Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs

Anonymous ACL submission

Abstract

Language models learn rare syntactic phenomena, but it has been argued that they rely on rote memorization, as opposed to grammatical generalization. Training on a corpus of human-scale in size (100M words), we iteratively trained transformer language models on systematically manipulated corpora and then evaluated their learning of a particular rare grammatical phenomenon: the English Article+Adjective+Numeral+Noun (AANN) construction ("a beautiful five days"). We compared how well this construction was learned on the default corpus relative to a counterfactual corpus in which the AANN sentences were removed. AANNs were still learned better than systematically perturbed variants of the construction. Using additional counterfactual corpora, we suggest that this learning occurs through generalization from related constructions (e.g., "a few days"). An additional experiment showed that this learning is enhanced when there is more variability in the input. Taken together, our results provide an existence proof that models can learn rare grammatical phenomena by generalization from less rare phenomena. Code will be available at (url).

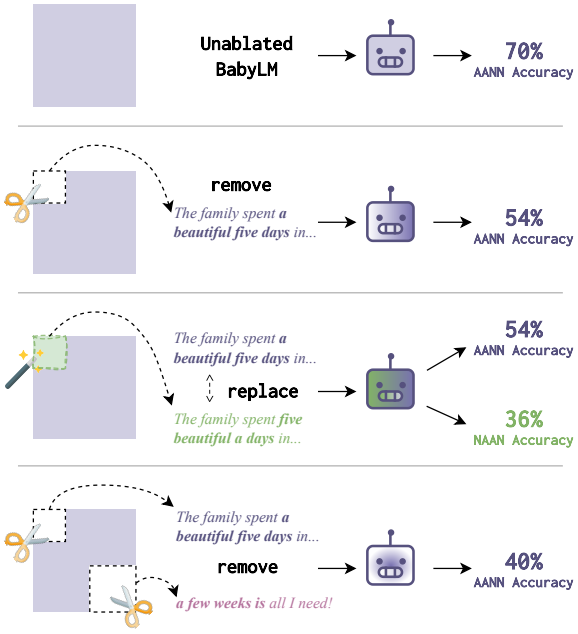


Figure 1: Some of our key experimental manipulations and resulting performance changes. We train LMs on varied input corpora and measure the learning of the *rare* AANN ("a beautiful five days") or related constructions. E.g., we train on the default BabyLM corpus, a corpus in which we remove all AANNs, a corpus in which we replace all AANNs with a corrupted version ("beautiful a five days") and measure learning of the corrupted version, and a corpus in which we remove AANNs and also remove related constructions like "a few weeks is".

1 Introduction

1.1 Motivation and Prior Work

Humans come to learn and use rare grammatical structures, even if they have encountered those structures only rarely or even not at all (Pullum and Scholz, 2002; Pearl, 2022). For instance, humans accept the grammaticality of the PiPP construction ("surprising though it may be...") even where the preposed element crosses a finite close boundary ("surprising though I know it may be that...") (Pullum, 2017) and even though they may plausibly have *never* encountered such a sentence in their linguistic experience (see Potts, 2023, for corpus estimate). How people come to know an utterance is grammatical has occupied a central place in linguistics. Specifically, mastery of never-before-encountered grammatical structures has been taken to mean that people are endowed with innate linguistic knowledge (Chomsky, 1986, 1957, 1965).

Recent evidence, though, suggests that Large Language Models (LLMs) can learn complex grammar (Wilcox et al., 2018; Futrell et al., 2019; Linzen et al., 2016; Mahowald et al.) even from human-scale amounts of input (Warstadt et al., 2023; Eldan and Li, 2023; Huebner et al., 2021). This raises the possibility that input data, along

with an appropriately sophisticated or weakly biased statistical learning mechanism, is sufficient for learning rare constructions by allowing for models to emergently learn appropriate grammatical abstraction (Baroni, 2022; Piantadosi, 2023; Misra and Kim, 2023).

But modern LLMs often have access to much more training input than people do and thus might memorize in a way that humans can't (Warstadt and Bowman, 2022; Warstadt et al., 2023). The possibility that LLMs are "stochastic parrots" (Bender et al., 2021), heavily reliant on memorization, is a common criticism of using LLMs to study human language (e.g., Chomsky et al., 2023).

There are different levels of memorization, though, requiring different levels of abstraction. Consider the AANN construction: "a beautiful five days in Florida" (Solt, 2007; Keenan, 2013; Dalrymple and King, 2019), which is rarer than the default "five beautiful days in Florida". A model that strictly memorizes this phrase might come to know that "a beautiful five days in Florida" is grammatical but have no idea that "a beautiful four days in Florida" is grammatical if it never appeared in its training. A model that generalizes just a bit more might know that "a beautiful five days in New York" is also grammatical by generalizing that any U.S. state can fill the slot. Knowing that "an astonishing 200 pages" is acceptable requires generalization beyond mere lexical items. And knowing that "a blue five pencils" is *not* acceptable (because colors are "stubbornly distributive", Schwarzschild 2011) requires yet more knowledge. Even for an idealized learner, it is difficult to precisely formulate how these kinds of generlizations emerge.

There is increasing evidence that LLMs can generate novel linguistic utterances (McCoy et al., 2023), and also make subtle judgments on relatively rare constructions like these (Weissweiler et al., 2022; Potts, 2023), including the AANN (Mahowald, 2023). If they do so by memorizing examples *verbatim* from an unrealistically large training corpus, that would not be particularly informative for human processing. But, if they do learn rare grammatical phenomena from smaller amounts of data and can generalize beyond just those verbatim instances, that would raise the question of how they do it and if it can inform theorizing about humans. For instance, in the context of the PiPP construction, Potts (2023) speculates that the comparative construction (e.g., "They are happier than we are.") "may be the key to all of this [i.e., learning the PiPP]" because such constructions are "incredibly common" yet share abstract structure with the PiPP. If LLMs learn rare grammatical structures in part by learning and generalizing structures from much more common constructions, that would be powerful evidence for abstraction in LLMs and raise the possibility that even quite general learners could learn very rare phenomena without strong innate priors, drawing in part on the long-posited linguistic hypothesis that apparently distinct grammatical phenomena often share underlying structure.

