the richness of several linguistic corpora, asking how well these corpora support the acquisition of wh-dependencies by a linguistically-neutral learner.

3 Evaluating the APS for wh-movement

How rich is the input, then, when it comes to filler-gap dependencies of the wh kind? We follow WFL in evaluating this question using four language models which achieved state-of-the-art results on various NLP benchmarks: JRNN (Jozefowicz et al., 2016), GRNN (Gulordava et al., 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020). According to WFL's estimates, each of these models was exposed to amounts of data ranging from eight years of linguistic experience (GRNN) to 10 and 100 human lifetimes (GPT-2 and 3); see Table 1. Indeed, WFL admit that this linguistic experience is probably above and beyond that of human children, and could thus weaken their argument against APS in case of successful learning by the models. However, in the current work the size of the training corpora contributes to our 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.

3.1 ANNs succeed in very simple cases

In simple cases such as (3) above, the ANNs considered by WFL succeed directly, in terms of assigning a higher probability to the grammatical continuation than to the ungrammatical one: Figure 1 plots raw surprisal values for sentences (3a) and (3b) that make up a $\Delta_{+filler}$ pair. All models assign a lower surprisal value (i.e., a higher probability) to the grammatical continuation 'yesterday' in the gapped sentence than to 'Mary'.

WFL further show that the ANNs go beyond the basic knowledge that fillers and gaps go hand in hand. Specifically, they provide evidence that suggests that ANNs are aware of *islands* (Ross, 1967): configurations in which a gap is bad even if there is a filler upstream. Examples include subjects as in (4) and coordinate structures as in (5):

(4) * I know who [Mary's talking to ___] insulted John.

(5) * I know who Mary [talked to ___ yesterday] and [will insult you tomorrow].

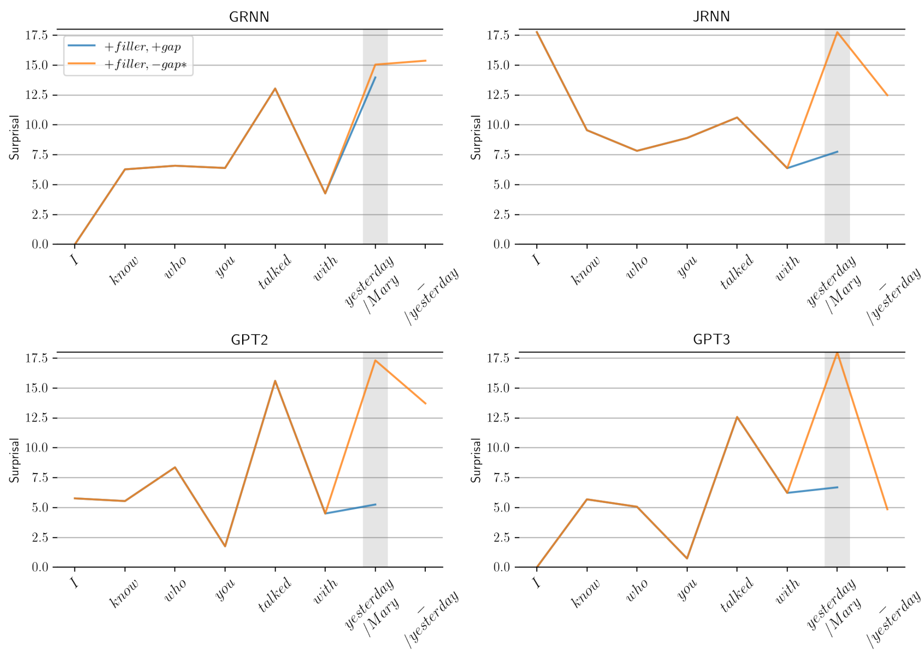


Figure 1: Raw surprisal values outputted by the four ANNs under consideration for the grammatical (3a), in blue, and ungrammatical (3b), in orange. All models correctly output lower surprisal values for the grammatical continuation.

WFL consider the predictions of the ANNs in cases that involve islands. While, as discussed above, a filler upstream generally increases the ANNs' expectation of a gap downstream, this expectation should be reduced within islands. In (4), for example, the expectation of a gap immediately after the subject-internal 'to' should be low. Consequently, it should be surprising to see 'insulted' as the following word, since it indicates a gap. And it should be less surprising to see 'John', since it indicates that there was no gap. WFL show that in many cases of this kind, the ANNs indeed show a reduced expectation of a gap within islands.[7]

The ANNs' performance suggests that they have learned something nontrivial regarding wh-movement. Indeed, WFL take this performance to indicate that the ANNs have acquired an adequate knowledge of the relevant dependencies. If true, this suggests that linguistically-neutral learners can learn the intricacies of wh-movement from the input data. In other words, the input data are not impoverished after all with respect

[7]Things are more involved than this brief sketch suggests. In (4), for example, the sentence itself is indeed ungrammatical, but the ungrammaticality cannot be determined at the site of the subject-internal gap, since it is possible to rescue this prefix with a further gap in the matrix object position, a matter that we return to shortly. And in (5), where ungrammaticality can be determined at the gap site, a non-gap (e.g., a name such as 'Kim') will also lead to grammaticality. Such considerations complicate the interpretation of probabilistic expectations in simple islands. As far as we can tell, these complications do not arise in the cases of parasitic gaps and across-the-board movement that we focus on below.

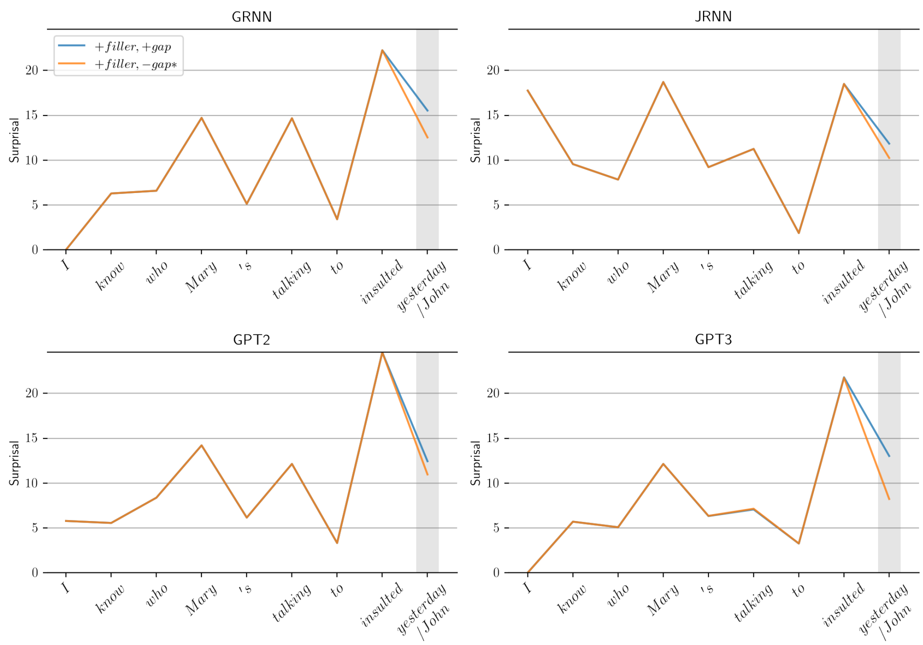


Figure 2: Raw surprisal values for the ungrammatical sentence (4) which violates a subject island, in orange, and its grammatical variant (6), in blue, where a parasitic gap makes it possible to escape the island. For measuring the model's expectation for a gap, surprisal is measured at the adjunct 'yesterday', 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.

to wh-movement, and an APS in this domain falls apart.

We wish to probe the performance of the ANNs — and through that, the richness of the stimulus — further. We will argue that the ANNs do not, in fact, achieve adequate knowledge of wh-movement and therefore do not debunk the APS in this domain. More tentatively, we will provide evidence suggesting that the ANNs' failure is attributable to the input data. If correct, the APS in this domain stands.

3.2 ANNs fail on slightly more complex (but still simple) cases

We start with a well-studied nuance of islands: in various cases, an otherwise impossible gap inside an island is made possible by a separate gap elsewhere. Compare the bad (4) above to the good (6), and the bad (5) above to the good (7).

(6) I know who Mary's talking to __ insulted __ .

(7) I know who Mary talked to __ yesterday and will insult __ tomorrow.

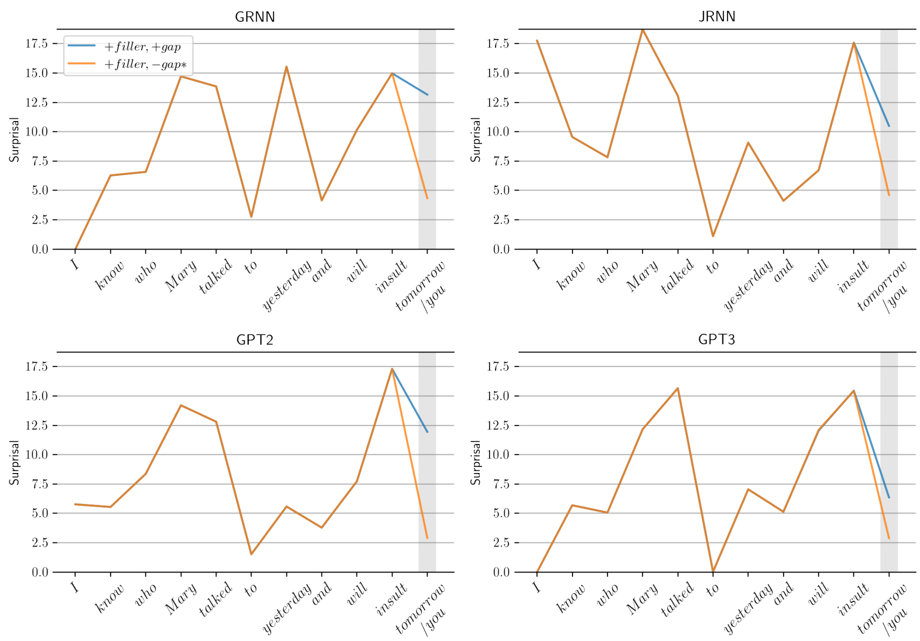


Figure 3: Raw surprisal values for the ungrammatical sentence (5) which violates the coordinate structure constraint (orange), and its grammatical variant (7) where ATB movement makes it possible to avoid the constraint (blue). All networks wrongly assign a higher surprisal value to the grammatical continuation 'tomorrow' rather than to 'you'.

