BabyLM's First Constructions: Causal probing provides a signal of learning

Joshua Rozner[1], **Leonie Weissweiler**[2], **Cory Shain**[1]

[1]Stanford University [2]Uppsala University
rozner@stanford.edu, leonie.weissweiler@lingfil.uu.se, cory.shain@gmail.com

Abstract

Construction grammar posits that language learners acquire constructions (form-meaning pairings) from the statistics of their environment. Recent work supports this hypothesis by showing sensitivity to constructions in pretrained language models (PLMs), including one recent study (Rozner et al., 2025) demonstrating that constructions shape RoBERTa's output distribution. However, models under study have generally been trained on developmentally implausible amounts of data, casting doubt on their relevance to human language learning. Here we use Rozner et al.'s methods to evaluate construction learning in masked language models from the 2024 BabyLM Challenge. Our results show that even when trained on developmentally plausible quantities of data, models learn diverse constructions, even hard cases that are superficially indistinguishable. We further find correlational evidence that constructional performance may be functionally relevant: models that better represent constructions perform better on the BabyLM benchmarks.[1]

1 Introduction

Construction Grammars (CxG, Goldberg 1995, 2003, 2006; Fillmore 1988; Croft 2001) define constructions as form-meaning pairings and typically assume few innate constraints on the inventory of constructions (constructicon). Thus, a central question in CxG concerns how learners might abstract constructions over time from experience with language (distributional learning; Goldberg 2003; Bybee 2006; Tomasello 2005; Diessel 2004, 2019). Some studies have demonstrated the feasibility of distributional learning of constructions in simplified settings (Casenhiser and Goldberg, 2005; Dunn, 2017). Recent advances in statistical modeling of language (Zhao et al., 2025) have produced pretrained language models (PLMs) that directly instantiate (to a good approximation) the probability distribution over strings (and thus, much of linguistic usage), and a growing literature has explored the use of PLMs as tools for testing usage-based linguistic theories (Piantadosi, 2024; Goldberg, 2024; Millière, 2024; Weissweiler et al., 2025).

Causal Excess Construction	It was **so** *big* **that** it fell over
Affective Adjective Phrase	I was **so** *happy* **that** I was freed
Epistemic Adjective Phrase	I was **so** *certain* **that** I saw you
Idiom	*keep your* **nose clean**
much less let alone	He has not been put on trial (**let alone/much less**) found guilty
Comparative Correlative	**The** *more* **the** *merrier*
Conative	He kicked **at** the ball
Way-manner	We *made our* **way** home
Causative-with	She loaded the truck **with** books
NPN	*questions* **upon** *questions*

Figure 1: Examples of the evaluated constructions. **Bold** reflects *fixed* words and *italics* reflect *schematic* slots constrained to a set of words.

Recently, motivated by collostructional analysis (Stefanowitsch and Gries, 2003) and causal approaches to model study (e.g., Feder et al., 2021), Rozner et al. (2025) hypothesized that constructions should modulate *affinities* (statistical interactions) between words in PLMs' output distributions. For example, in the *let alone* example in Figure 1, RoBERTa assigns 99% probability to both words when either is masked. Rozner et al. hypothesized, following related arguments in linguistics (Croft and Cruse 2004, p. 248-53; Hoffmann 2022, p. 169), that such affinity patterns might hold of constructions more generally. They developed causal probing methods and deployed them on RoBERTa (Liu et al., 2019), showing that diverse construction types are in fact represented

[1]All code and data are provided at https://github.com/jsrozner/cxs_are_revealed.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2238–2250
November 4-9, 2025

in the distribution (cf. Zhou et al., 2024; Bonial and Tayyar Madabushi, 2024; Scivetti et al., 2025; Weissweiler et al., 2024).

Nonetheless, studies of PLM construction learning have largely used models trained on developmentally implausible quantities of training data. For example, though the RoBERTa model used by Rozner et al. is less performant than many of the models studied in prior work, it is still trained on roughly 160GB of text (∼30B words), much more than the 100 million word upper bound estimated for humans by age 13 (Hart et al., 1997; Gilkerson et al., 2017). This limits the degree to which patterns of model learning can support inferences about human learning (Warstadt and Bowman, 2022; Frank, 2023), and it remains an open question whether models trained on developmentally plausible quantities of data can learn constructions (van Schijndel et al., 2019; Yedetore et al., 2023; Mahowald et al., 2024; Millière, 2024).

In this work, we address the particular question of whether there exist settings—namely choices of architecture, training parameters, and word quantity ($\leq 100M$)—in which statistical learners acquire constructions from developmentally plausible *quantities* of data. We evaluate eight models from the 2024 BabyLM Challenge (Hu et al., 2024b), using the experiments from Rozner et al. as a test suite for constructional knowledge. All evaluations are done in the masked language setting (i.e. with bidirectional context). Some of the models perform quite well, indicating that developmentally plausible quantities of data are indeed sufficient to recover knowledge of many constructions, as predicted by the usage-based view. We further find correlational evidence that constructional performance may be functionally relevant: models that better represent constructions perform better on the BabyLM evaluations.

2 Methods

Affinity measures Rozner et al. developed two *affinity* methods (see also Wu et al., 2020; Hoover et al., 2021) and used them to show that constructions manifest as constraints in a `RoBERTa`'s output distribution. Their affinity measures use bidirectional context (constructions often depend on *subsequent* context), so in this study we test only models that support masked language modeling.