To that end, our goal in this paper is to study a relatively rare grammatical phenomenon in LMs trained on controlled input corpora that are (a) of human realistic scale, and (b) systematically manipulated with respect to the target constructions as well as key related constructions. Our hypothesis is that **generalization abilities of LMs on such rare phenomena come from abstractions and structures learned from more frequent phenomena**—and that knowledge of more frequent phenomena *are* the "keys to all of this."

As a case study, we focus on the aforementioned AANN construction, although we highlight how the methods used here could serve as a blueprint for work on other rare grammatical phenomena. Our method is to train different instantiations of a standard transformer model on the 100M-word BabyLM corpus, which we systematically manipulate—via removal and replacement—to shed light on the extent to which frequent and related phenomena encountered during training can give rise to abstractions that allow for generalizations in LMs. To test for generalization, we subjected our different LMs to a series of acceptability tests on sentences which do not appear in the training corpus and which specifically target the special properties of the AANN. This approach of training on a systematically manipulated corpus has been fruitfully used to debias models (Lu et al., 2020; Maudslay et al., 2019), study whether LMs rely on lexical knowledge in learning syntactic rules (Wei et al., 2021), explore the effect of permuting words on pretrained models (Sinha et al., 2021), and test whether LMs can learn languages judged to be hard for humans (Kallini et al., 2024), among others.

1.2 Summary of findings

First, we find BabyLM-trained LMs to successfully generalize to novel instances of the AANN construction. Performance unsurprisingly drops for LMs that were trained *without* being exposed to even a

single AANN during training, but perhaps surprisingly, not by all that much—they are substantially above chance. This suggests that certain items present in the training data might give rise to LMs' non-trivial performance in judging acceptability of AANNs. This finding is further strengthened by the fact that LMs trained on counterfactual variants of the AANN—e.g., ANAN and NAAN, obtained by shuffling word order and are far less likely to share structure with training data items—are unable to generalize to those constructions as well as they do to AANNs.

Next, we investigated what might enable LMs' learning of the AANN, by further systematically manipulating their training data to hold out utterances conforming to specific linguistic and statistical phenomena. Through our manipulations, we find LMs become worse at predicting novel instances of the AANN as more frequent, non-AANN-but-AANN-related phenomena are held out. For example, phenomena such as the treatment of measure noun phrases as singular (e.g., ***a few days is*** *all we need*)—similar to how AANNs treat a plural NP as singular—end up making unseen AANNs less likely by 31% on average. Importantly, these results could not be explained simply by loss of data—LMs that were trained with these phenomena left in but without an equivalently large chunk of the training data removed were almost as good as LMs that never saw AANNs. This further strengthens the conclusion that the hypothesized linguistic phenomena did indeed affect generalization of the targeted construction. Notably, LMs are largely affected by these manipulations when they do not see *any* AANNs during training, highlighting the expected non-trivial role of encountering some instances of AANNs to show stronger generalization.

Finally, we characterized the aforementioned interplay between the properties of the encountered AANNs and the LMs generalizations on novel instances. Here we found LMs that observed AANNs with more variability on the adjective, numeral, and noun slots to show better generalization than did LMs that saw more restricted-but-repeating instances of AANNs. This importantly mimicked analogous findings of inductive inference in humans across linguistics (Goldberg, 1995, 2005; Suttle and Goldberg, 2011; Baayen, 2009; O'Donnell, 2015) and cognitive psychology studies (Osherson et al., 1990; Xu and Tenenbaum, 2007).

Taken together, these results provide an existence proof that a weakly biased but sophisticated general-purpose statistical learner can learn rare and complex linguistic phenomena, in part because of key related phenomena seen during training. While our analyses suggest potential links between "constructions" (Goldberg, 1995), our findings are also compatible with theories that think of rare phenomena as derivationally related (Chomsky, 1965) to more frequent and well-attested structures (much as Potts (2023) posits shared syntactic structure as the key to the PiPP).

2 General Methods

Here we describe methods that we use to characterize learning of the targeted rare construction, the AANN. As mentioned in §1, we take these methods as a general blueprint for studying other grammatical phenomena.

2.1 Data and Model

Corpus Throughout the paper, we use the BabyLM-strict corpus (Warstadt et al., 2023) as our base training set, often with systematic ablations. We use BabyLM-strict because of its human-realistic scale and tractable size (100M tokens), which allows us to (1) detect and control the instances of the target construction as well as related linguistic phenomena; and (2) train a large number of LMs in a reasonable timeframe.

Construction Detection We use regexes over a POS-tagged version of the BabyLM training data to detect AANN instances.[1] We detect 2,301 AANNs in the BabyLM corpus, **occurring in about 0.02% of all utterances**.

Model Our LMs are instances of OPT LM (Zhang et al., 2022), an autoregressive transformer architecture. Our LMs have 12 layers and attention heads, use a vocabulary size of 16,384, and are trained for a maximum of 20 epochs using the `transformers` library (Wolf et al., 2020). The results we report for a given LM are averaged over three different runs (with different random seeds). We list other hyper-parameters and architectural details in Appendix B.