As mentioned above and illustrated in (4), a gap inside a subject is bad. The gap becomes acceptable when, in (6), a second gap is added in the direct-object position. This phenomenon is known as a *parasitic gap* (PG): the gap inside the subject island becomes acceptable parasitically, based on the direct-object gap. We also mentioned that a coordinate structure is generally an island, as illustrated in (5). But wh-movement across a coordinate structure becomes acceptable when, in (7), a gap is added to the second conjunct. This phenomenon is known as *across-the-board movement* (ATB).[8]

Are the ANNs aware of PG and ATB? Figures 2-3 illustrate that all the ANNs under consideration prefer the ungrammatical continuation over the grammatical one in seemingly simple cases. This seems to indicate that the ANNs have failed to acquire a knowledge of the relevant constructions, which in turn suggests that the input is insufficiently rich.

Recall our earlier discussion, however. While the failure of the ANNs might be due to the input being too impoverished, which is the point we wish to examine, it might

[8]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.

also be due to inadequacies of the ANNs themselves. Above we mentioned two possible kinds of inadequacy: (a) the ANNs might be too weak or somehow biased against the dependencies under consideration, which could prevent them from acquiring the relevant knowledge; and, (b) regardless of whether they have acquired a knowledge of the relevant dependency, they might give us confusing answers because of irrelevant factors such as lexical frequency.

We will partially address the first worry in Section 3.3 below by showing that at least one of the networks does, in fact, succeed when the input data are sufficiently rich. Before that, we will control for the second worry by modifying the criterion for success, as discussed above. If the ANNs acquired the knowledge of the pattern but are hiding this knowledge through the effects of irrelevant facts relating to lexical frequency, checking whether $\Delta_{+filler} < \Delta_{-filler}$ might allow us to see this. The following tables illustrate the ingredients of the Δ's for PG (8) and ATB (9). Underlined words indicate the $\pm$filler alternations. Words in bold indicate the critical region where surprisal is measured for comparison within each condition. In each row of tables 8 and 9, the presence of a gap in the $+$gap condition (first column) becomes evident when a reader reaches the words 'today' or 'tomorrow' (i.e. the direct object is missing); conversely, in the $-$gap condition (second column) in each case, the direct object 'John' is where the absence of a gap becomes evident.

(8) PG

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Mary's talking to insulted **today**	*I know <u>who</u> Mary's talking to insulted **John** today
$-filler$	*I know <u>that</u> Mary's talking to <u>me</u> insulted **today**	I know <u>that</u> Mary's talking to <u>me</u> insulted **John** today

(9) ATB

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Mary talked to yesterday and will insult **tomorrow**	*I know <u>who</u> Mary talked to yesterday and will insult **John** tomorrow
$-filler$	*I know <u>that</u> Mary talked to <u>Frank</u> yesterday and will insult **tomorrow**	I know <u>that</u> Mary talked to <u>Frank</u> yesterday and will insult **John** tomorrow

In order to go beyond a handful of hand-picked examples and test the performance of the networks on PG and ATB sentences more broadly, we manually constructed context-free grammars to generate a variety of paradigms similar to those in (8) and (9). 6,144 sentence tuples were generated for PG and 5,552 for ATB. Excerpts from the CFGs and the sentences they generate are given in Figure 4. The full grammars are given in Appendix A.[9] For each such combination of sentences, surprisal was measured at the critical positions and the $\Delta_{\pm filler}$ values were computed. As explained in

[9]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.

Section 2, $\Delta_{+filler} < \Delta_{-filler}$ indicates that the network trends in the right direction: it is more confident that a gap should appear when a filler is present than when it is not.

Figure 5 plots the results of examining the Δ's. At first blush the results look encouraging for some ANNs: for PG, the most successful model is GPT-3, whose preferences for about 98% of the cases go in the correct direction. The rest of the models range from 69% to 88% success rates. For ATB, the networks' preferences go in the right direction 90%-99% of the time.

Upon closer inspection, however, the performance of the ANNs appears less successful. Even in the case of the best performing model, 1% is a non-negligible proportion, representing failure on many grammatical sentences. Figure 6 visualizes the models' outputs for some of these sentences. A more exhaustive list of model failures is given in Appendix B. We note that predicting an island violation is not immediately problematic for the models: exceptions to islands were already pointed out by Ross (1967) and have been discussed extensively in subsequent work. Moreover, one might be lenient toward ANN failures in cases that involve rare vocabulary choices or unusual structural properties. As the reader can verify, however, many of the ANNs' failures occur with simple examples with frequent lexical choices and no unusual properties in which human judgments are clear and directly conflict with the ANN predictions.

That such failure occurs even with the modified criterion for success — that is, with checking whether $\Delta_{+filler} < \Delta_{-filler}$ — strongly suggests that the ANNs did not, in fact, acquire an adequate knowledge of PG and ATB. We conclude that the behavior of the ANNs that WFL consider does not, in fact, debunk the APS in the domain of wh-movement.

3.3 GRNN seems to succeed when retrained on a richer corpus

As mentioned above, we wish to make a further, somewhat tentative step and provide evidence that the ANNs have failed not because of their inability to acquire the knowledge of wh-dependencies but rather because the input is insufficiently rich. If this conclusion is correct, it will constitute an APS in the domain of wh-movement, one that is based on generally successful learning models and large linguistic corpora.

In order to make the point that the failure of the ANNs is due to the input and not to the ANNs themselves, we checked what happens when the corpus is clearly not impoverished. To do so, we took one of the learners, GRNN, and retrained it on a corpus that was identical to the original training data (English Wikipedia) but with an addition of many instances of PG and ATB, generated from the same manually-constructed CFG described above.[10] Overall 5,440 extra sentences were generated for the PG case and 2,980 sentences for ATB.

To make sure that the model would not simply memorize the training sentences, we generated the training and test sets using the following recipe: for each node in a CFG derivation, part of the terminal values (65%) are selected for generating test sentences. The test set consists of all combination of these values. From the set of test terminal values, we compute all pairs of lexical choices and remove all sentences which contain

[10] Retraining GPT-2 and GPT-3 using this regime is currently impossible: the original training data for these models have not been released, and GPT-3 itself is proprietary. For JRNN such retraining is possible in principle but not feasible due to the large computation power required, see Jozefowicz et al. (2016).

PG Grammar

$S \rightarrow \langle PREAMBLE \rangle \, \langle \pm F \rangle \, \langle \pm G \rangle$
$\langle PREAMBLE \rangle \rightarrow$ *I know*
$\langle +F \rangle \rightarrow$ *who* $\langle NAME1 \rangle \langle GEN \rangle \, \langle SUBJ \rangle \, \langle V \rangle$
$\langle -F \rangle \rightarrow$ *that* $\langle NAME1 \rangle \langle GEN \rangle \, \langle SUBJ \rangle \, \langle \underline{NAME2} \rangle \, \langle V \rangle$
$\langle +G \rangle \rightarrow$ ⟨**ADJUNCT**⟩
$\langle -G \rangle \rightarrow$ ⟨**NAME3**⟩ $\langle ADJUNCT \rangle$
$\langle NAME1 \rangle \rightarrow$ *Alice* | *Bob* | ... | *John*
$\langle GEN \rangle \rightarrow$ *'s*
$\langle SUBJ \rangle \rightarrow$ *talking to* | *friendship with* | ... | *praising of*
$\langle V \rangle \rightarrow$ *bothered* | *excited* | ... | *annoyed*
$\langle ADJUNCT \rangle \rightarrow$ *today* | *yesterday* | ... | *lately*
...
⇒ I know who John's friendship with bothered **yesterday**. $_{(+filler,+gap)}$
⇒ *I know who John's friendship with bothered **William** yesterday. $_{(+filler,-gap)}$
⇒ *I know that John's friendship with Mary bothered **yesterday**. $_{(-filler,+gap)}$
⇒ I know that John's friendship with Mary bothered **William** yesterday. $_{(-filler,-gap)}$

ATB Grammar

$S \rightarrow \langle PREAMBLE \rangle \, \langle \pm F \rangle \, \langle CONN \rangle \, \langle \pm G \rangle$
$\langle PREAMBLE \rangle \rightarrow$ *I know*
$\langle +F \rangle \rightarrow$ *what* $\langle NAME1 \rangle \, \langle V1 \rangle$
$\langle -F \rangle \rightarrow$ *that* $\langle NAME1 \rangle \, \langle V1 \rangle \, \langle \underline{OBJ1} \rangle$
$\langle +G \rangle \rightarrow \langle V2 \rangle$ ⟨**ADJUNCT**⟩
$\langle -G \rangle \rightarrow \langle V2 \rangle$ ⟨**OBJ**⟩ $\langle ADJUNCT \rangle$
$\langle CONN \rangle \rightarrow$ *yesterday and will*
$\langle V1 \rangle \rightarrow$ *looked for* | *found* | ... | *went shopping for*
$\langle OBJ1 \rangle \rightarrow$ *food* | *candy* | ... | *bread*
$\langle V2 \rangle \rightarrow$ *devour* | *serve* | ... | *donate*
$\langle OBJ2 \rangle \rightarrow$ *it* | *fish* | ... | *snacks*
$\langle ADJUNCT \rangle \rightarrow$ *tomorrow* | *soon* | ... | *tonight*
...
⇒ I know what John looked for yesterday and will devour **tomorrow**. $_{(+filler,+gap)}$
⇒ *I know what John looked for yesterday and will devour **it** tomorrow. $_{(+filler,-gap)}$
⇒ *I know that John looked for food yesterday and will devour **tomorrow**. $_{(-filler,+gap)}$
⇒ I know that John looked for food yesterday and will devour **it** tomorrow. $_{(-filler,-gap)}$

Figure 4: Context-free grammars used to generate PG and ATB sentences for the experiments in Section 3, and sample sentences generated from each grammar. Words that alternate according to the $\pm filler$ condition are underlined; words in bold mark the position where the $\pm gap$ condition becomes evident. The full grammars are given in Appendix A.