Given a string, s, $s \setminus \mathcal{I}$ is defined as the string with the word indices in $\mathcal{I}$ masked, and $\mathcal{P}^i_{s\setminus\mathcal{I}}$ is the *probability distribution* given by the model for the ith position in the *masked string*, $s \setminus \mathcal{I}$. *Global affinity* is then defined simply as the probability assigned to a word given bidirectional context:

$$\mathcal{P}^{(i)}_{s\setminus\{i\}}(w_i)$$

Local affinity measures pairwise interactions by comparing the change in a model's output distribution for a masked position, when another word in the context is also masked, using Jensen-Shannon Divergence (JSD; Lin 1991):

$$a_{i,j} = \text{JSD}(\mathcal{P}^{(j)}_{s\setminus\{j\}}, \mathcal{P}^{(j)}_{s\setminus\{i,j\}})$$

Evaluations Evaluations follow Rozner et al. in using affinity to characterize how well the constructions in Figure 1 are reflected in models' output distributions. Most evaluations look for high global affinities where constructions should constrain word distributions. Local affinity is used to characterize long-range dependency in the CEC in § 3.1. These tests span a wide variety of previously-studied constructions, from fixed (specific-word) to schematic (abstract category), enabling evaluation across degrees of grammatical abstraction: (i) distinguishing the *causal excess construction* from other constructions with the same surface form, (ii) distinguishing literal from figurative usages in potentially idiomatic expressions, (iii) recognizing the fixed (substantive) word constraints in six constructions from the Construction Grammar Schematicity corpus (CoGS, Bonial and Tayyar Madabushi 2024), and (iv) recognizing the abstract (schematic) category constraints of the noun-preposition-noun construction (NPN; e.g., ***day** after **day***, where the bolded slots must be identical nouns) and the comparative correlative (CC).

BabyLMs used for evaluation We test models from the strict and strict-small tracks of the 2024 BabyLM Challenge, in which models were limited to 100M or 10M words, but had no other restrictions on training nor data. For both tracks, we test the best-performing model, `GPT-BERT` (Charpentier and Samuel, 2024), which is a hybrid model that uses both causal and masked language modeling. We select three additional models from both tracks—roughly the next-best-performing models that also support masked language modeling. This gives us one `GPT-BERT`, two `LTG-BERT`, and one `RoBERTa` model for each of the two tracks, so eight total BabyLMs: $\texttt{GPT-BERT}_{\texttt{100M}}$ and $\texttt{GPT-BERT}_{\texttt{10M}}$

(Charpentier and Samuel, 2024), LTG-BERT$_{100M}$ and LTG-BERT$_{10M}$ (the 2024 BabyLM baselines; Hu et al. 2024b), BERTtime$_{100M}$ and BERTtime$_{10M}$ (Theodoropoulos et al., 2024), and RoBERTa architectures ELI5$_{100M}$ and QE CL$_{10M}$ (Lucas et al., 2024; Nguyen et al., 2024). We additionally test four non-BabyLM models in order to provide an estimate of ceiling performance in our evaluations (subscripts indicate large or base version): RoBERTa$_L$ (same as tested by Rozner et al.), RoBERTa$_B$, BERT$_L$, and BERT$_B$ (Liu et al., 2019; Devlin et al., 2019). See Appendix B for more details, including Table 2 for model descriptions.

3 Experiments and Results

Results for all experiments are given in Table 1. Examples of the constructions are found in Figure 1, and additional experiment details are in Appendix C.

3.1 Superficially indistiguishable constructions: the CEC vs. EAP/ AAP

The causal excess construction (CEC; *I was so happy that I cried*) is superficially indistinguishable from epistemic and affective adjective phrases (EAP, AAP; *I was so certain/happy that I saw you*) but admits only *so* as the adverb: *"I was *very happy that I cried"* has an entirely different meaning (Kay and Sag, 2012). Therefore, models that have learned the CEC should assign high affinity to *so* in the CEC but not in the EAP/ AAP. Prior work argued for the difficulty of distinguishing the CEC from the EAP/AAP (Zhou et al., 2024), but Rozner et al. show that the CEC is well-distinguished in the distribution: simply thresholding global affinity scores from RoBERTa$_L$ at 0.78 (i.e., no classifier is trained) correctly characterizes 98% of examples in the Zhou et al. CEC dataset.

To quantify whether this distinction is also learned by BabyLMs, we use the same dataset and compute a receiver operating characteristic (ROC) curve for CEC vs. EAP/AAP classification using global affinity on *so* as the classifier score (again, the contextual probability of *so* is directly treated as the classification score). We report **CEC AUC**, the area under curve, which reflects how likely models are to assign higher affinity to *so* in the CEC than in the EAP/AAP. We also report a **CEC so-that** score, the percentage of multi-that examples (e.g., *I was so happy that$_1$ I won **that**$_2$ I smiled*) where *so* has higher *local affinity* for the causal *that$_2$* than any distractor *that*. A higher score reflects that a model's distribution for *so* exhibits the correct (potentially long-range) dependency.

Results The CEC is well-learned by GPT-BERT$_{100M}$ especially, but also by several other models. Moreover, so-that accuracy tends to be relatively high even on models that poorly classify the CEC (e.g., ELI5$_{100M}$ and LTG-BERT$_{10M}$ have AUC < 0.5 and are thus worse than random), suggesting that models may learn to attend to the correct *that* (i.e. learn what long range dependencies to attend to) before learning to put most probability mass on *so* (i.e. become "confident" that the CEC requires *so*).