2.2 Acceptability data

To test our LMs on their knowledge of the AANN, we use data from Mahowald (2023), which consists of 12,960 templatically generated sentences that

[1] We use `spaCy` for getting POS-tags, see Appendix C for the exact details of the detection pipeline.

contain AANNs. Out of these, 3,420 contain acceptability ratings provided by 190 human participants, ranging from 1 (unacceptable) to 10 (acceptable). We use 7 as the threshold for clear acceptability, in that we only keep instances where human participants rated the acceptability of the construction in context to be greater than 7. We additionally discarded instances where the AANNs appear in the BabyLM training set ($n = 4$), as testing on these would not shed light on the LMs' generalization behavior. This leaves us with 2,027 items.

For each AANN instance in the dataset, Mahowald (2023) has also made available its corresponding corrupted versions, which focus on (1) adjective-numeral order; (2) presence of the article; (3) presence of the adjective; and (4) presence of the numeral. A hypothetical example of these corruptions is shown in Table 1 under the "AANN" column. A model that has knowledge of the AANN should find the well-formed instance to be more likely than each of its corrupted versions. Below we describe methods to measure likelihood and assess accuracy on these tests.

2.3 Scoring and Accuracy

We use the Syntactic Log-odds Ratio (SLOR) proposed by Pauls and Klein (2012); Lau et al. (2017), to score sentences in our tests. Given a sentence containing a prefix followed by our target construction $\mathcal{C}$ and an optional suffix, SLOR is computed as the log of the ratio between the probability of the construction given the prefix as estimated by the LM, and that estimated by a unigram model, normalized by the length of the construction. Given a language model m and a unigram estimator u:

$$\text{SLOR}_{\text{prefix}}(\mathcal{C}) = \frac{1}{\mid \mathcal{C} \mid} \log \frac{p_m(\mathcal{C} \mid \text{prefix})}{p_u(\mathcal{C})} \quad (1)$$

Importantly, we train the unigram estimator for a given corpus using the same tokenizer used to train our autoregressive LMs on that corpus. We use SLOR in lieu of the usual normalized log-probability measure, ensuring that the comparison between two models cannot be explained simply by the difference in unigram frequencies due to our manipulations. Log-probabilities were computed using `minicons` (Misra, 2022).

An instance within our test set is considered to be correct *iff* the SLOR values of the well-formed construction is greater than that for *all* four corrupted instances. The accuracy, then, is the proportion of correct instances within the test set. Since this involves four pairwise comparisons, chance performance is 6.25%.

2.4 Ablations

Common to subsequent experiments (§4 and §5) is the fact that we hold out certain parts of the BabyLM corpus—parts that conform to a certain linguistic or statistical hypothesis. This creates a gap between the experience of LMs trained on these ablated versions of the corpus, and that of the LM trained on the full BabyLM data. To circumvent this issue, we up-sample non-hypothesis-conforming utterances in BabyLM after performing our ablations, in a manner such that the LM still encounters the exact same number of tokens.

3 Experiment 1: LMs learn about AANNs without having seen a single instance

LMs learn about AANNs... To investigate the extent to which LMs trained on BabyLM learn the AANN construction, we measure their accuracy on our tests described in §2.2. From Figure 2, we observe that the BabyLM-trained LMs obtain accuracies around 70%, which is substantially above chance. **This suggests that LMs can reasonably acquire generalizations to AANNs from exposure to positive evidence that makes up only 0.02% of their training experience.**

For comparison to larger, state-of-the-art LMs, we test Llama-2-7B (Touvron et al., 2023) and GPT-2 XL (Radford et al., 2019) on the AANNs, and find them to achieve 83% and 78%, respectively.[2] Similarly, as a comparison to shallower LMs, we tested on 2-and 4-gram LMs trained on BabyLM and found both of them to achieve 0% accuracy, eliminating the possibility that the observed results are due to surface-level statistics.

...without having seen a single instance... Given that BabyLM-trained LMs learn the AANN construction, how well would an LM generalize to AANNs *without having seen a single positive instance*? To this end, we compare accuracies in the previous experiment to that obtained by LMs trained on BabyLM with our **2,301 detected AANNs removed** (i.e., NO AANN).

From Figure 2, we find LMs trained with the NO AANN condition to achieve an average accu-

[2] Qualitatively, performance is on these models is worse than data reported in Mahowald (2023), which used GPT-3. Because log-probabilities are not available from GPT-3, a direct comparison is not possible.

Context	AANN	ANAN	NAAN
WELL-FORMED	a whopping ninety LMs	a ninety whopping LMs	ninety whopping a LMs
		Corruptions	
ORDER-SWAP	a ninety whopping LMs	a whopping ninety LMs	whopping ninety a LMs
NO ARTICLE	whopping ninety LMs	ninety whopping LMs	ninety whopping LMs
NO MODIFIER	a ninety LMs	a ninety LMs	ninety a LMs
NO NUMERAL	a whopping LMs	a whopping LMs	whopping a LMs

Table 1: Well-formed and corrupted examples of the AANN construction and its counterfactual versions (ANAN and NAAN). Corruption types are taken from Mahowald (2023).

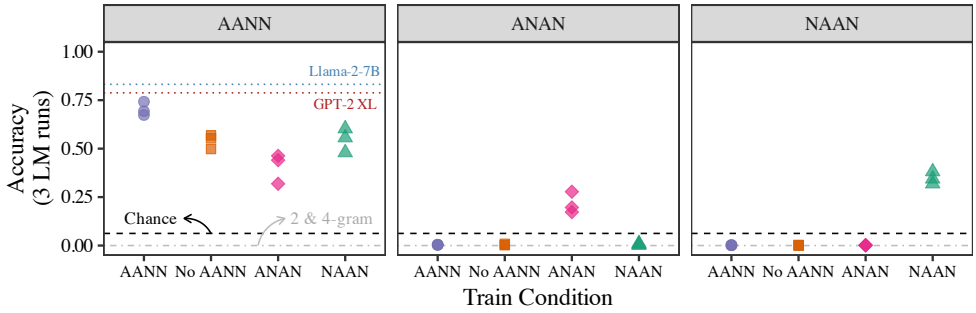


Figure 2: Accuracies on tests for AANN and its counterfactuals (ANAN and NAAN), achieved by LMs trained on BabyLM with various AANN-manipulations (AANN as is, NO AANN, ANAN, NAAN). The dashed line represents chance performance (6.25%), the dot-dashed line represents accuracies for SLORs computed using 2- and 4-gram LMs trained on BabyLM. Accuracies for GPT-2-XL (Radford et al., 2019) and Llama-2-7B (Touvron et al., 2023) are computed using **log-probabilities**, since unigram frequencies were unavailable for these LMs' corpora.