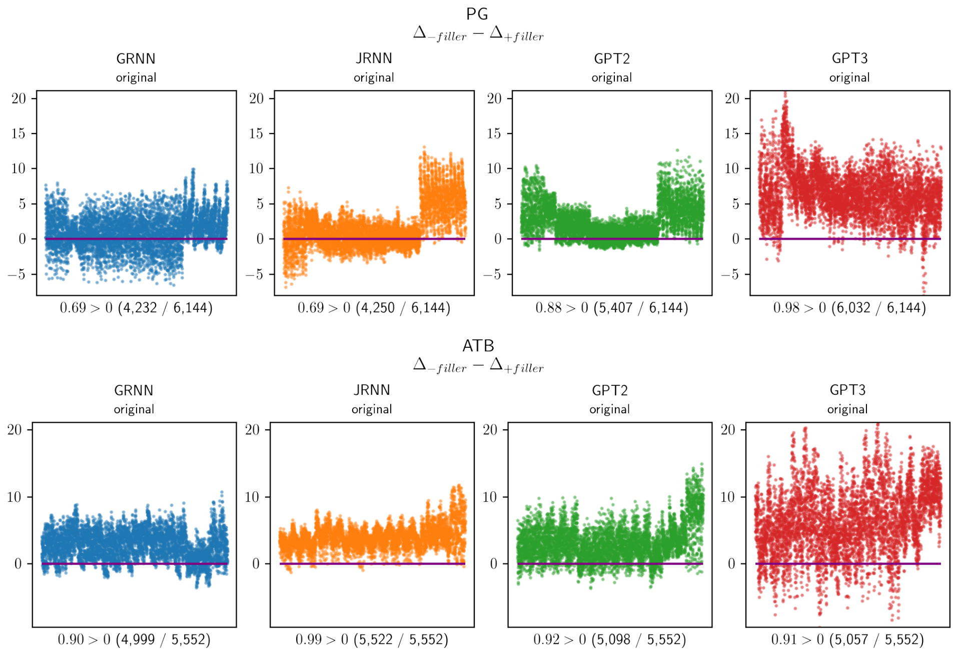


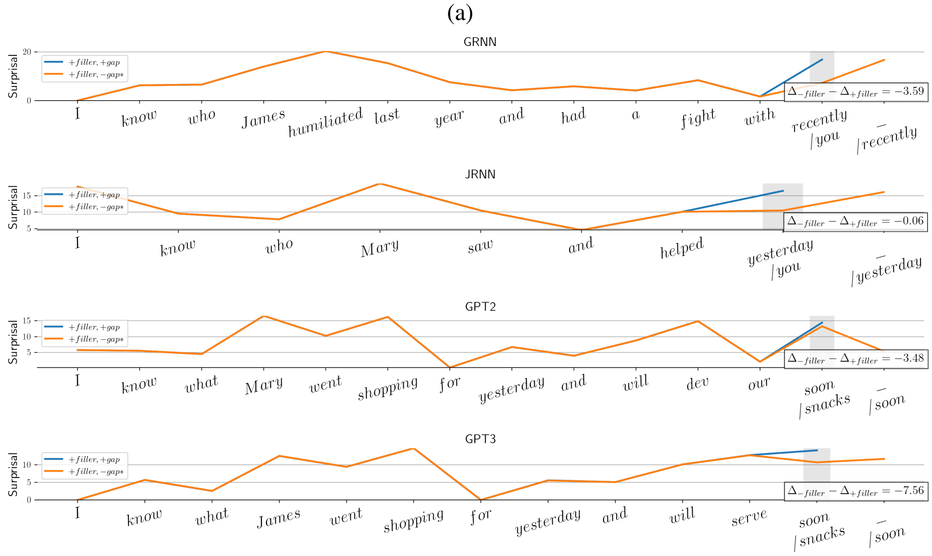
Figure 5: Model performance on PG and ATB sentences generated from context-free grammars. Each point represents $\Delta_{-filler} - \Delta_{+filler}$ for one four-tuple of sentences such as (8) and (9). Positive values thus indicate that a network's outputs trend in the right direction, i.e. that it is more confident that a gap should appear when a filler is present than when it is not. Below each plot is the success rate, representing the ratio of tuples for which the model's predictions trend in the right direction.

these pairs from the remaining training set. In this way, the model is never exposed to co-occurrences of lexical choices which appear in the test set; good performance on the test set thus indicates that the model went beyond memorizing lexical combinations. Figure 7 illustrates the training and test data generation process.

The model's performance on the training and test set, before and after retraining, is visualized in Figure 8. For PG the model improves from 77% to 100%, suggesting that the network is perfectly capable of acquiring the dependencies under consideration when the corpus is sufficiently rich. For ATB the model improved from 88% to 98%.

The dramatic improvement in performance after retraining is compatible with the model being perfectly capable of learning PG and ATB given enough data. In this case, the model's initial failure when trained on an equivalent of eight years of linguistic experience constitutes an APS. WFL's methodology, then, far from undermining the APS from wh-movement, would actually further support it. Alternatively, given the non-perfect results for ATB, it is conceivable that the model is simply incapable of learning phenomena such as ATB. If this is the case, however, then it is not an adequate model for evaluating the APS in the first place.

(a)



(b)

Figure 6: Surprisal values for example sentences automatically generated for PG (a) and ATB (b). The plotted values are for the $\Delta_{+filler}$ pairs, and the complete $\Delta_{-filler} - \Delta_{+filler}$ calculation of the four-tuple of sentences is given for each plot. An extensive list of model failures is given in Appendix B.

$\langle NAME1 \rangle$ $\rightarrow$ *Alice* | *Bob* | ... | *John* $\langle SUBJ \rangle$ $\rightarrow$ *talking to* | *friendship with* | ... | *praising of* $\langle V \rangle$ $\rightarrow$ *bothered* | *excited* | ... | *annoyed* $\langle ADJUNCT \rangle$ $\rightarrow$ *today* | *yesterday* | ... | *lately* ...
Training set
I know who Alice's praising of annoyed lately. I know who John's talking to annoyed lately. I know who John's praising of bothered lately. I know who John's praising of annoyed today. ...
Test set
I know who Alice's talking to bothered today. I know who Alice's talking to bothered yesterday. ... I know who Bob's friendship with excited yesterday.

Figure 7: Example generation of training and test data containing extra PG and ATB sentences, used to test GRNN's ability to learn these phenomena given more exposure. For each lexical category in the CFG (here for PG), a subset of its lexical choices is reserved for the test set (in red). Sentences in which these items co-occur are removed from the training set. This ensures that the model cannot simply memorize co-occurrence frequencies in order to predict the grammatical continuation of a sentence at the relevant gap site, and must use at least some structural cues.

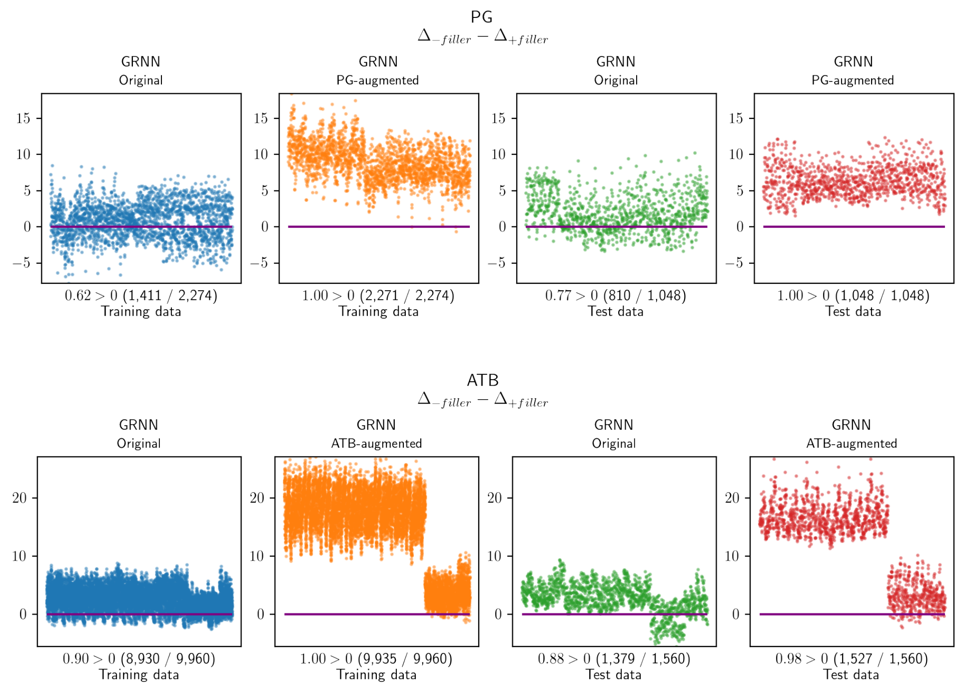


Figure 8: GRNN performance on the training and test sets before and after retraining with additional sentences for PG and ATB. Each point represents $\Delta_{-filler} - \Delta_{+filler}$ for one four-tuple of sentences such as (8) and (9). Positive values indicate that the network's outputs trend in the right direction. The retrained model shows a dramatic improvement for both PG and ATB, suggesting that the network is capable of acquiring the dependencies under consideration when the corpus is sufficiently rich.

4 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 WFL show us how. WFL conclude 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: when ANNs' knowledge of wh-movement is probed beyond the most basic aspects of these dependencies, the models make predictions that are clearly wrong. We illustrated this with parasitic gaps and across-the-board movement.

Is it possible that some future linguistically-neutral learner will succeed where the four ANNs have failed? Of course. But we note that the architectures we have considered are extremely successful and have shown an impressive ability to learn many other aspects of linguistic data. And they have been provided with very generous amounts of input, ranging from the equivalent of about eight years (JRNN) to many thousands of years (GPT-3). Given that none of the ANNs achieved systematic success — and given that at least one network seemed capable of achieving systematic success 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. In that case, humans' knowledge of these aspects means that humans are innately endowed in ways that are not linguistically neutral. In other words, the APS from wh-dependencies stands.

References

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.

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

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.

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.

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.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings* *of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205. New Orleans, Louisiana: Association for Computational Linguistics.

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

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.

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

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

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.

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.

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.

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.

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.

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

Wilcox, Ethan, Richard Futrell, and Roger Levy. 2021. Using computational models to test syntactic learnability. Ms., available at https://ling.auf.net/lingbuzz/006327, November 2021.

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.