3.2 Figurative vs. literal usages

Rozner et al. show that global affinity provides signal in discriminating literal and figurative usages in potentially idiomatic expressions, possibly because the non-compositionality of figurative use leads to greater constraint and thus higher affinity (e.g., one can *spill the beans* but in the same context, one would not *spill the water*, so *beans* should have high affinity). Following their approach, we compute **Idioms AUC** using global affinity to classify figurative vs. literal usages of potentially idiomatic expressions in MAGPIE, a corpus of ~50,000 sentences (Haagsma et al., 2020). Higher scores reflect that models on average recognize greater constraint in figurative usages. In humans, the long tail of idioms is acquired after other constructions (Sprenger et al., 2019; Carrol, 2023); here we are interested in corresponding model behavior.

Results Whereas GPT-BERT$_{100M}$ achieves fairly good performance on many of our constructional evaluations (Table 1), performance on the MAGPIE dataset is barely above chance (AUC 55.3). Perhaps more interesting is that many BabyLMs are substantially *below* chance: the worst model, BERTtime$_{100M}$, provides a classification signal nearly as good as RoBERTa$_B$ if the classifier is flipped. Why might this be the case? Though humans (and to some extent RoBERTa) recognize noncompositional uses and their triggering contexts, our results suggest that BabyLMs trained on less data have not yet learned the long tail of idioms. It makes sense that uncommon, non-compositional usages would be *surprising* to a BabyLM trained on less data and that they would thus exhibit low affinities.

		Classification Accuracy			Global Affinity on Fixed Slots								Schem. Slots	
	BabyLM	CEC		Idioms	Much Less		Let Alone		Con	Way	Caus	CC	CC	NPN
Model	Macro Avg	AUC	so-that	AUC	much	less	let	alone	at	way	with	the	adj/adv	noun (upon)
GPT-BERT$_{100M}$	75.7	93.5	93.5	55.3	62.3	52.2	94.7	94.2	19.1	62.8	91.5	91.6	99.7	81.3
LTG-BERT$_{100M}$	64.0	84.2	83.9	39.9	22.4	7.7	43.9	38.5	19.4	40.2	80.1	69.2	96.5	65.7
BERTtime$_{100M}$	63.1	85.0	87.1	34.4	10.4	5.0	1.5	3.5	14.3	34.2	87.4	92.9	93.8	42.8
ELI5$_{100M}$	59.4	46.5	71.0	37.1	4.1	1.8	9.4	8.6	3.0	3.6	25.1	60.6	81.8	0.2
GPT-BERT$_{10M}$	70.4	85.2	87.1	41.7	5.9	2.6	45.3	53.2	10.7	32.1	84.4	85.8	96.4	17.7
BERTtime$_{10M}$	61.0	74.8	83.9	34.5	4.5	3.0	0.1	0.1	11.7	27.5	76.6	84.4	77.1	29.7
QE CL$_{10M}$	58.8	60.6	80.6	40.5	4.2	1.9	2.2	1.5	14.6	25.7	59.2	57.4	49.1	0.5
LTG-BERT$_{10M}$	57.6	41.1	67.7	36.6	5.6	2.8	0.2	0.2	5.2	24.2	29.1	30.9	30.7	0.3
RoBERTa$_L$		99.4	100.0	69.2	93.5	99.1	98.5	99.6	43.4	84.8	95.0	98.9	99.9	94.4
RoBERTa$_B$		98.8	93.5	66.6	92.6	95.7	99.4	99.9	37.3	74.7	94.3	97.1	99.5	91.5
BERT$_L$		97.9	96.8	62.5	94.8	87.1	91.3	99.3	57.3	74.1	94.1	98.0	99.9	94.6
BERT$_B$		96.0	100.0	56.8	79.1	76.7	88.4	95.3	49.8	65.8	89.6	98.8	98.2	85.8

Table 1: Results. All scores between 0 and 100. Blue scores are classifier accuracy; green shows average affinities, except for CC adj/adv. Left to right: BabyLM Macro-Avg reported on HuggingFace; AUC for classifying CEC vs. EAP/AAP using global aff on *so*; so-that: % of multithat sentences where *so* has higher local aff with the causal *that* than any other *that*; Idioms: AUC for classifying potentially idiomatic expressions as figurative vs. literal using global aff; Fixed Slots: Avg global aff for indicated fixed slot; Schem-CC: For the adj/adv slot, the % of probability mass that the model places on comparative adj/adv; Schem-NPN: Avg global aff for nouns in NPNs with P=*upon*

3.3 Fixed slots in partially substantive constructions

Rozner et al. compute global affinity on the fixed words in ∼50 examples of each of six partially schematic constructions (see Figure 1) from the Construction Grammar Schematicity corpus (CoGs; Bonial and Tayyar Madabushi, 2024) and show that the fixed words often have high global affinities, since the constructional context constrains them. We evaluate whether BabyLMs learn to assign high affinities to fixed words in these constructions and whether there are any differences in degree of acquisition.

Results RoBERTa converges to nearly 100% affinity on all fixed words except the **conative** and **way-manner**; Rozner et al. note that the conative is relatively rare and that way-manner allows other completions. BabyLMs place nontrivial probability on both *at* and *way* but still less than the RoBERTa and BERT models, suggesting that the idiomatic form of the construction may require greater exposure to become entrenched. **CC** and **causative-with** are acquired by a number of models, whereas **let-alone** and **much-less** appear harder to learn. We look more closely at this latter divergence in § 3.6.