racy of 54%, which is a noticeable drop compared to the 70% obtained by the LMs trained on the complete BabyLM corpus, but importantly 47.75 points above chance performance (and, as we show below, well above comparable baselines with other constructions). This is a non-trivial result, since it suggests that LMs can learn the acceptability of a construction's instances, without having seen a single positive occurrence, which indicates that there exist *some* systematic patterns in the corpus that are driving this generalization behavior.

...more strongly than they learn counterfactual AANN variants To further contextualize the above results, we consider two counterfactual cases, where we replaced AANNs in BabyLMs with instances involving the same lexical items, but in a word order that violates English grammar: (1) ANAN (e.g., *a 90 whopping LMs*); and (2) NAAN (e.g., *90 whopping a pages*). This allows us to test if the results before are genuinely because LMs recognize the nuances of the AANN construction. If LMs are able to learn these counterfactual constructions just as well as the the LMs in the previous experiments learned AANNs, then the generalization claims from the previous experiments would be weakened.

To test for such possibilities, we create counterfactual versions of the Mahowald (2023) stimuli, where we apply analogous corruptions to the counterfactual variants of AANN, such that they are violated in a similar manner as in the AANN tests. Examples for the three types of instances in our tests can be found in Table 1. We then evaluate the previous two LMs (trained on BabyLM with and without seeing any AANNs) with LMs trained on BabyLM with these counterfactual variants on judging the acceptability of AANNs, ANANs, and NAANs. Figure 2 shows these results.

From Figure 2, we make two high-level observations. First, and most importantly, **LMs that see ANANs and NAANs do not learn those constructions as well as they learn AANNs**—especially the LM that saw no AANNs (54% AANN accuracy compared to 36% NAAN accuracy obtained by the NAAN-trained LM). Second, these LMs end up learning AANNs more strongly relative to how well they learn constructions that they observe in lieu of the AANN—e.g., NAAN trained LM achieves around 54% accuracy on AANNs *even though* NAANs *appeared frequently in the data and no* AANNs *did*. This, combined with the results of the previous two sub-experiments suggests strongly that LMs pick up on cues from other—

Phenomenon/Manipulation	Example/`Desc.`	Freq.
AANN	**a fine eighteen months**	2,301
DT ANN	**the usual forty dollars fine**	14,347
A few/couple/dozen/etc. NNS	**a few plums**	55,226
Measure NNS with SG verbs and/or indef. articles	**6 months is a long time**	62,597
A/An + ADJ/NUM balancing	`enforce freq. balance`	571,874
Random removal (**control**)	`randomized ablation`	571,874

Table 2: Manipulated Phenomena, their examples/descriptions, and their frequency in the BabyLM corpus.

likely related—constructions encountered during training in order to assign non-trivial probability to unseen instances of AANNs.

4 Experiment 2: Keys to Learning AANNs

Our previous experiments reveal that, keeping everything else the same, LMs learn the AANN construction far more accurately than they do its counterfactual variants. Furthermore, we also see strong AANN acceptability judgments in LMs that have never encountered a single instance. This suggests that there are instances in the training data that contribute to the learning of the construction.

What might these be? Below we enumerate four hypotheses, each of which tackles subtly different aspects of the AANN construction, then measure the effect of these phenomena, by separately holding them out during training and computing the average SLOR of the well-formed instances of the AANN tests. The effect of a particular phenomenon on the acceptability of AANNs can therefore be measured by comparing SLORs before and after holding out instances of that phenomenon. Methods for detecting the hypothesized phenomena can be found in Appendix C. As control, we additionally also hold out a random set of utterances, which do not conform to the target phenomena of interest. Note again that for all these cases, we ensure the LMs see the same number of tokens during training, by up-sampling other, non-AANN containing sentences. Table 2 lists the hypotheses we consider, along with an example of their utterance and frequency of occurrence, in the BabyLM corpus.

The presence of "the ANN" Phrases like "*the beautiful five days*" are common in corpora, which are not as unusual because "the" regularly takes plural nouns. We hypothesize that the acceptability of these structures affects the acceptability of AANNs, since an LM might analogize from the general observation that 'a' or 'an' can substitute 'the' (e.g., a ball vs. the ball). Therefore, we consider all cases wheere a determiner precedes the contiguous sequence of article, numeral, plural noun.

A few/couple/dozen/etc. NNS Another related phenomenon that is more common relative to the AANN construction involves phrases such as "*a few days*" or "*a couple bottles*". To an LM learner, they might provide evidence that measure noun phrases with plural nouns can be attached to an indefinite article (a/an; Solt, 2007), as is the case in AANNs.

Measure NNS treated as singular We consider yet another phenomenon involving phrases that treat measure nouns as singular, this time in terms of agreement, e.g., "*Five miles **is** a long way to go*", and "*1,000 pages **is** a lot for a dissertation.*" These cases might provide further evidence to the model that measure noun phrases with plural nouns can be treated as a singular unit (Solt, 2007), thereby affecting the acceptability of the AANN. These are less prevalent compared to the cases involving **a few/couple/dozen NNS**, but still far more frequent than the AANN, therefore, we combine the two as a general case of treating measure NPs as singular.