A Appendix: context-free grammars

A.1 PG

A.1.1 Grammar 1

$\langle S\rangle \rightarrow \langle S\_FG\rangle$
$\langle S\_FG\rangle \rightarrow \langle PREAMBLE\rangle\ \langle F\rangle\ \langle G\rangle$
$\langle S\_XG\rangle \rightarrow \langle UNGRAMMATICAL\rangle\ \langle PREAMBLE\rangle\ \langle XF\rangle\ \langle G\rangle$
$\langle S\_FX\rangle \rightarrow \langle UNGRAMMATICAL\rangle\ \langle PREAMBLE\rangle\ \langle F\rangle\ \langle XG\rangle$
$\langle S\_XX\rangle \rightarrow \langle PREAMBLE\rangle\ \langle XF\rangle\ \langle XG\rangle$
$\langle UNGRAMMATICAL\rangle \rightarrow$ '*'
$\langle NAME1\rangle \rightarrow$ 'Michael' | 'Ashley' | 'Daniel' | 'John' | 'Brandon' | 'William' | 'Nicole' | 'Eric' | 'Melissa' | 'Timothy'
$\langle NAME2\rangle \rightarrow$ 'Christopher' | 'Jennifer' | 'David'
$\langle NAME3\rangle \rightarrow$ 'Jessica' | 'Joshua' | 'James'
$\langle NAME4\rangle \rightarrow$ 'Matthew' | 'you'
$\langle PREAMBLE\rangle \rightarrow$ 'I know'
$\langle F\rangle \rightarrow$ 'who' $\langle NAME1\rangle\ \langle GEN\rangle\ \langle ADJ\rangle\ \langle SUBJ\rangle\ \langle V\rangle$
$\langle XF\rangle \rightarrow$ 'that' $\langle NAME1\rangle\ \langle GEN\rangle\ \langle ADJ\rangle\ \langle SUBJ\rangle\ \langle NAME2\rangle\ \langle V\rangle$
$\langle G\rangle \rightarrow \langle ADJUNCT\rangle$
$\langle XG\rangle \rightarrow \langle NAME4\rangle\ \langle ADJUNCT\rangle$
$\langle GEN\rangle \rightarrow$ ''s'
$\langle SUBJ\rangle \rightarrow$ 'talking to' | 'attitude towards' | 'friendship with' | 'praising of'
$\langle ADJ\rangle \rightarrow$ 'recent' | 'current'
$\langle V\rangle \rightarrow$ 'bothered' | 'distracted' | 'worried' | 'annoyed'
$\langle ADJUNCT\rangle \rightarrow$ 'recently' | 'yesterday' | 'lately'

A.1.2 Grammar 2

$\langle S\rangle \rightarrow \langle S\_FG\rangle$
$\langle S\_FG\rangle \rightarrow \langle PREAMBLE\rangle\ \langle F\rangle\ \langle G\rangle$
$\langle S\_XG\rangle \rightarrow \langle UNGRAMMATICAL\rangle\ \langle PREAMBLE\rangle\ \langle XF\rangle\ \langle G\rangle$
$\langle S\_FX\rangle \rightarrow \langle UNGRAMMATICAL\rangle\ \langle PREAMBLE\rangle\ \langle F\rangle\ \langle XG\rangle$
$\langle S\_XX\rangle \rightarrow \langle PREAMBLE\rangle\ \langle XF\rangle\ \langle XG\rangle$
$\langle UNGRAMMATICAL\rangle \rightarrow$ '*'
$\langle NAME1\rangle \rightarrow$ 'Michael' | 'Ashley' | 'Daniel' | 'John' | 'Brandon' | 'William' | 'Nicole' | 'Eric' | 'Melissa' | 'Timothy'
$\langle NAME2\rangle \rightarrow$ 'Christopher' | 'Jennifer' | 'David'
$\langle NAME3\rangle \rightarrow$ 'Jessica' | 'Joshua' | 'James'
$\langle NAME4\rangle \rightarrow$ 'Matthew' | 'you'
$\langle PREAMBLE\rangle \rightarrow$ 'I know'
$\langle F\rangle \rightarrow$ 'what' $\langle SUBJ\rangle\ \langle V1\rangle\ \langle ADV\rangle\ \langle V2\rangle$
$\langle XF\rangle \rightarrow$ 'that' $\langle SUBJ\rangle\ \langle V1\rangle\ \langle OBJ1\rangle\ \langle ADV\rangle\ \langle V2\rangle$
$\langle G\rangle \rightarrow \langle ADJUNCT\rangle$
$\langle XG\rangle \rightarrow \langle OBJ2\rangle\ \langle ADJUNCT\rangle$
$\langle SUBJ\rangle \rightarrow$ 'the attempt to'
$\langle V1\rangle \rightarrow$ 'repair' | 'fix' | 'overhaul' | 'rebuild'

$\langle OBJ1 \rangle \rightarrow$ 'the car' | 'the bike' | 'the washing machine' | 'the drier' | 'the ceiling' | 'the apartment'
$\langle ADV \rangle \rightarrow$ 'eventually' | 'finally'
$\langle V2 \rangle \rightarrow$ 'damaged' | 'destroyed' | 'ruined' | 'wrecked'
$\langle OBJ2 \rangle \rightarrow$ 'it'
$\langle ADJUNCT \rangle \rightarrow$ 'nevertheless' | 'nonetheless'

A.1.3 Grammar 3

$\langle S \rangle \rightarrow \langle S\_FG \rangle$
$\langle S\_FG \rangle \rightarrow \langle PREAMBLE \rangle \ \langle F \rangle \ \langle G \rangle$
$\langle S\_XG \rangle \rightarrow \langle UNGRAMMATICAL \rangle \ \langle PREAMBLE \rangle \ \langle XF \rangle \ \langle G \rangle$
$\langle S\_FX \rangle \rightarrow \langle UNGRAMMATICAL \rangle \ \langle PREAMBLE \rangle \ \langle F \rangle \ \langle XG \rangle$
$\langle S\_XX \rangle \rightarrow \langle PREAMBLE \rangle \ \langle XF \rangle \ \langle XG \rangle$
$\langle UNGRAMMATICAL \rangle \rightarrow$ '*'
$\langle NAME1 \rangle \rightarrow$ 'Michael' | 'Ashley' | 'Daniel' | 'John' | 'Brandon' | 'William' | 'Nicole' | 'Eric' | 'Melissa' | 'Timothy'
$\langle NAME2 \rangle \rightarrow$ 'Christopher' | 'Jennifer' | 'David'
$\langle NAME3 \rangle \rightarrow$ 'Jessica' | 'Joshua' | 'James'
$\langle NAME4 \rangle \rightarrow$ 'Matthew' | 'you'
$\langle PREAMBLE \rangle \rightarrow$ 'I know'
$\langle F \rangle \rightarrow$ 'who' $\langle SUBJ \rangle \ \langle NAME1 \rangle \ \langle ADV \rangle \ \langle V1 \rangle$
$\langle XF \rangle \rightarrow$ 'that' $\langle SUBJ \rangle \ \langle NAME1 \rangle \ \langle ADV \rangle \ \langle V1 \rangle \ \langle NAME3 \rangle$
$\langle G \rangle \rightarrow \langle V2 \rangle \ \langle ADJUNCT \rangle$
$\langle XG \rangle \rightarrow \langle V2 \rangle \ \langle NAME4 \rangle \ \langle ADJUNCT \rangle$
$\langle SUBJ \rangle \rightarrow$ 'the' $\langle SUBJ1 \rangle$ 'that'
$\langle SUBJ1 \rangle \rightarrow$ 'fact' | 'idea' | 'rumor'
$\langle ADV \rangle \rightarrow$ 'secretly' | 'really' | 'absolutely' | 'actually' |
$\langle V1 \rangle \rightarrow$ 'liked' | 'loved' | 'hated' | 'fancied'
$\langle V2 \rangle \rightarrow$ 'surprised' | 'shocked' | 'irritated'
$\langle ADJUNCT \rangle \rightarrow$ 'today' | 'yesterday' | 'recently'

A.1.4 Grammar 4

$\langle S \rangle \rightarrow \langle S\_FG \rangle$
$\langle S\_FG \rangle \rightarrow \langle PREAMBLE \rangle \ \langle F \rangle \ \langle G \rangle$
$\langle S\_XG \rangle \rightarrow \langle UNGRAMMATICAL \rangle \ \langle PREAMBLE \rangle \ \langle XF \rangle \ \langle G \rangle$
$\langle S\_FX \rangle \rightarrow \langle UNGRAMMATICAL \rangle \ \langle PREAMBLE \rangle \ \langle F \rangle \ \langle XG \rangle$
$\langle S\_XX \rangle \rightarrow \langle PREAMBLE \rangle \ \langle XF \rangle \ \langle XG \rangle$
$\langle UNGRAMMATICAL \rangle \rightarrow$ '*'
$\langle NAME1 \rangle \rightarrow$ 'Michael' | 'Ashley' | 'Daniel' | 'John' | 'Brandon' | 'William' | 'Nicole' | 'Eric' | 'Melissa' | 'Timothy'
$\langle NAME2 \rangle \rightarrow$ 'Christopher' | 'Jennifer' | 'David'
$\langle NAME3 \rangle \rightarrow$ 'Jessica' | 'Joshua' | 'James'
$\langle NAME4 \rangle \rightarrow$ 'Matthew' | 'you'
$\langle PREAMBLE \rangle \rightarrow$ 'I know'
$\langle F \rangle \rightarrow$ 'what the' $\langle SUBJ \rangle$ 'to' $\langle V1 \rangle$

$\langle XF\rangle \rightarrow$ 'that the' $\langle SUBJ\rangle$ 'to' $\langle V1\rangle$ $\langle OBJ1\rangle$
$\langle G\rangle \rightarrow \langle V3\rangle$ $\langle V2\rangle$ $\langle ADJUNCT\rangle$
$\langle XG\rangle \rightarrow \langle V3\rangle$ $\langle V2\rangle$ $\langle OBJ2\rangle$ $\langle ADJUNCT\rangle$
$\langle SUBJ\rangle \rightarrow$ 'political campaign' | 'recommendation' | 'legislation' | 'suggestion'
$\langle V1\rangle \rightarrow$ 'preserve' | 'help' | 'save'
$\langle OBJ1\rangle \rightarrow$ 'nature' | 'the environment' | 'the rain forests' | 'biodiversity'
$\langle V3\rangle \rightarrow$ 'made people' | 'caused people to'
$\langle V2\rangle \rightarrow$ 'harm' | 'hurt'
$\langle OBJ2\rangle \rightarrow$ 'animals' | 'wildlife' | 'plants' | 'trees'
$\langle ADJUNCT\rangle \rightarrow$ 'nevertheless' | 'nonetheless'