3.4 Category constraint in the comparative correlative

Using comparative correlative examples from CoGS we compute the percentage of each model's top-p (p=0.85) completions for the **CC adj/adv** slot that are comparative (e.g., *The **more** the **merrier***). This score represents global affinity at a *category* level, rather than for a single word. The high scores on numerous models indicate that the abstract category constraint that the slot be an adjective or adverb is well-learned. In some cases, the abstract constraint seems to be better captured than the fixed word constraint on *the* (**CC the**).

3.5 Generalization of the form of the NPN

The noun-preposition-noun construction (NPN) is a schematic construction (slots are not constrained to be fixed words, but rather to categories of word—namely two matching nouns and a preposition), so a PLM that shows high affinity for the noun slot has generalized an abstract constraint. Following Rozner et al. we aim to test generalization to *unseen* NPNs by generating a new dataset of 400 NPN sentences (100 nouns, singly-tokenized in all BabyLMs for fair comparison, using each of 4 prepositions: after, upon, by, to), and the last author, blind to affinity scores, rates the acceptability of each sentence. We report average global affinity for nouns in the NPN with *upon* as the preposition (**NPN noun, upon**) where acceptability is ≥ 4 and where the particular NPN is seen 0 times in the GPT-BERT$_{100M}$ training corpus (see Appendix C.5 for additional details and results). High affinities reflect that models learn that each noun slot must match the other noun; average affinity on the noun slot of 81% for *unseen* NPNs in GPT-BERT$_{100M}$ is a strong signal of acquisition.

3.6 Much less and let alone: A brief corpus analysis

Table 1 shows that both **let-alone** and **much-less** are hard for most of the BabyLMs except for GPT-BERT, which much better learns **let-alone** than **much-less**. This is interesting because these constructions have similar semantics. Suprisingly, a review of the $\text{GPT-BERT}_{\text{100M}}$ training data shows that the bigram *much less* occurs almost twice as often as *let alone* (765 vs. 439). However, closer examination reveals that the bigram *much less* occurs in other settings (e.g., *John worked **much less** than Mary*). We use RoBERTa (which reliably distinguishes *constructional* usages of both **let-alone** and **much-less**; see Table 1) to classify usage (global affinity ≥ 0.9 on both words). Whereas almost all of the *let alone* usages are constructional; only 100 ($\sim$13%) of the *much less* usages are. Thus, constructional usage of *much less* is both less frequent and more confusable with other usages, likely impeding acquisition relative to *let alone*.

3.7 Better construction learning is associated with better downstream performance

To assess the functional relevance of construction learning, we include the **BabyLM macro average** in the first column (see Table 3 for details of this score). We compute the correlation of each construction score with the BabyLM macro average and then average, giving an average correlation of $r = 0.78 \pm 0.10$ SD, which shows that performance on the constructional tests in the masked language setting correlates with BabyLM performance.

4 Related Work and Discussion

Prior work on grammatical knowledge in cognitively plausible models has tended to focus on syntactic minimal pairs (Warstadt et al., 2023; Wilcox et al., 2025; Bunzeck et al., 2025; Marvin and Linzen, 2018) using datasets like BLiMP and CLAMS (Warstadt et al., 2020; Mueller et al., 2020). Studies of more complex constructions have tended to use probing and prompting (Rozner et al. 2025; see also Weissweiler et al., 2023; Millière, 2024 for discussion). More recently, BabyLM-like models have been trained on curated corpora to test how relatively rare constructions are acquired from limited data (e.g., Misra and Mahowald, 2024; Leong and Linzen, 2024). Though prior work has questioned whether complex constructions can be learned from cognitively plausible quantities of data (van Schijndel et al., 2019; Yedetore et al., 2023; Mahowald et al., 2024; Millière, 2024), our results provide evidence that even difficult constructions can be acquired from developmentally plausible quantities of data and that the extent of acquisition correlates with general performance.

Prior work has considered the patterns of acquisition of syntactic capacities over model training (Zhang et al., 2021; Warstadt and Bowman, 2022; Wilcox et al., 2025). In this work, we see that when compared to RoBERTa, a larger model trained on $300 - 3000\times$ more data, BabyLMs show divergences in their acquisition of different constructions. In more recent work Bunzeck et al. (2025) report that the distribution of constructions (basic sentence types; e.g., wh-question, copula, imperative) does not substantially impact learning trajectories when models are evaluated on lexical, syntactic, and semantic minimal pairs; although Bunzeck and Zarrieß (2024) provide evidence that the shape of learning curves does vary across different syntactic phenomena in BLiMP. In this study we do find a correlation between corpus composition and performance on *much less* and *let alone*, which makes sense, given that substantive constructions with fixed words must be seen to be learned. The low AUC we observe in BabyLMs for classifying potentially idiomatic expressions likely reflects the same effect: learning the long tail of idioms requires exposure. Future work should investigate the interplay between corpus composition and acquisition dynamics across the spectrum of constructions.

5 Conclusion

Models trained on cognitively plausible quantities of data acquire diverse constructions. This result provides empirical support for the feasibility of distributional learning. Moreover, we found that differences in construction acquisition in BabyLMs exhibit some similarities to human learning. Prior studies of BabyLMs have tended to focus on simpler constructional distinctions via minimal pairs; here we used targeted distributional evaluation to study acquisition of more complex constructions. Given that acquisition correlates with other functional behaviors, future work should examine acquisition dynamics, including interactions between simple and complex constructions during learning.

Limitations

This is a computational modeling study and thus comes with the usual caveats: the "subjects" are models, and inferences to humans should be drawn with care. This study improves on the cognitive plausibility of prior work only in the *amount*—but not the *kind*—of linguistic experience; some recent findings suggest that content may play surprisingly little role in shaping LMs' linguistic abstractions (Feng et al., 2024). 100M words corresponds roughly to the number of words seen by a 13 year-old, and we consider it likely that 13 year-olds have acquired many of the constructions we study in this paper, except possibly rare idioms, though further work might compare the course of acquisition in humans and BabyLMs.