Balancing the frequencies of A/An + ADJ/NUM A more surface-level reason why "a beautiful five days" might be more natural to LMs than is "a five beautiful days", could be that adjectives are more likely to follow indefinite articles than are numerals. For instance, adjectives are ≈14.6 times more likely to follow indefinite articles in the BabyLM corpus than are numerals. To measure this effect, we hold out instances such that adjectives and numerals are equally likely to follow an indefinite article. This ends up being the largest portion of the data that we hold out.

Control: Random removal A potential confound in the above ablations could be that the SLOR values of the AANNs go down merely due to loss of content—this could be despite the fact that we add back additional tokens from BabyLM (such that all LMs see the exact same amount of tokens). To account for this, we additionally consider a control condition where we remove as many tokens as in the largest ablation (i.e., the **A/An + ADJ/NUM** case) such that none of the above phenomena are taken out. Insofar as the LMs rely on the aforementioned phenomena, results on LMs trained with this ablation should be closer to the original BabyLM-trained LM's results.

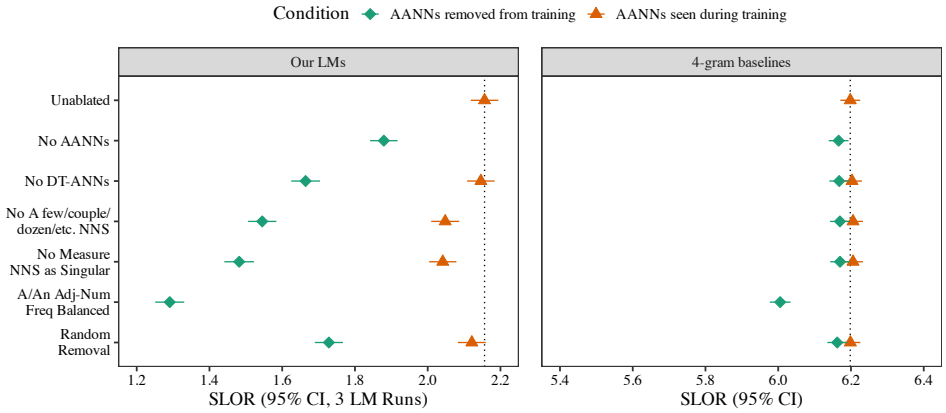


Figure 3: SLORs on AANNs from Mahowald (2023) for our LMs (left) and a 4-gram baseline (right) trained on BabyLM and its ablated versions. SLOR for **unablated BabyLM-trained LM** shown with dotted line.

4.1 Analysis and Results

In our experiments, we individually ablate out each of the aforementioned phenomena under two settings: (1) AANNs are removed during training in addition to the phenomena; and when possible, (2) AANNs are seen during training. (1) is a stricter setting, since here the LMs see neither the target phenomenon nor a single instance of the AANN. Comparing average SLORs obtained in this condition to that obtained for the NO AANN can shed light on the the extent to which the target phenomenon is critical in allowing LMs to assign non-trivial probabilities on unseen AANNs, *zero-shot*. On the other hand, (2) still allows for the LMs to perform lexical generalization (Kim et al., 2022) from seen AANN instances—where they may arrive at strong acceptability measures on the test AANNs by performing slot filling, without necessarily relying on the hypothesized phenomena.

Figure 3 shows the average SLORs obtained across various ablations under the two settings. As a baseline, we compare our results to 4-gram LMs[3] trained on corresponding ablations of the BabyLM corpus. We observe that holding out all our hypothesized phenomena has non-trivial effects on our LMs' ratings of unseen well-formed AANNs, and that these effects are different for when AANNs are seen during training, or are held out. When AANNs are held out along with the phenomena, we see substantial drops in the average SLOR values assigned by LMs on unseen AANNs relative to that assigned by LMs in the NO AANN condition. Specifically, balancing the frequency of adjectives and numerals following an article has the greatest effect, followed by the two cases where measure nouns are treated as singular, followed by the removal of all cases that involve any determiner + adjective + numeral + noun sequence. This suggests that, in the absence of even a single AANN during training, these phenomena are critical for LMs to assign probability to AANNs. Simply ablating large amounts of data cannot explain these results, since LMs trained on our controlled condition show higher SLOR values than in all hypothesis-informed ablations. These patterns are absent in 4-gram LMs, suggesting that they do not arise as a result of shallow, surface statistics—with the exception of differences observed for the article+adjective/numeral ablation. Overall, this finding indicates that **LMs can demonstrate a completely novel phenomenon (AANN) by relying on other related—and more frequent—phenomena.**

When AANNs *are* seen during training, however, we observe LMs' results on unseen AANNs to show more similar SLOR values with respect to the the LMs trained on the unablated BabyLM corpus, although they are still significantly reduced in some cases (e.g., singular measure nouns). We conclude that when LMs see evidence of the AANN construction, they do learn from it. But key related phenomena where measure nouns are treated as singular do seem to show some notable effects even when AANNs are present, suggesting that these enable additional learning even when AANNs are present.

5 Experiment 3: The Role of Variability

Results from the previous experiment highlight the importance of seen AANNs in order for LMs to generalize to unseen instances. What properties of these seen instances make LMs generalize?

[3]Trained using KenLM (Heafield, 2011)

More broadly, there is a longstanding question as to how the nature of the instances of a construction provided during learning affect its (partial) productivity (Goldberg, 2005, 2019). In the context of AANNs, we consider the role of *variability* on the open slots of the construction as a factor that might play a role in LMs' productivity on unseen instances. The idea that slot-variability could affect learnability is motivated by theoretical claims in usage-based linguistics (Bybee, 1995), as well as existing research on novel constructions (Suttle and Goldberg, 2011), morphological productivity (Baayen, 2009; O'Donnell, 2015), and inductive inferences in cognitive psychology (Osherson et al., 1990; Xu and Tenenbaum, 2007). The idea is that encountering a slot with a wide variety of lexical items serves as a cue that the slot is flexible.