A.2 ATB

A.2.1 Grammar 1

$\langle S\rangle \rightarrow \langle S\_FG\rangle$
$\langle S\_FG\rangle \rightarrow \langle PREAMBLE\rangle$ $\langle F\rangle$ $\langle G\rangle$
$\langle S\_XG\rangle \rightarrow \langle UNGRAMMATICAL\rangle$ $\langle PREAMBLE\rangle$ $\langle XF\rangle$ $\langle G\rangle$
$\langle S\_FX\rangle \rightarrow \langle UNGRAMMATICAL\rangle$ $\langle PREAMBLE\rangle$ $\langle F\rangle$ $\langle XG\rangle$
$\langle S\_XX\rangle \rightarrow \langle PREAMBLE\rangle$ $\langle XF\rangle$ $\langle XG\rangle$
$\langle UNGRAMMATICAL\rangle \rightarrow$ '*'
$\langle NAME1\rangle \rightarrow$ 'Michael' | 'Ashley' | 'Daniel' | 'John' | 'Brandon' | 'William' | 'Nicole' | 'Eric' | 'Melissa' | 'Timothy'
$\langle NAME2\rangle \rightarrow$ 'Christopher' | 'Jennifer' | 'David'
$\langle NAME3\rangle \rightarrow$ 'Jessica' | 'Joshua' | 'James'
$\langle NAME4\rangle \rightarrow$ 'Matthew' | 'you'
$\langle PREAMBLE\rangle \rightarrow$ 'I know'
$\langle F\rangle \rightarrow$ 'what' $\langle NAME1\rangle$ $\langle V1\rangle$
$\langle XF\rangle \rightarrow$ 'that' $\langle NAME1\rangle$ $\langle V1\rangle$ $\langle OBJ1\rangle$
$\langle CONN\rangle \rightarrow$ 'yesterday and will'
$\langle G\rangle \rightarrow \langle CONN\rangle$ $\langle V2\rangle$ $\langle ADJUNCT\rangle$
$\langle XG\rangle \rightarrow \langle CONN\rangle$ $\langle V2\rangle$ $\langle OBJ2\rangle$ $\langle ADJUNCT\rangle$
$\langle V1\rangle \rightarrow$ 'looked for' | 'searched everywhere for' | 'found' | 'bought' | 'purchased' | 'went shopping for'
$\langle OBJ1\rangle \rightarrow$ 'food' | 'bread' | 'meat' | 'cheese' | 'candy'
$\langle V2\rangle \rightarrow$ 'devour' | 'serve' | 'donate' | 'distribute'
$\langle OBJ2\rangle \rightarrow$ 'it' | 'fish' | 'snacks'
$\langle ADJUNCT\rangle \rightarrow$ 'tomorrow' | 'soon' | 'tonight' | 'today' | 'shortly' | 'quickly'

A.2.2 Grammar 2

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

$\langle UNGRAMMATICAL\rangle \rightarrow$ ‘*’
$\langle NAME1\rangle \rightarrow$ ‘Michael’ | ‘Ashley’ | ‘Daniel’ | ‘John’ | ‘Brandon’ | ‘William’ | ‘Nicole’ | ‘Eric’ | ‘Melissa’ | ‘Timothy’
$\langle NAME2\rangle \rightarrow$ ‘Christopher’ | ‘Jennifer’ | ‘David’
$\langle NAME3\rangle \rightarrow$ ‘Jessica’ | ‘Joshua’ | ‘James’
$\langle NAME4\rangle \rightarrow$ ‘Matthew’ | ‘you’
$\langle PREAMBLE\rangle \rightarrow$ ‘I know’
$\langle F\rangle \rightarrow$ ‘who’ $\langle NAME1\rangle$ $\langle V1\rangle$
$\langle XF\rangle \rightarrow$ ‘that’ $\langle NAME1\rangle$ $\langle V1\rangle$ $\langle NAME2\rangle$
$\langle CONN\rangle \rightarrow$ ‘last year and’
$\langle G\rangle \rightarrow \langle CONN\rangle$ $\langle V2\rangle$ $\langle ADJUNCT\rangle$
$\langle XG\rangle \rightarrow \langle CONN\rangle$ $\langle V2\rangle$ $\langle NAME4\rangle$ $\langle ADJUNCT\rangle$
$\langle V1\rangle \rightarrow$ ‘talked to’ | ‘called’ | ‘texted’ | ‘yelled at’ | ‘humiliated’
$\langle V2\rangle \rightarrow$ ‘argued with’ | ‘had a fight with’ | ‘made peace with’ | ‘stopped talking to’ | ‘fell in love with’ | ‘started to like’
$\langle ADJUNCT\rangle \rightarrow$ ‘today’ | ‘recently’ | ‘lately’

A.2.3 Grammar 3

$\langle S\rangle \rightarrow \langle S\_FG\rangle$
$\langle S\_FG\rangle \rightarrow \langle PREAMBLE\rangle$ $\langle F\rangle$ $\langle G\rangle$
$\langle S\_XG\rangle \rightarrow \langle UNGRAMMATICAL\rangle$ $\langle PREAMBLE\rangle$ $\langle XF\rangle$ $\langle G\rangle$
$\langle S\_FX\rangle \rightarrow \langle UNGRAMMATICAL\rangle$ $\langle PREAMBLE\rangle$ $\langle F\rangle$ $\langle XG\rangle$
$\langle S\_XX\rangle \rightarrow \langle PREAMBLE\rangle$ $\langle XF\rangle$ $\langle XG\rangle$
$\langle UNGRAMMATICAL\rangle \rightarrow$ ‘*’
$\langle NAME1\rangle \rightarrow$ ‘Michael’ | ‘Ashley’ | ‘Daniel’ | ‘John’ | ‘Brandon’ | ‘William’ | ‘Nicole’ | ‘Eric’ | ‘Melissa’ | ‘Timothy’
$\langle NAME2\rangle \rightarrow$ ‘Christopher’ | ‘Jennifer’ | ‘David’
$\langle NAME3\rangle \rightarrow$ ‘Jessica’ | ‘Joshua’ | ‘James’
$\langle NAME4\rangle \rightarrow$ ‘Matthew’ | ‘you’
$\langle PREAMBLE\rangle \rightarrow$ ‘I know’
$\langle F\rangle \rightarrow$ ‘who’ $\langle NAME1\rangle$ $\langle V1\rangle$
$\langle XF\rangle \rightarrow$ ‘that’ $\langle NAME1\rangle$ $\langle V1\rangle$ $\langle NAME2\rangle$
$\langle CONN\rangle \rightarrow$ ‘and’
$\langle G\rangle \rightarrow \langle CONN\rangle$ $\langle V2\rangle$ $\langle ADJUNCT\rangle$
$\langle XG\rangle \rightarrow \langle CONN\rangle$ $\langle V2\rangle$ $\langle NAME4\rangle$ $\langle ADJUNCT\rangle$
$\langle V1\rangle \rightarrow$ ‘saw’ | ‘spotted’ | ‘noticed’ | ‘looked at’
$\langle V2\rangle \rightarrow$ ‘helped’ | ‘played with’ | ‘started to like’ | ‘fell in love with’
$\langle ADJUNCT\rangle \rightarrow$ ‘today’ | ‘yesterday’ | ‘recently’ | ‘lately’

B Appendix: model failures

Worst 10 four-tuples of sentences per phenomenon (PG, ATB), per model (GRNN, JRNN, GPT2, GPT3).

B.1 PG – GRNN

(10)

	$+gap$	$-gap$
$+filler$	I know who the rumor that Mary really fancied irritated **recently (17.60)**	*I know who the rumor that Mary really fancied irritated **you (8.37)** recently
$-filler$	*I know that the rumor that Mary really fancied Jennifer irritated **recently (13.68)**	I know that the rumor that Mary really fancied Jennifer irritated **you (11.02)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.57$		

(11)

	$+gap$	$-gap$
$+filler$	I know who the rumor that James really fancied irritated **recently (16.84)**	*I know who the rumor that James really fancied irritated **you (8.68)** recently
$-filler$	*I know that the rumor that James really fancied Jennifer irritated **recently (13.11)**	I know that the rumor that James really fancied Jennifer irritated **you (11.28)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.33$		

(12)

	$+gap$	$-gap$
$+filler$	I know who the rumor that Mary really liked irritated **recently (16.73)**	*I know who the rumor that Mary really liked irritated **you (9.18)** recently
$-filler$	*I know that the rumor that Mary really liked Jennifer irritated **recently (13.84)**	I know that the rumor that Mary really liked Jennifer irritated **you (12.60)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.32$		

(13)

	$+gap$	$-gap$
$+filler$	I know who the fact that James actually fancied irritated **recently (17.94)**	*I know who the fact that James actually fancied irritated **you (8.02)** recently
$-filler$	*I know that the fact that James actually fancied Jennifer irritated **recently (13.79)**	I know that the fact that James actually fancied Jennifer irritated **you (10.06)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.18$		

(14)

	$+gap$	$-gap$
$+filler$	I know who the fact that James really fancied irritated **recently (17.39)**	*I know who the fact that James really fancied irritated **you (7.61)** recently
$-filler$	*I know that the fact that James really fancied Jennifer irritated **recently (13.26)**	I know that the fact that James really fancied Jennifer irritated **you (9.54)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.07$		

(15)

	$+gap$	$-gap$
$+filler$	I know who the idea that James actually fancied irritated **recently (17.52)**	*I know who the idea that James actually fancied irritated **you (8.12)** recently
$-filler$	*I know that the idea that James actually fancied Jennifer irritated **recently (13.52)**	I know that the idea that James actually fancied Jennifer irritated **you (10.16)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.03$		

(16)

	$+gap$	$-gap$
$+filler$	I know who the idea that James really fancied irritated **recently (17.19)**	*I know who the idea that James really fancied irritated **you (7.79)** recently
$-filler$	*I know that the idea that James really fancied Jennifer irritated **recently (13.14)**	I know that the idea that James really fancied Jennifer irritated **you (9.72)** recently
$\Delta_{-filler} - \Delta_{+filler} = -5.98$		