We evaluated only models that support bidirectional context, which is implausible for process models of language comprehension (e.g., Frazier and Fodor, 1978), as the constructions we evaluate depend on subsequent context. Models also differ from humans in other important ways including learning dynamics and architecture (see e.g., Frank, 2023). Though likelihood scores over entire sentences do correlate with human grammaticality judgements (Hu et al., 2024a), no direct work has yet been done to correlate bidirectional affinity scores with human behaviors.

Our study focused on whether model distributions reflect certain formal and semantic distinctions between constructions, but does not allow us to conclude that models "know" constructions in every sense relevant to humans (e.g., that they can recognize and reason about their truth conditions; see e.g., Zhou et al. 2024; Weissweiler et al. 2022 for countervailing evidence). While we leave study of these dimensions of semantic knowledge to future work, we believe that the ability to make the distinctions we study is a critical component of grammar learning and highly relevant to the distributional hypothesis for human language acquisition (see e.g., Tomasello, 2005, p. 30).

Finally, we have framed our study against the theoretical background of construction grammar because we are motivated by considerations about learning that arise from that background. Our results do not in themselves argue for CxG over other theories of natural language syntax, and we do not mean to imply that our findings cannot be accommodated by other views of the nature of human linguistic knowledge. We have simply provided evidence that much constructional information is available to a sufficiently performant statistical learner, as is typically required by CxG theories.

Acknowledgements

Leonie Weissweiler was supported by a postdoctoral fellowship from the German Research Foundation (DFG, `WE 7627/1-1`).

References

Claire Bonial and Harish Tayyar Madabushi. 2024. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in German BabyLMs? In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 169–186, Vienna, Austria. Association for Computational Linguistics.

Bastian Bunzeck and Sina Zarrieß. 2024. Fifty shades of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly. In *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pages 39–55, Gothenburg, Sweden. Association for Computational Linguistics.

Joan Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, pages 711–733.

Gareth Carrol. 2023. Old dogs and new tricks: Assessing idiom knowledge amongst native speakers of different ages. *Journal of Psycholinguistic Research*, 52(6):2287–2302.

Devin Casenhiser and Adele E Goldberg. 2005. Fast mapping between a phrasal form and meaning. *Developmental science*, 8(6):500–508.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *Preprint*, arXiv:2502.10645.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA.

William Croft and D Alan Cruse. 2004. *Cognitive linguistics*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Holger Diessel. 2004. *The acquisition of complex sentences*. Cambridge University Press.

Holger Diessel. 2019. *The grammar network*. Cambridge University Press.

Jonathan Dunn. 2017. Computational learning of construction grammars. *Language and cognition*, 9(2):254–292.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Charles J Fillmore. 1988. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 35–55.

Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.

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

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Adele E Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, USA.

Adele E. Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Betty Hart, Todd R Risley, and John R Kirby. 1997. Meaningful differences in the everyday experience of young american children. *Canadian Journal of Education*, 22(3):323.

Thomas Hoffmann. 2022. *Construction grammar*. Cambridge University Press.

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

Jacob Louis Hoover, Wenyu Du, Alessandro Sordoni, and Timothy J. O'Donnell. 2021. Linguistic dependencies and statistical dependence. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2963, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024a. Language Models Align with Human Judgments on Key Grammatical Constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024b. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Paul Kay and Ivan A Sag. 2012. Cleaning up the big mess: Discontinuous dependencies and complex determiners. In *Sign-based construction grammar*, chapter 5, pages 229–256. Citeseer.

Cara Su-Yi Leong and Tal Linzen. 2024. Testing learning hypotheses using neural networks by manipulating learning data. *Preprint*, arXiv:2407.04593.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

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. *Preprint*, arXiv:1907.11692.

Evan Lucas, Dylan Gaines, Tagore Rao Kosireddy, Kevin Li, and Timothy C. Havens. 2024. Using curriculum masking based on child language development to train a large language model with limited training data. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 221–228, Miami, FL, USA. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Raphaël Millière. 2024. Language models as models of language. *Preprint*, arXiv:2408.07144.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Hiep Nguyen, Lynn Yip, and Justin DeBenedetto. 2024. Automatic quality estimation for data selection and curriculum learning. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 212–220, Miami, FL, USA. Association for Computational Linguistics.

Steven T. Piantadosi. 2024. Modern language models refute chomsky's approach to language. In Edward Gibson and Moshe Poliak, editors, *From fieldwork to linguistic theory: A tribute to Dan Everett*. Language Science Press.

Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025. Constructions are revealed in word distributions. In *The 2025 Conference on Empirical Methods in Natural Language Processing*. To appear.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025. Assessing language comprehension in large language models using construction grammar. *CoRR*, abs/2501.04661.

Simone A Sprenger, Amélie la Roi, and Jacolien Van Rij. 2019. The development of idiom knowledge across the lifespan. *Frontiers in Communication*, 4:29.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.

Nikitas Theodoropoulos, Giorgos Filandrianos, Vassilis Lyberatos, Maria Lymperaiou, and Giorgos Stamou. 2024. BERTtime stories: Investigating the role of synthetic story data in language pre-training. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 308–323, Miami, FL, USA. Association for Computational Linguistics.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. 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 5831–5837, Hong Kong, China. 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.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest*

2023), pages 85–95, Washington, D.C. 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. *Preprint*, arXiv:2403.06965.