We hypothesize that instances of AANNs that provide natural evidence of greater open-slot variability—i.e. evidence that many different adjectives, numerals, and nouns can go in their respective positions in the AANN construction—would lead LMs to assign greater likelihood to unseen AANNs. On the other hand, LMs that only encounter a restricted set of instances might overfit, and be more conservative in extending the coverage of possible AANNs to novel combinations of the slot-fillers. To test this, we divided our set of 2,301 AANN-containing utterances in the BabyLM corpus into two roughly equal subsets—one that contained AANNs which were individually highly frequent but restricted in the types of adjective/numeral/nouns, and the other where the AANNs were individually less frequent, but showed more variability in those slots. We obtain these subsets by performing a median split based on the number of unique occurrences of each adjective/numeral/noun triple, resulting in a set of 1149 low variability, and 1152 high variability instances. Details about the slot fillers and examples from each set are provided in Appendix E. We then trained LMs on the BabyLM corpus containing utterances involving either of these two cases. Figure 4 shows the resulting average SLORs obtained from this experiment, along with those obtained by LMs trained on unablated BabyLM and the NO AANN conditions.

From Figure 4, we see that LMs that only saw AANNs that were highly variable in their open-slots demonstrated SLORs that were comparable and sometimes greater than those obtained by LMs that saw all AANNs. By contrast, LMs that only saw AANNs with low variability were as good as

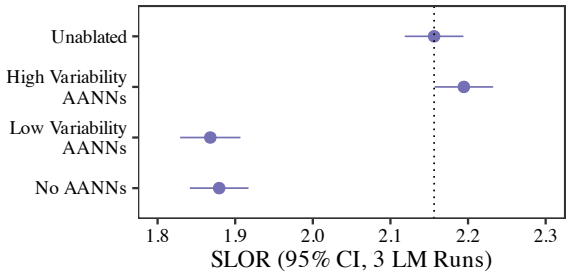


Figure 4: SLORs on AANNs from Mahowald (2023) for LMs trained on BabyLM with low and high variability in the *observed* instances of AANN. SLOR for **unablated BabyLM-trained LM** shown with dotted line.

LMs that never saw any AANNs. Therefore, LMs were sensitive to the nature of range of fillers that went into the construction's open slots, showing relatively greater productivity when they observed evidence that the slots were highly variable. This is compatible with our hypothesis that slot-variability might affect the extent to which LMs "permit" productive uses of a construction.

6 Conclusion

There is increasing interest in computational linguistics in how language models can handle what has been variously called the "long tail" of language (Prange et al., 2021), "extremely rare constructions" (Potts, 2023), "exceptions to syntactic rules" (Leong and Linzen, 2023), "rare linguistic phenomena" (Weissweiler et al., 2024), *inter alia*. Studies of such phenomena are important first because LLMs (and LMs and statistical models in general) are known to be extremely sensitive to frequency and to perform far better in data-rich environments and, second, because the human ability to generalize to rare phenomena is central to knowledge of language.

We found that LMs trained on a human-scale of data can learn a rare construction like the AANN. We found that this learning occurs even without veridical instances of the AANN construction in the training data, and that it is mediated by occurrences of other related constructions in training. As such, these results join a growing body of work showing the ability of LLMs to learn constructions (Tayyar Madabushi et al., 2020; Tseng et al., 2022; Li et al., 2022; Veenboer and Bloem, 2023).

7 Limitations

In future work, it would be valuable to extend this method to a wider range of constructions. But scal-

ing this approach up is not straightforward since it requires identifying and extracting idiosyncratic constructions, and—more onerously—developing testable hypotheses about what makes them learnable from limited amounts of data. While this is a limitation, it also calls for more synergistic collaborations between theoretical and computational linguists.

Another limitation is that our method requires repeated training of LMs from scratch which can be computationally expensive. Alternate methods could be to ablate knowledge of particular hypotheses using representational editing methods such as AlterRep (Ravfogel et al., 2021), etc.

Unlike Weissweiler et al. (2022), we do not test the ability to interpret these constructions for downstream tasks. Instead, our ablations target linguistic form alone. Extending these results to semantic tasks would be quite informative.

Finally, this work only studies a rare construction in English, and on LMs that are trained on English text data. While this is a limitation of the paper, the paradigm introduced can be readily used in future work to study hypotheses and perform indirect evidence elicitation in multi-lingual LMs.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

R Harald Baayen. 2009. 43. corpus linguistics in morphology: morphological productivity. *Corpus linguistics. An international handbook*, pages 900–919.

Marco Baroni. 2022. On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language*, pages 1–16. CRC Press.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Joan Bybee. 1995. Regular morphology and the lexicon. *Language and cognitive processes*, 10(5):425–455.

N. Chomsky. 1957. *Syntactic Structures*. The Hague: Mouton.

N. Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

N. Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. Praeger Publishers.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam Chomsky: The false promise of ChatGPT. *The New York Times*.

Mary Dalrymple and Tracy Holloway King. 2019. An amazing four doctoral dissertations. *Argumentum*, 15(2019). Publisher: Debreceni Egyetemi Kiado.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

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

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele E Goldberg. 2005. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

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

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. *arXiv preprint arXiv:2401.06416*.

Richard S Kayne. 2007. On the syntax of quantity in english. *Linguistic theory and south Asian languages: Essays in honour of Ka Jayaseelan*, 102:73.