(17)

	$+gap$	$-gap$
$+filler$	I know who the fact that Mary really fancied irritated **recently (17.82)**	*I know who the fact that Mary really fancied irritated **you (7.51)** recently
$-filler$	*I know that the fact that Mary really fancied Jennifer irritated **recently (13.99)**	I know that the fact that Mary really fancied Jennifer irritated **you (9.53)** recently
$\Delta_{-filler} - \Delta_{+filler} = -5.85$		

(18)

	$+gap$	$-gap$
$+filler$	I know who the rumor that James actually fancied irritated **recently (17.43)**	*I know who the rumor that James actually fancied irritated **you (9.50)** recently
$-filler$	*I know that the rumor that James actually fancied Jennifer irritated **recently (13.89)**	I know that the rumor that James actually fancied Jennifer irritated **you (11.75)** recently
$\Delta_{-filler} - \Delta_{+filler} = -5.79$		

(19)

	$+gap$	$-gap$
$+filler$	I know who the idea that Mary really fancied irritated **recently (17.61)**	*I know who the idea that Mary really fancied irritated **you (7.45)** recently
$-filler$	*I know that the idea that Mary really fancied Jennifer irritated **recently (14.10)**	I know that the idea that Mary really fancied Jennifer irritated **you (9.70)** recently
$\Delta_{-filler} - \Delta_{+filler} = -5.76$		

B.2 PG – JRNN

(20)

	$+gap$	$-gap$
$+filler$	I know who James's recent attitude towards worried **recently (16.94)**	*I know who James's recent attitude towards worried **Michael (10.85)** recently
$-filler$	*I know that James's recent attitude towards Robert worried **recently (12.69)**	I know that James's recent attitude towards Robert worried **Michael (13.46)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.87$		

(21)

	$+gap$	$-gap$
$+filler$	I know who James's current attitude towards worried **recently (17.31)**	*I know who James's current attitude towards worried **Michael (10.81)** recently
$-filler$	*I know that James's current attitude towards Robert worried **recently (13.50)**	I know that James's current attitude towards Robert worried **Michael (13.41)** recently
$\Delta_{-filler} - \Delta_{+filler} = -6.41$		

(22)

	$+gap$	$-gap$
$+filler$	I know who James's recent attitude towards worried **yesterday (16.01)**	*I know who James's recent attitude towards worried **Michael (10.85)** yesterday
$-filler$	*I know that James's recent attitude towards Robert worried **yesterday (12.47)**	I know that James's recent attitude towards Robert worried **Michael (13.46)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -6.16$		

(23)

	$+gap$	$-gap$
$+filler$	I know who James's current praising of worried **yesterday (15.75)**	*I know who James's current praising of worried **Michael (9.60)** yesterday
$-filler$	*I know that James's current praising of Robert worried **yesterday (12.79)**	I know that James's current praising of Robert worried **Michael (12.24)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -5.61$		

(24)

	$+gap$	$-gap$
$+filler$	I know who James's recent praising of bothered **yesterday (15.34)**	*I know who James's recent praising of bothered **Michael (7.61)** yesterday
$-filler$	*I know that James's recent praising of Robert bothered **yesterday (12.72)**	I know that James's recent praising of Robert bothered **Michael (10.42)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -5.44$		

(25)

	$+gap$	$-gap$
$+filler$	I know who James's current attitude towards worried **yesterday (15.98)**	*I know who James's current attitude towards worried **Michael (10.81)** yesterday
$-filler$	*I know that James's current attitude towards Robert worried **yesterday (13.19)**	I know that James's current attitude towards Robert worried **Michael (13.41)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -5.39$		

(26)

	$+gap$	$-gap$
$+filler$	I know who James's current praising of bothered **yesterday (14.98)**	*I know who James's current praising of bothered **Michael (8.20)** yesterday
$-filler$	*I know that James's current praising of Robert bothered **yesterday (12.39)**	I know that James's current praising of Robert bothered **Michael (10.83)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -5.22$		

(27)

	$+gap$	$-gap$
$+filler$	I know who James's recent attitude towards annoyed **recently (17.35)**	*I know who James's recent attitude towards annoyed **Michael (9.36)** recently
$-filler$	*I know that James's recent attitude towards Robert annoyed **recently (13.94)**	I know that James's recent attitude towards Robert annoyed **Michael (11.14)** recently
$\Delta_{-filler} - \Delta_{+filler} = -5.19$		

(28)

	$+gap$	$-gap$
$+filler$	I know who Mary's current praising of worried **yesterday (15.81)**	*I know who Mary's current praising of worried **Michael (10.47)** yesterday
$-filler$	*I know that Mary's current praising of Robert worried **yesterday (13.26)**	I know that Mary's current praising of Robert worried **Michael (13.06)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -5.14$		

(29)

	$+gap$	$-gap$
$+filler$	I know who James's recent praising of worried **yesterday (16.03)**	*I know who James's recent praising of worried **Michael (9.74)** yesterday
$-filler$	*I know that James's recent praising of Robert worried **yesterday (12.58)**	I know that James's recent praising of Robert worried **Michael (11.42)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -5.13$		

B.3 PG – GPT2

(30)

	$+gap$	$-gap$
$+filler$	I know who Mary's current talking to worried **yesterday (13.77)**	*I know who Mary's current talking to worried **Michael (13.38)** yesterday
$-filler$	*I know that Mary's current talking to Patricia worried **yesterday (12.63)**	I know that Mary's current talking to Patricia worried **Michael (13.79)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -1.55$		

(31)

	$+gap$	$-gap$
$+filler$	I know who the idea that James actually hated surprised **recently (17.98)**	*I know who the idea that James actually hated surprised **you (4.83)** recently
$-filler$	*I know that the idea that James actually hated Jennifer surprised **recently (18.14)**	I know that the idea that James actually hated Jennifer surprised **you (6.41)** recently
$\Delta_{-filler} - \Delta_{+filler} = -1.42$		

(32)

	$+gap$	$-gap$
$+filler$	I know who the idea that James actually hated surprised **today (15.84)**	*I know who the idea that James actually hated surprised **you (4.83)** today
$-filler$	*I know that the idea that James actually hated John surprised **today (15.73)**	I know that the idea that James actually hated John surprised **you (6.12)** today
$\Delta_{-filler} - \Delta_{+filler} = -1.39$		

(33)

	$+gap$	$-gap$
$+filler$	I know who the idea that Mary actually hated surprised **today (15.25)**	*I know who the idea that Mary actually hated surprised **you (4.72)** today
$-filler$	*I know that the idea that Mary actually hated John surprised **today (15.30)**	I know that the idea that Mary actually hated John surprised **you (6.11)** today
$\Delta_{-filler} - \Delta_{+filler} = -1.34$		

(34)

	$+gap$	$-gap$
$+filler$	I know who the idea that Mary actually fancied surprised **today (14.85)**	*I know who the idea that Mary actually fancied surprised **you (5.14)** today
$-filler$	*I know that the idea that Mary actually fancied John surprised **today (15.05)**	I know that the idea that Mary actually fancied John surprised **you (6.66)** today
$\Delta_{-filler} - \Delta_{+filler} = -1.32$		

(35)

	$+gap$	$-gap$
$+filler$	I know who Mary's current talking to worried **recently (12.88)**	*I know who Mary's current talking to worried **Michael (13.38)** recently
$-filler$	*I know that Mary's current talking to Patricia worried **recently (11.98)**	I know that Mary's current talking to Patricia worried **Michael (13.79)** recently
$\Delta_{-filler} - \Delta_{+filler} = -1.31$		

(36)

	$+gap$	$-gap$
$+filler$	I know who the idea that James really hated surprised **today (15.66)**	*I know who the idea that James really hated surprised **you (4.72)** today
$-filler$	*I know that the idea that James really hated John surprised **today (15.67)**	I know that the idea that James really hated John surprised **you (6.04)** today
$\Delta_{-filler} - \Delta_{+filler} = -1.30$		

(37)

	$+gap$	$-gap$
$+filler$	I know who the idea that Mary actually fancied surprised **recently (17.65)**	*I know who the idea that Mary actually fancied surprised **you (5.14)** recently
$-filler$	*I know that the idea that Mary actually fancied Jennifer surprised **recently (18.15)**	I know that the idea that Mary actually fancied Jennifer surprised **you (6.94)** recently
$\Delta_{-filler} - \Delta_{+filler} = -1.30$		

(38)

	$+gap$	$-gap$
$+filler$	I know who the idea that James really hated surprised **recently (17.69)**	*I know who the idea that James really hated surprised **you (4.72)** recently
$-filler$	*I know that the idea that James really hated Jennifer surprised **recently (18.02)**	I know that the idea that James really hated Jennifer surprised **you (6.33)** recently
$\Delta_{-filler} - \Delta_{+filler} = -1.27$		

(39)

	$+gap$	$-gap$
$+filler$	I know who the idea that Mary actually hated surprised **recently (17.89)**	*I know who the idea that Mary actually hated surprised **you (4.72)** recently
$-filler$	*I know that the idea that Mary actually hated Jennifer surprised **recently (18.15)**	I know that the idea that Mary actually hated Jennifer surprised **you (6.25)** recently
$\Delta_{-filler} - \Delta_{+filler} = -1.27$		

B.4 PG – GPT3

(40)

	$+gap$	$-gap$
$+filler$	I know what the legislation to help made people hurt **nonetheless (21.70)**	*I know what the legislation to help made people hurt **animals (7.82)** nonetheless
$-filler$	*I know that the legislation to help the environment made people hurt **nonetheless (16.85)**	I know that the legislation to help the environment made people hurt **animals (11.03)** nonetheless
$\Delta_{-filler} - \Delta_{+filler} = -8.06$		

(41)

	$+gap$	$-gap$
$+filler$	I know what the legislation to help made people hurt **nevertheless (21.44)**	*I know what the legislation to help made people hurt **animals (7.82)** nevertheless
$-filler$	*I know that the legislation to help nature made people hurt **nevertheless (15.21)**	I know that the legislation to help nature made people hurt **animals (8.53)** nevertheless
$\Delta_{-filler} - \Delta_{+filler} = -6.94$		

(42)