Leonie Weissweiler, Kyle Mahowald, and Adele Goldberg. 2025. Linguistic generalizations are not rules: Impacts on evaluation of lms. *Preprint*, arXiv:2502.13195.

Ethan G Wilcox, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions are so difficult that Even large language models get them right for the wrong reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811, Torino, Italia. ELRA and ICCL.

A Methodological Details

Experiments were run either on an M3 Macbook Pro or on a single Nvidia RTX A6000 cluster GPU. Classification of idioms (MAGPIE) is the only experiment requiring any meaningful amount of computation and takes roughly 3 hours on a single RTX GPU.

B BabyLM Supplement

Table 2 gives details for all models tested, and Table 3 reproduces each BabyLM's scores that make up the macro average score reported in Table 1.

The 2024 BabyLM Challenge is the second iteration of the BabyLM Challenge (Warstadt et al., 2023). There were 18 models submitted to the strict track and 32 submitted to the strict small track. Of these models many support only autoregressive modeling and are thus not suited to testing using the bidirectional methods developed by Rozner et al. (in general, identifying constructions depends on bidirectional context). The best-performing model in both the strict and strict-small tracks was `GPT-BERT`, a hybrid `GPT` (autoregressive)/ `BERT` (masked language model; MLM) architecture (Charpentier and Samuel, 2024), which was based on `LTG-BERT` (Samuel et al., 2023). As `GPT-BERT` can be run as an MLM it can be evaluated using Rozner et al.'s approach.

We select three additional models for both tracks to give eight total BabyLM models for evaluation. For both tracks, we take the next top-performing MLMs (i.e., excluding purely autoregressive models), though we additionally exclude DeBERTa models because their tokenization scheme was not easily adapted to work with the experiment pipeline. As this would have given us four `LTG-BERT`-style models for the strict track, we include a slightly worse-performing RoBERTa architecture for the final strict model. Models are obtained from HuggingFace (Wolf et al., 2019) at `https://huggingface.co/spaces/babylm/leaderboard-2024`, accessed March 18, 2025.

In the 2024 BabyLM Challenge, submissions were limited to a total *quantity* of words but were not subject to any other limitations on training approaches. The best-performing BabyLM model, `GPT-BERT`, though limited to 100M or 10M words for the strict and strict-small tracks, still trained for many epochs over the same data (Hu et al., 2024b; Wilcox et al., 2025). In the 2025 BabyLM Challenge, a limit is placed on the the number of times data can be seen (Charpentier et al., 2025). For this study, as we are not treating PLMs as models of the learning process, but only as lower bounds on the linguistic knowledge that can be acquired from particular quantities of input, the particular details of training (e.g., more epochs) were not at issue. Nonetheless, such divergences are interesting and would be relevant to any subsequent work that looks more closely at, e.g., the mechanisms or trajectories of acquisition.

C Evaluations

As the methods developed by Rozner et al. (2025) are for singly-tokenized words, all words evaluated using either affinity method are singly-tokenized. This constraint affects only the results for figurative vs. literal usage and NPN generalization studies, details of which we provide here.

C.1 CEC vs. EAP/AAP

Rozner et al. hypothesize that, as *so* is necessary in the CEC but not in the EAP/AAP (**It was **very** big that it fell over*), PLMs might identify this constraint in their output distribution. Rozner et al. find that global affinity in $\texttt{RoBERTa}_{\texttt{L}}$ on *so* in fact robustly distinguishes the CEC from EAP/AAP and that in multi-that sentences (e.g., *I was so happy that I won **that** I smiled*), the distribution for *so* is always more affected by the causal that.

The CEC dataset has 323 possible examples. After cleaning and labeling, we have 292. Four are invalid, leaving us with a total of 22 EAP, 73 AAP, and 193 CEC (288 total). Rozner et al. identify a couple mislabeled examples, and we use their corrected labels.

The accuracy we report for $\texttt{RoBERTa}_{\texttt{L}}$ is slightly lower than that reported by Rozner et al.. As we use a HuggingFace FastTokenizer, our evaluation included 15 examples that were omitted in Rozner et al.'s original study. As these omitted examples had more awkward punctuation (tokenization issues were what led to their omission in the original study) that might have made them harder to correctly classify.

CEC multithat for so-that local affinity study We use the 31 entry multi-that dataset from Rozner et al..

C.2 Figurative vs. literal usages

Here we provide two examples drawn from the MAGPIE dataset for *nuts and bolts* (same as

Model	Arch.	# Par.	Training Data		HuggingFace ID
			Words	Source	
GPT-BERT$_{100M}$	GPT-BERT	119 M	100 M	BabyLM, FineWeb-Edu, Cosmopedia	ltg/gpt-bert-babylm-base
GPT-BERT$_{10M}$		30 M	10 M		ltg/gpt-bert-babylm-small
LTG-BERT$_{100M}$	LTG-BERT	99 M	100 M	BabyLM	babylm/ltgbert-100m-2024
LTG-BERT$_{10M}$		99 M	10 M		babylm/ltgbert-10m-2024
BERTtime$_{100M}$	LTG-BERT	98 M	100 M	BabyLM, TinyStories, GPT-Neo completions	nikitastheo/BERTtime-Stories-100m-nucleus-1
BERTtime$_{10M}$		24 M	10 M		nikitastheo/BERTtime-Stories-10m-nucleus-1-balanced
ELI5$_{100M}$	RoBERTa	44 M	100 M	BabyLM + reddit ELI5	3van/RoBERTa_100M_ELI5_CurriculumMasking
QE CL$_{10M}$	RoBERTa	43 M	10 M	BabyLM - filtered QE	jdebene/BabyLM2024
RoBERTa$_L$	RoBERTa	304 M	∼30 B	Bookscorpus, English Wiki, cc-news	FacebookAI/roberta-large
RoBERTa$_B$		86 M	∼30 B		FacebookAI/roberta-base
BERT$_L$	BERT	304 M	∼3 B	Bookscorpus, English Wiki, openwebtext, stories	google-bert/bert-large-cased
BERT$_B$		86 M	∼3 B		google-bert/bert-base-cased