Caitlin Keenan. 2013. "A pleasant three days in Philadelphia": Arguments for a pseudopartitive analysis. *University of Pennsylvania Working Papers in Linguistics*, 19(1):11.

Najoung Kim, Tal Linzen, and Paul Smolensky. 2022. Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models. *arXiv preprint arXiv:2212.10769*.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 133–144, Amherst, MA. Association for Computational Linguistics.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

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

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Hillsdale, New Jersey.

Kyle Mahowald. 2023. A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Kanishka Misra and Najoung Kim. 2023. Abstraction via exemplars? a representational case study on lexical category inference in bert. In *BUCLD 48: Proceedings of the 48th annual Boston University Conference on Language Development*, Boston, USA.

Timothy J O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Daniel N Osherson, Edward E Smith, Ormond Wilkie, Alejandro Lopez, and Eldar Shafir. 1990. Category-based Induction. *Psychological Review*, 97(2):185.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Lisa Pearl. 2022. Poverty of the stimulus without tears. *Language Learning and Development*, 18(4):415–454.

Steven Piantadosi. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Christopher Potts. 2023. Characterizing English Preposing in PP constructions. Ms., Stanford University.

Jakob Prange, Nathan Schneider, and Vivek Srikumar. 2021. Supertagging the long tail with tree-structured decoding of complex categories. *Transactions of the Association for Computational Linguistics*, 9:243–260.

Geoffrey K Pullum. 2017. Theory, data, and the epistemology of syntax. In *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, pages 283–298. de Gruyter.

Geoffrey K Pullum and Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):9–50.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Roger Schwarzschild. 2011. Stubborn distributivity, multiparticipant nouns and the count/mass distinction. In *Proceedings of NELS*, volume 39, pages 661–678. Graduate Linguistics Students Association, University of Massachusetts. Issue: 2.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephanie Solt. 2007. Two types of modified cardinals. In *International Conference on Adjectives. Lille*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Laura Suttle and Adele E Goldberg. 2011. The partial productivity of constructions as induction. *Linguistics*, 49(6):1237–1269.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369, Marseille, France. European Language Resources Association.

Tim Veenboer and Jelke Bloem. 2023. Using collostructional analysis to evaluate BERT's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12937–12951, Toronto, Canada. Association for Computational Linguistics.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024. Hybrid human-llm corpus construction and llm evaluation for rare linguistic phenomena. *arXiv preprint arXiv:2403.06965*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages

211–221, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114(2):245.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Dataset Access and Licensing

The AANN acceptability dataset by Mahowald (2023) is released using the MIT License and was accessed from the author's public github repo.[4] The BabyLM dataset[5] does not have a single license of its own but instead inherits the licenses of its constituents: CHILDES (MacWhinney, 2000), BNC Dialogue portion,[6] Children's Book Test (Hill et al., 2015), Children's Stories Text Corpus,[7] Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2020), OpenSubtitles (Lison and Tiedemann, 2016), QCRI Educational Domain Corpus (Abdelali et al., 2014), Wikipedia,[8] Simple Wikipedia,[9] Switchboard Dialog Act Corpus (Stolcke et al., 2000). Since this dataset was specifically released to train LMs, we work under the assumption that our LMs do not violate its license policies. We will follow the inherited licenses' policies while making the trained LMs and ablated BabyLM data public, and refrain from releasing them if we find them to be in violation of the policies.

[4] https://github.com/mahowak/aann-public

[5] accessed from https://babylm.github.io/

[6] http://www.natcorp.ox.ac.uk

[7] https://www.kaggle.com/datasets/edenbd/children-stories-text-corpus

[8] https://dumps.wikimedia.org/enwiki/20221220/

[9] https://dumps.wikimedia.org/simplewiki/20221201/

B LM training details

As mentioned in the main text (see §2), we use the OPT architecture (Zhang et al., 2022) to train our LMs on all versions of the BabyLM corpus. This was the best performing autoregressive LM in the BabyLM Competition (Warstadt et al., 2023). For each instance of the BabyLM (ablated or otherwise), we tune the learning rate[10] based on the validation set, and use the best learning rate as a result of the tuning to train an additional two language models using different seeds. As a result, for each ablation of the BabyLM corpus, we run 6 LM training experiments, which amounts to a whopping 90 LMs for all our experiments. Table 3 contains further details of the training.

(Hyper)parameter	Value
Architecture	OPT (Zhang et al., 2022)
Embed size	768
FFN dimension	3,072
Num. layers	12
Attention heads	12
Vocab size	16,384
Max. seq. length	256
Batch size	32
Warmup steps	32,000
Epochs	20
Total parameters	97M
Training size	100M tokens
Compute	1x NVIDIA A40
Training time	21 hours

Table 3: LM Training details

C Detecting AANNs and related phenomena

In this section, we briefly describe our methods to extract constructions and phenomena relevant to this paper from the BabyLM corpus (Warstadt et al., 2023). Our methods primarily rely on: 1) the surface form of the sentences in the corpus; 2) their corresponding part-of-speech (POS) tag sequences; and in a few cases, 3) their depedency parses. For the latter two, we used `spacy` (Honnibal et al., 2020), specifically, its `en_web_trf` model, which is based on the RoBERTa-base LM (Liu et al., 2019). Next we describe how we used these artifacts to detect our target constructions:

[10] We searched the following set: `{1e-4, 3e-4, 1e-3, 3e-3}`

C.1 AANNs

To detect AANNs we primarily rely on POS-tagged sequences, and construct a regex pattern over them[11] which is able to robustly detect AANNs:

```
pattern = r'\b(DT)(?:(?:\s(RB))*\s(JJ|JJR|JJS)
    (?:\s(CC))*)+(\s(CD|JJ|JJR|JJS|NN|CD\sCD)
    (?:\s(TO|CC)\s(CD))*)(\s(NNS|NNPS|(NN\sNNS)
    |((NN|NNS)_IN_NNS)))+'
```

where we restrict the determiner (DT) to be either 'a', 'an', or 'another'. This regex permits multiple adjectives (*an **exhilarating and marvelous** three months*) optional adverbs (*an **excruciatingly** painful two semesters*), multi-word noun phrases with plural head-nouns (*a refreshing two **glasses of aperol spritz***), numeral-expressions involving subordinate clauses (*a measly **three to five** days*), among other potential edge cases. We additionally use the following adjectives as proxies for numerals, as per the guidelines of Kayne (2007) and Solt (2007):

```
numeral_proxies = ['few', 'dozen', 'couple', '
    several', 'many', 'more']
```

For instance, the following examples would all count as instances of the AANN:

(1) a. a beautiful **few** days.
    b. an amazing **dozen** eggs.
    c. a pictorial **several** pages.
    d. a great **many** days.
    e. an awful last **couple** of days.
    f. a few **more** inches.

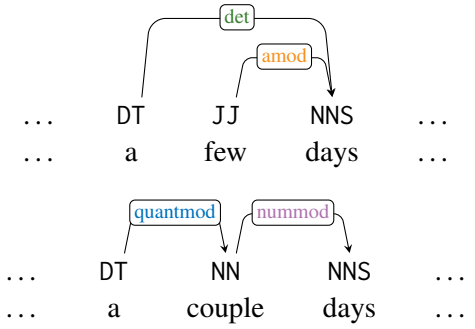
Once detected, we map the found constructions to their respective positions within the AANN format, which allows us to measure metrics such as slot variability, etc.

C.2 DT ANNs

We follow the exact same procedure as the one for AANNs, but no longer restrict the determiner position to only be an indefinite determiner.

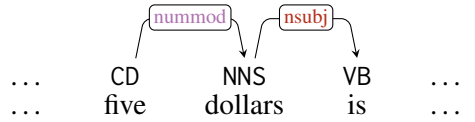
C.3 A few/couple/dozen NOUNs

An important phenomenon that we consider to be related to the AANN involves cases such as: "*that only lasted a few days*" and "*could you bring me a couple liters?*", etc., where the plural nouns are attached to an indefinite article. To detect such cases, we consider the following two dependency configurations, where we have an indefinite determiner (*a, an, another*) with either a det relation with the plural noun (NNS or NNPS) or a quantmod relation with a noun which has a nummod with the plural noun. In the former case, we usually have an amod relation between the noun and the adjective.



C.4 Measure NNS with Singular Verbs

Similar to the previous case, another phenomenon which might be related to the AANN constructions is when measure noun-phrases with plural nouns are treated as singular via their agreement with a verb—e.g., "*five dollars **is** plenty!*" To detect such cases, we again rely on the following dependency configuration, where we have a plural noun (NNS or NNPS) attached to a cardinal number (CD) via the nummod dependency relation, and at the same time also attached to singular verbs via the nsubj dependency relation (i.e., are subjects of singular verbs).



D A/An + ADJ/NUM frequency balancing

A corpus analysis of BabyLM, along with its POS-tagged version suggests that the sequence "a/an/another (JJ|JJR|JJS)" occurs 613,985 times while "a/an/another CD" occurs only 42,111 times – this suggests that adjectives are approximately 14.6 more likely to follow an indefinite article than are numerals. We therefore balance these values by removing 571,874 instances where adjectives follow an indefinite article. This constitutes the largest-sized ablation in this work.

E Variability Analysis

In §5 we compared AANN-generalization of LMs trained on BabyLM versions which differed in the amount of variability that was present in the AANNs

[11] In reality this was constructed over several iterations, taking into account many different possible realizations of the construction in free text.

that the models were exposed to. In particular, we operationalized variability in terms of the slot-fillers of the adjective/numeral/noun slots. Figure 5 shows statistics of the two roughly equal subsets of the AANN-containing utterances in BabyLM. From figure 5, we see that low-variability AANNs were individually more frequent than the high-variability ones.

Variability	Sentences	Freq.
High	A colossal 5 stories	1
	A cold few days	1
	A leisurely six weeks	1
	A long 8 years	1
	A paltry hundred thousand pounds	1
Low	A few more minutes	98
	A few more days	70
	A good six months	5
	A good 4 years	4
	A rough couple days	4

Table 4: Example AANN instances from BabyLM in high and low-variability subsets, as well as their individual frequencies.

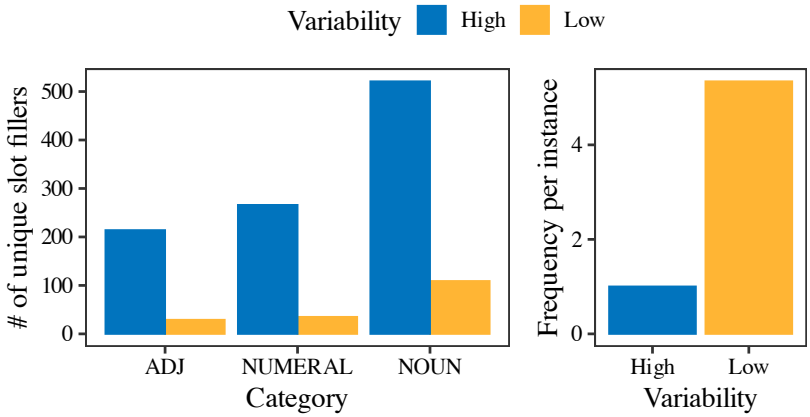


Figure 5: Frequency statistics of instances of the AANN that appear in BabyLM, divided between High- and Low-variability instances, with variability quantified using the number of slot-fillers on the adjective/numeral/noun positions.