	$+gap$	$-gap$
$+filler$	I know what the legislation to help caused people to hurt **nonetheless (16.56)**	*I know what the legislation to help caused people to hurt **animals (4.30)** nonetheless
$-filler$	*I know that the legislation to help the environment caused people to hurt **nonetheless (15.67)**	I know that the legislation to help the environment caused people to hurt **animals (9.60)** nonetheless
$\Delta_{-filler} - \Delta_{+filler} = -6.20$		

(43)

	$+gap$	$-gap$
$+filler$	I know what the legislation to help caused people to hurt **nevertheless (16.43)**	*I know what the legislation to help caused people to hurt **animals (4.30)** nevertheless
$-filler$	*I know that the legislation to help the environment caused people to hurt **nevertheless (16.80)**	I know that the legislation to help the environment caused people to hurt **animals (9.52)** nevertheless
$\Delta_{-filler} - \Delta_{+filler} = -4.85$		

(44)

	$+gap$	$-gap$
$+filler$	I know what the legislation to help made people harm **nonetheless (17.80)**	*I know what the legislation to help made people harm **animals (2.95)** nonetheless
$-filler$	*I know that the legislation to help the environment made people harm **nonetheless (19.87)**	I know that the legislation to help the environment made people harm **animals (9.01)** nonetheless
$\Delta_{-filler} - \Delta_{+filler} = -3.99$		

(45)

	$+gap$	$-gap$
$+filler$	I know who Mary's current attitude towards distracted **yesterday (21.98)**	*I know who Mary's current attitude towards distracted **Michael (20.67)** yesterday
$-filler$	*I know that Mary's current attitude towards Robert distracted **yesterday (14.76)**	I know that Mary's current attitude towards Robert distracted **Michael (16.49)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -3.04$		

(46)

	$+gap$	$-gap$
$+filler$	I know what the legislation to help made people harm **nevertheless (17.67)**	*I know what the legislation to help made people harm **animals (2.95)** nevertheless
$-filler$	*I know that the legislation to help nature made people harm **nevertheless (15.82)**	I know that the legislation to help nature made people harm **animals (3.97)** nevertheless
$\Delta_{-filler} - \Delta_{+filler} = -2.86$		

(47)

	$+gap$	$-gap$
$+filler$	I know who James's current attitude towards distracted **yesterday (21.14)**	*I know who James's current attitude towards distracted **Michael (19.43)** yesterday
$-filler$	*I know that James's current attitude towards Patricia distracted **yesterday (13.05)**	I know that James's current attitude towards Patricia distracted **Michael (14.17)** yesterday
$\Delta_{-filler} - \Delta_{+filler} = -2.84$		

(48)

	$+gap$	$-gap$
$+filler$	I know who Mary's current talking to worried **lately (18.75)**	*I know who Mary's current talking to worried **Michael (17.11)** lately
$-filler$	*I know that Mary's current talking to Patricia worried **lately (16.68)**	I know that Mary's current talking to Patricia worried **Michael (17.81)** lately
$\Delta_{-filler} - \Delta_{+filler} = -2.78$		

(49)

	$+gap$	$-gap$
$+filler$	I know who the rumor that Mary absolutely loved surprised **today (14.95)**	*I know who the rumor that Mary absolutely loved surprised **Michael (14.09)** today
$-filler$	*I know that the rumor that Mary absolutely loved John surprised **today (16.62)**	I know that the rumor that Mary absolutely loved John surprised **Michael (18.19)** today
$\Delta_{-filler} - \Delta_{+filler} = -2.43$		

B.5 ATB – GRNN

(50)

	$+gap$	$-gap$
$+filler$	I know who James humiliated last year and had a fight with **recently (16.93)**	*I know who James humiliated last year and had a fight with **you (7.24)** recently
$-filler$	*I know that James humiliated Robert last year and had a fight with **recently (17.14)**	I know that James humiliated Robert last year and had a fight with **you (11.04)** recently
$\Delta_{-filler} - \Delta_{+filler} = -3.59$		

(51)

	$+gap$	$-gap$
$+filler$	I know who James humiliated last year and made peace with **recently (19.37)**	*I know who James humiliated last year and made peace with **you (7.57)** recently
$-filler$	*I know that James humiliated Robert last year and made peace with **recently (19.18)**	I know that James humiliated Robert last year and made peace with **you (10.74)** recently
$\Delta_{-filler} - \Delta_{+filler} = -3.35$		

(52)

	$+gap$	$-gap$
$+filler$	I know who Mary humiliated last year and had a fight with **recently (16.83)**	*I know who Mary humiliated last year and had a fight with **you (7.38)** recently
$-filler$	*I know that Mary humiliated Robert last year and had a fight with **recently (17.18)**	I know that Mary humiliated Robert last year and had a fight with **you (11.08)** recently
$\Delta_{-filler} - \Delta_{+filler} = -3.35$		

(53)

	$+gap$	$-gap$
$+filler$	I know who Mary humiliated last year and made peace with **recently (19.25)**	*I know who Mary humiliated last year and made peace with **you (7.54)** recently
$-filler$	*I know that Mary humiliated Robert last year and made peace with **recently (19.35)**	I know that Mary humiliated Robert last year and made peace with **you (10.80)** recently
$\Delta_{-filler} - \Delta_{+filler} = -3.15$		

(54)

	$+gap$	$-gap$
$+filler$	I know who Mary humiliated last year and stopped talking to **recently (17.26)**	*I know who Mary humiliated last year and stopped talking to **you (6.54)** recently
$-filler$	*I know that Mary humiliated Robert last year and stopped talking to **recently (17.62)**	I know that Mary humiliated Robert last year and stopped talking to **you (9.97)** recently
$\Delta_{-filler} - \Delta_{+filler} = -3.07$		

(55)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> James humiliated last year and had a fight with **lately (21.30)**	*I know <u>who</u> James humiliated last year and had a fight with **you (7.24)** lately
$-filler$	*I know <u>that</u> James humiliated Robert last year and had a fight with **lately (22.04)**	I know <u>that</u> James humiliated Robert last year and had a fight with **you (11.04)** lately
$\Delta_{-filler} - \Delta_{+filler} = -3.06$		

(56)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> James humiliated last year and stopped talking to **recently (16.98)**	*I know <u>who</u> James humiliated last year and stopped talking to **you (6.57)** recently
$-filler$	*I know <u>that</u> James humiliated Robert last year and stopped talking to **recently (17.24)**	I know <u>that</u> James humiliated Robert last year and stopped talking to **you (9.73)** recently
$\Delta_{-filler} - \Delta_{+filler} = -2.90$		

(57)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> Mary humiliated last year and had a fight with **lately (21.13)**	*I know <u>who</u> Mary humiliated last year and had a fight with **you (7.38)** lately
$-filler$	*I know <u>that</u> Mary humiliated Robert last year and had a fight with **lately (21.95)**	I know <u>that</u> Mary humiliated Robert last year and had a fight with **you (11.08)** lately
$\Delta_{-filler} - \Delta_{+filler} = -2.88$		

(58)

	$+gap$	$-gap$
$+filler$	I know <u>who</u> James humiliated last year and made peace with **lately (21.44)**	*I know <u>who</u> James humiliated last year and made peace with **you (7.57)** lately
$-filler$	*I know <u>that</u> James humiliated Robert last year and made peace with **lately (21.79)**	I know <u>that</u> James humiliated Robert last year and made peace with **you (10.74)** lately
$\Delta_{-filler} - \Delta_{+filler} = -2.81$		

(59)

	$+gap$	$-gap$
$+filler$	I know who Mary yelled at last year and had a fight with **recently (16.67)**	*I know who Mary yelled at last year and had a fight with **you (6.77)** recently
$-filler$	*I know that Mary yelled at Robert last year and had a fight with **recently (16.91)**	I know that Mary yelled at Robert last year and had a fight with **you (9.82)** recently
$\Delta_{-filler} - \Delta_{+filler} = -2.81$		

B.6 ATB – JRNN

(60)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will distribute **shortly (14.32)**	*I know what James looked for yesterday and will distribute **fish (14.48)** shortly
$-filler$	*I know that James looked for candy yesterday and will distribute **shortly (15.35)**	I know that James looked for candy yesterday and will distribute **fish (16.78)** shortly
$\Delta_{-filler} - \Delta_{+filler} = -1.27$		

(61)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will distribute **tonight (11.85)**	*I know what James looked for yesterday and will distribute **fish (14.48)** tonight
$-filler$	*I know that James looked for candy yesterday and will distribute **tonight (13.16)**	I know that James looked for candy yesterday and will distribute **fish (16.78)** tonight
$\Delta_{-filler} - \Delta_{+filler} = -0.98$		

(62)

	$+gap$	$-gap$
$+filler$	I know what Mary looked for yesterday and will distribute **shortly (13.41)**	*I know what Mary looked for yesterday and will distribute **fish (14.75)** shortly
$-filler$	*I know that Mary looked for candy yesterday and will distribute **shortly (14.84)**	I know that Mary looked for candy yesterday and will distribute **fish (17.11)** shortly
$\Delta_{-filler} - \Delta_{+filler} = -0.92$		

(63)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will distribute **soon (12.13)**	*I know what James looked for yesterday and will distribute **fish (14.48)** soon
$-filler$	*I know that James looked for candy yesterday and will distribute **soon (13.65)**	I know that James looked for candy yesterday and will distribute **fish (16.78)** soon
$\Delta_{-filler} - \Delta_{+filler} = -0.77$		

(64)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will donate **shortly (14.25)**	*I know what James looked for yesterday and will donate **fish (15.02)** shortly
$-filler$	*I know that James looked for candy yesterday and will donate **shortly (14.49)**	I know that James looked for candy yesterday and will donate **fish (16.03)** shortly
$\Delta_{-filler} - \Delta_{+filler} = -0.77$		

(65)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will donate **tonight (11.27)**	*I know what James looked for yesterday and will donate **fish (15.02)** tonight
$-filler$	*I know that James looked for candy yesterday and will donate **tonight (11.57)**	I know that James looked for candy yesterday and will donate **fish (16.03)** tonight
$\Delta_{-filler} - \Delta_{+filler} = -0.70$		