Table 2: Model Details

Model	Evals reported by BabyLM 2024				
	BliMP	BLiMP Supplement	(Super) GLUE	EWoK	Macro Avg
GPT-BERT$_{100M}$	86.1	76.8	81.5	58.4	75.7
LTG-BERT$_{100M}$	69.2	66.5	68.4	51.9	64
BERTtime$_{100M}$	65.6	65	72.7	49.2	63.1
ELI5$_{100M}$	60.2	56.8	67.7	53	59.4
GPT-BERT$_{10M}$	81.2	69.4	76.5	54.6	70.4
BERTtime$_{10M}$	63.2	59.3	71.1	50.4	61
QE CL$_{10M}$	61.9	58.3	64.4	50.8	58.8
LTG-BERT$_{10M}$	60.6	60.8	60.3	48.9	57.6

Table 3: BabyLM scores

Rozner et al.):

Literal usage: They would include orders for routine raw materials such as steel stock; screws; *nuts and bolts*; lubricants and fuel oil.

Figurative usage: Jay comes from a different end of the spectrum to Dave Ambrose, but the two both like to talk *nuts and bolts*.

Magpie has 48,395 unique sentences with a total of 129,397 words used in potentially idiomatic expressions that are labeled as figurative or literal uses (average of 2.6 words per sentence/ example). Of the 48,395 examples, we omit 3,944 *sentences* that do not have $\geq 99\%$ confidence in annotation. This leaves us with 44,451 sentences with a total of 119,401 words. Among those 119,401 words, 2,016 have wrong offsets, giving us 117,385 words for the analysis. Of the 45,450 (117,385) sentences (words), 10,313 (23,484) are labeled as literal and 34,138 (95,917) are labeled as idiomatic.

For the result presented in the main paper, we omit from consideration any word or sentence that could not be processed for *any* model (43,124 words). In general, such processing failures are a result of differences in tokenization behavior: the Rozner et al. methods are developed for singly-tokenized words and each BabyLM has a different set of singly-tokenized words. This allows us to make a fair comparison across all models. This leaves us with 74,261 words for the analysis in the main paper.

Whereas Rozner et al. omit sentences with fewer than 10 words of context and words with less than four characters, we include all data. The 69.2 score for RoBERTa$_L$ matches their corresponding result: they report an AUC for the whole dataset of 0.71 with omission and 0.69 without omission. This correspondence suggests that even with this study's restriction to a common vocabulary over all models, the underlying trend in classification behavior is not substantially affected.

Model	50th %ile	80th %ile	CC-score
GPT-BERT$_{100M}$	2.2	3.1	99.7
LTG-BERT$_{100M}$	2.6	8.5	96.5
BERTtime$_{100M}$	4.6	25.9	93.8
ELI5$_{100M}$	15.6	185.8	81.8
GPT-BERT$_{10M}$	2.6	6.2	96.4
BERTtime$_{10M}$	8.6	48.9	77.1
QE CL$_{10M}$	59.6	818.0	49.1
LTG-BERT$_{10M}$	18.3	157.7	30.7
RoBERTa$_L$	2.1	2.6	99.9
RoBERTa$_B$	2.2	3.4	99.5
BERT$_L$	2.1	2.7	99.9
BERT$_B$	2.5	5.4	98.2

Table 4: CC: Average number of words (sorted by model likelihood) needed to get to nth %-ile of output distribution for CC adj/adv slot. CC-score same as in Table 1 for comparison.

C.3 Fixed slots in partially substantive constructions

Examples of the six partially substantive constructions (from Bonial and Tayyar Madabushi 2024), with fixed words italicized. (These are the same as in Figure 1.)

Causative-with: She loaded the truck *with* books.
Comparative correlative: *The* more *the* merrier. (In our analysis the two *the* words are considered as a single class.)
Conative: He kicked *at* the ball.
Let-alone: None of these arguments is particularly strong, *let alone* conclusive.
Much-less: He has not been put on trial, *much less* found guilty.
Way-manner: We made our *way* home.

C.4 Category constraint in the comparative correlative

We replicate the procedure of Rozner et al.: Using the 54 CC examples from the CoGS dataset, we mask each comparative adjective/adverb, obtain the set of highest probability outputs at the masked position that sum to 85% probability mass, and calculate a *comparative score*: the percentage of this set that is a comparative adjective/adverb.

To calculate the percentage of the output distribution nucleus that is a comparative adj/adv, we order the outputs by probability and iterate through them until reaching a total probability mass of $p \geq 0.85$ (a nucleus using 0.85). Rozner et al. use a nucleus of 0.98, but since the BabyLM models have output distributions with much higher entropy than RoBERTa (see Table 4), we consider a smaller nucleus to avoid summing probabilities over the whole vocabulary.

For each sampled word, we substitute it into the original sentence and use Spacy (Honnibal et al., 2020) to check whether it is a comparative adverb or comparative adjective. Whereas Rozner et al. use the transformer version of Spacy's tagger, we use the small, non-transformer model since the higher entropy distribution in the BabyLMs causes us to calculate for many more possible fills. Given that the tagger module sees the whole sentence (and may already "know" the CC construction), it may be biased to label words as comparative even if they are not. The final score is the proportion of the sample (the 85% nucleus) that is a comparative adjective or adverb.

Of the 108 ($= 54 \times 2$) candidate slots, across all twelve models, an average of 5.7 words cannot be processed (due to multi-tokenization).

C.5 Generalization of the form of the NPN

NPN dataset generation We follow Rozner et al.'s procedure in generating a new NPN dataset (below adapted from their Appendix). We use GPT-4 via the OpenAI API, version gpt-4-0613, temperature 0.7, max tokens 100. Total cost to produce 400 sentences is less than $5. We prompt as follows, where "{phrase}" is the particular targeted NPN (e.g., *day by day*):

> An NPN construction is one like "day by day" or "face to face". It has a repeated singular noun with a preposition in the middle. Other prepositions are also possible: "book upon book", "week over week", "year after year". Please use "{phrase}" in an NPN construction, placing "{phrase}" in the middle of the sentence. Make sure the sentence establishes a context in which the noun makes sense. Please provide only the sentence in the response.

We verify that each generation matches the desired form noun+prep+noun.

To obtain acceptability judgements, we randomly sort all sentences and the last author annotates with a score between 1 and 5, inclusive.

Results In the main text we reported only the average affinity score for NPNs using *upon* as preposition. In Table 5 we report average scores for

- all NPNs (all four prepositions: upon, after, by, to),

Model	upon			after			by			to		
	All	Freq	Acc	All	Freq	Acc	All	Freq	Acc	All	Freq	Acc
GPT-BERT$_{100M}$	73.4	81.3	83.7	55.6	57.2	61.6	42.3	55.1	56.8	60.5	70.2	70.6
LTG-BERT$_{100M}$	61.0	65.7	68.1	39.2	41.8	45.1	49.6	63.0	64.9	31.1	38.3	38.2
BERTtime$_{100M}$	38.7	42.8	45.7	21.7	20.3	22.8	33.8	42.7	44.7	8.3	10.5	10.3
ELI5$_{100M}$	0.2	0.2	0.2	0.2	0.2	0.2	0.7	0.3	0.8	0.2	0.2	0.2
GPT-BERT$_{10M}$	19.3	17.7	23.2	14.8	13.8	17.3	26.1	29.7	35.4	9.2	11.8	11.7
BERTtime$_{10M}$	26.0	29.7	30.2	14.7	12.5	15.3	33.6	45.0	45.8	2.8	3.4	3.4
QE CL$_{10M}$	0.6	0.5	0.5	0.3	0.1	0.1	1.3	0.2	2.2	0.2	0.2	0.2
LTG-BERT$_{10M}$	0.4	0.3	0.3	0.1	0.0	0.0	0.5	0.2	0.6	0.3	0.3	0.3
RoBERTa$_{L}$	88.6	94.4	95.2	72.6	80.7	82.7	52.8	68.3	72.5	86.3	94.1	94.2
RoBERTa$_{B}$	84.5	91.5	92.8	69.0	73.5	76.4	54.9	70.9	75.6	79.4	88.9	89.0
BERT$_{L}$	90.2	94.6	95.2	73.1	81.1	83.1	52.6	66.4	69.2	80.1	87.0	87.1
BERT$_{B}$	79.1	85.8	87.2	55.6	60.0	63.6	46.6	55.8	63.2	55.9	63.1	63.3

Table 5: Results for NPN. All includes all 100 NPNs for each. Freq limits to acceptability ≥ 4 and further restricts to NPNs that occur 0 times in the GPT-BERT$_{100M}$ dataset. Acc limits to NPNs with acceptability ≥ 4. Freq tends to have worse performance than Acc because it excludes NPNs that models were more likely to have seen, even if they are just as acceptable.

- all NPNs with acceptability ≥ 4 and filtered to those which occur zero times in the GPT-BERT$_{100M}$ training data (same as in Table 1)
- all NPNs when filtered to acceptability ≥ 4.

(It is possible that some of the NPNs which have zero-frequency in the GPT-BERT$_{100M}$ training data have non-zero-frequency in the other models' data.)

When filtering to acceptable NPNs, we have upon: 72, after: 76, by: 64, to: 52. Of these 264, 219 are seen zero times in GPT-BERT$_{100M}$'s training data.

Our results for RoBERTa$_{L}$ generally agree with the prior results of Rozner et al.: NPNs with *upon* are most well-generalized. Our results seem to show better generalization for NPNs using *to*, which could result from using the simpler common vocabulary across the BabyLMs.

Our overall NPN results show (i) gradient generalization of all NPNs, and (ii) sensitivity to acceptability (more acceptable have higher affinity), agreeing with Rozner et al.'s results. In general, removing NPNs that were in GPT-BERT$_{100M}$'s training data reduces average affinity scores. This makes sense given that there is overlap in the datasets for most of the models that were trained (all contain some part of the BabyLM dataset). This implies an increase in affinity for NPNs that are observed in the training data. This could be either an effect of memorization or it could reflect that NPNs in the training data have some underlying property that makes them more acceptable to the models, as measured by affinity.

D Use of AI Assistant

ChatGPT was used to produce initial versions of some python matplot code. Any code produced was subsequently adapted, reviewed, and/or modified. ChatGPT was not used to write any part of this paper.