(66)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will distribute **today (8.08)**	*I know what James looked for yesterday and will distribute **fish (14.48)** today
$-filler$	*I know that James looked for candy yesterday and will distribute **today (9.71)**	I know that James looked for candy yesterday and will distribute **fish (16.78)** today
$\Delta_{-filler} - \Delta_{+filler} = -0.67$		

(67)

	$+gap$	$-gap$
$+filler$	I know what Mary looked for yesterday and will distribute **soon (11.25)**	*I know what Mary looked for yesterday and will distribute **fish (14.75)** soon
$-filler$	*I know that Mary looked for candy yesterday and will distribute **soon (13.07)**	I know that Mary looked for candy yesterday and will distribute **fish (17.11)** soon
$\Delta_{-filler} - \Delta_{+filler} = -0.54$		

(68)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will donate **quickly (14.93)**	*I know what James looked for yesterday and will donate **fish (15.02)** quickly
$-filler$	*I know that James looked for candy yesterday and will donate **quickly (15.46)**	I know that James looked for candy yesterday and will donate **fish (16.03)** quickly
$\Delta_{-filler} - \Delta_{+filler} = -0.47$		

(69)

	$+gap$	$-gap$
$+filler$	I know what Mary looked for yesterday and will donate **shortly (13.48)**	*I know what Mary looked for yesterday and will donate **fish (15.36)** shortly
$-filler$	*I know that Mary looked for candy yesterday and will donate **shortly (14.18)**	I know that Mary looked for candy yesterday and will donate **fish (16.53)** shortly
$\Delta_{-filler} - \Delta_{+filler} = -0.47$		

B.7 ATB – GPT2

(70)

	$+gap$	$-gap$
$+filler$	I know what Mary looked for yesterday and will devour **soon (12.60)**	*I know what Mary looked for yesterday and will devour **snacks (15.87)** soon
$-filler$	*I know that Mary looked for cheese yesterday and will devour **soon (10.28)**	I know that Mary looked for cheese yesterday and will devour **snacks (17.21)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.66$		

(71)

	$+gap$	$-gap$
$+filler$	I know what Mary went shopping for yesterday and will devour **soon (14.43)**	*I know what Mary went shopping for yesterday and will devour **snacks (13.26)** soon
$-filler$	*I know that Mary went shopping for cheese yesterday and will devour **soon (12.70)**	I know that Mary went shopping for cheese yesterday and will devour **snacks (15.01)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.48$		

(72)

	$+gap$	$-gap$
$+filler$	I know what James looked for yesterday and will devour **soon (12.95)**	*I know what James looked for yesterday and will devour **snacks (16.07)** soon
$-filler$	*I know that James looked for cheese yesterday and will devour **soon (10.68)**	I know that James looked for cheese yesterday and will devour **snacks (16.79)** soon
$\Delta_{-filler} - \Delta_{+filler} = -3.00$		

(73)

	$+gap$	$-gap$
$+filler$	I know what Mary found yesterday and will devour **soon (14.09)**	*I know what Mary found yesterday and will devour **snacks (16.73)** soon
$-filler$	*I know that Mary found cheese yesterday and will devour **soon (12.44)**	I know that Mary found cheese yesterday and will devour **snacks (17.88)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.80$		

(74)

	$+gap$	$-gap$
$+filler$	I know what Mary went shopping for yesterday and will devour **shortly (16.70)**	*I know what Mary went shopping for yesterday and will devour **fish (15.86)** shortly
$-filler$	*I know that Mary went shopping for candy yesterday and will devour **shortly (15.57)**	I know that Mary went shopping for candy yesterday and will devour **fish (17.51)** shortly
$\Delta_{-filler} - \Delta_{+filler} = -2.78$		

(75)

	$+gap$	$-gap$
$+filler$	I know what James went shopping for yesterday and will devour **soon (14.52)**	*I know what James went shopping for yesterday and will devour **fish (16.14)** soon
$-filler$	*I know that James went shopping for candy yesterday and will devour **soon (13.05)**	I know that James went shopping for candy yesterday and will devour **fish (17.32)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.65$		

(76)

	$+gap$	$-gap$
$+filler$	I know what Mary looked for yesterday and will devour **quickly (15.56)**	*I know what Mary looked for yesterday and will devour **snacks (15.87)** quickly
$-filler$	*I know that Mary looked for cheese yesterday and will devour **quickly (14.41)**	I know that Mary looked for cheese yesterday and will devour **snacks (17.21)** quickly
$\Delta_{-filler} - \Delta_{+filler} = -2.49$		

(77)

	$+gap$	$-gap$
$+filler$	I know what Mary went shopping for yesterday and will devour **quickly (15.84)**	*I know what Mary went shopping for yesterday and will devour **snacks (13.26)** quickly
$-filler$	*I know <u>that</u> Mary went shopping for cheese yesterday and will devour **quickly (15.20)**	I know <u>that</u> Mary went shopping for cheese yesterday and will devour **snacks (15.01)** quickly
$\Delta_{-filler} - \Delta_{+filler} = -2.38$		

(78)

	$+gap$	$-gap$
$+filler$	I know what Mary looked for yesterday and will serve **soon (12.79)**	*I know what Mary looked for yesterday and will serve **snacks (14.84)** soon
$-filler$	*I know <u>that</u> Mary looked for cheese yesterday and will serve **soon (12.26)**	I know <u>that</u> Mary looked for cheese yesterday and will serve **snacks (16.61)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.30$		

(79)

	$+gap$	$-gap$
$+filler$	I know what Mary searched everywhere for yesterday and will devour **soon (12.69)**	*I know what Mary searched everywhere for yesterday and will devour **fish (16.47)** soon
$-filler$	*I know <u>that</u> Mary searched everywhere for candy yesterday and will devour **soon (11.03)**	I know <u>that</u> Mary searched everywhere for candy yesterday and will devour **fish (17.11)** soon
$\Delta_{-filler} - \Delta_{+filler} = -2.29$		

B.8 ATB – GPT3

(80)

	$+gap$	$-gap$
$+filler$	I know what James went shopping for yesterday and will serve **tonight (7.83)**	*I know what James went shopping for yesterday and will serve **snacks (10.67)** tonight
$-filler$	*I know <u>that</u> James went shopping for meat yesterday and will serve **tonight (9.19)**	I know <u>that</u> James went shopping for meat yesterday and will serve **snacks (21.53)** tonight
$\Delta_{-filler} - \Delta_{+filler} = -9.50$		

(81)

	$+gap$	$-gap$
$+filler$	I know what Mary went shopping for yesterday and will serve **tonight (6.79)**	*I know what Mary went shopping for yesterday and will serve **snacks (11.03)** tonight
$-filler$	*I know that Mary went shopping for meat yesterday and will serve **tonight (10.14)**	I know that Mary went shopping for meat yesterday and will serve **snacks (22.76)** tonight
$\Delta_{-filler} - \Delta_{+filler} = -8.38$		

(82)

	$+gap$	$-gap$
$+filler$	I know what Mary went shopping for yesterday and will distribute **today (10.08)**	*I know what Mary went shopping for yesterday and will distribute **snacks (11.31)** today
$-filler$	*I know that Mary went shopping for meat yesterday and will distribute **today (13.55)**	I know that Mary went shopping for meat yesterday and will distribute **snacks (22.82)** today
$\Delta_{-filler} - \Delta_{+filler} = -8.04$		

(83)

	$+gap$	$-gap$
$+filler$	I know what Mary went shopping for yesterday and will serve **today (8.36)**	*I know what Mary went shopping for yesterday and will serve **snacks (11.03)** today
$-filler$	*I know that Mary went shopping for meat yesterday and will serve **today (12.42)**	I know that Mary went shopping for meat yesterday and will serve **snacks (22.76)** today
$\Delta_{-filler} - \Delta_{+filler} = -7.67$		

(84)

	$+gap$	$-gap$
$+filler$	I know what James searched everywhere for yesterday and will donate **today (6.18)**	*I know what James searched everywhere for yesterday and will donate **it (1.78)** today
$-filler$	*I know that James searched everywhere for food yesterday and will donate **today (5.31)**	I know that James searched everywhere for food yesterday and will donate **it (8.48)** today
$\Delta_{-filler} - \Delta_{+filler} = -7.57$		

(85)

	$+gap$	$-gap$
$+filler$	I know what James went shopping for yesterday and will serve **soon (14.07)**	*I know what James went shopping for yesterday and will serve **snacks (10.67)** soon
$-filler$	*I know that James went shopping for meat yesterday and will serve **soon (17.37)**	I know that James went shopping for meat yesterday and will serve **snacks (21.53)** soon
$\Delta_{-filler} - \Delta_{+filler} = -7.56$		

(86)

	$+gap$	$-gap$
$+filler$	I know what James went shopping for yesterday and will serve **today (9.70)**	*I know what James went shopping for yesterday and will serve **snacks (10.67)** today
$-filler$	*I know that James went shopping for meat yesterday and will serve **today (13.12)**	I know that James went shopping for meat yesterday and will serve **snacks (21.48)** today
$\Delta_{-filler} - \Delta_{+filler} = -7.40$		

(87)

	$+gap$	$-gap$
$+filler$	I know what James went shopping for yesterday and will serve **shortly (13.84)**	*I know what James went shopping for yesterday and will serve **snacks (10.72)** shortly
$-filler$	*I know that James went shopping for meat yesterday and will serve **shortly (17.31)**	I know that James went shopping for meat yesterday and will serve **snacks (21.48)** shortly
$\Delta_{-filler} - \Delta_{+filler} = -7.29$		

(88)

	$+gap$	$-gap$
$+filler$	I know what Mary went shopping for yesterday and will distribute **tonight (11.46)**	*I know what Mary went shopping for yesterday and will distribute **snacks (11.45)** tonight
$-filler$	*I know that Mary went shopping for meat yesterday and will distribute **tonight (15.77)**	I know that Mary went shopping for meat yesterday and will distribute **snacks (22.82)** tonight
$\Delta_{-filler} - \Delta_{+filler} = -7.06$		

(89)

	$+gap$	$-gap$
$+filler$	I know what James bought yesterday and will serve **soon (12.75)**	*I know what James bought yesterday and will serve **fish (9.61)** soon
$-filler$	*I know <u>that</u> James bought candy yesterday and will serve **soon (19.06)**	I know <u>that</u> James bought candy yesterday and will serve **fish (22.58)** soon
$\Delta_{-filler} - \Delta_{+filler} = -6.66$		