Are BabyLMs Second Language Learners?

Lukas Edman[1,2] **Lisa Bylinina**[3] **Faeze Ghorbanpour**[1,2] **Alexander Fraser**[2,4]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning
[3]Institute for Language Sciences, Utrecht University
[4]School of Computation, Information and Technology, TU Munich

lukas@cis.lmu.de, e.g.bylinina@uu.nl, faeze.ghorbanpour@lmu.de, alexander.fraser@tum.de

Abstract

This paper describes a linguistically-motivated approach to the 2024 edition of the BabyLM Challenge (Warstadt et al., 2023). Rather than pursuing a first language learning (L1) paradigm, we approach the challenge from a second language (L2) learning perspective. In L2 learning, there is a stronger focus on learning explicit linguistic information, such as grammatical notions, definitions of words or different ways of expressing a meaning. This makes L2 learning potentially more efficient and concise. We approximate this using data from Wiktionary, grammar examples either generated by an LLM or sourced from grammar books, and paraphrase data. We find that explicit information about word meaning (in our case, Wiktionary) does not boost model performance, while grammatical information can give a small improvement. The most impactful data ingredient is sentence paraphrases, with our two best models being trained on 1) a mix of paraphrase data and data from the BabyLM pretraining dataset, and 2) exclusively paraphrase data.

1 Introduction

Language models (LMs) need a lot of data in order to learn to approximate human linguistic behaviour (Warstadt and Bowman, 2022). The amounts of linguistic data typically used for training recent LMs is is significantly larger than what is available for most of languages of the world, and also much more than what children are typically exposed to during their first language acquisition. A 13 year old is typically exposed to less than 100 million words of linguistic input, which is orders of magnitude less than the amount used in LM pretraining. And still, LMs fail to be quite as good in language as human learners. Can we teach our models to be more data-efficient? If yes, how?

There are two potential strategies. One is to study how children acquire language in a natural setting, and use their acquisitional trajectories and patterns as inspiration for LM training. This intuition is one of the motivations for the BabyLM Challenge (hence the name; other low-resource pretraining contexts are, of course, also relevant): the challenge encourages LM pretraining optimization advancements inspired by human linguistic development (Warstadt et al., 2023).

Another direction is to embrace the obvious differences between LM pretraining and the ways human learners acquire their native language. The architectures of current LMs are dramatically different from human brain anatomy, and training objectives and strategies have only limited psycholinguistic developmental parallels. Finally – and most importantly for our contribution – input for first language acquisition by human learners and for LM pretraining is hardly comparable not only when it comes to dataset size. While the amount of strictly linguistic input that children get is small compared to typical LM training data, children get this input in communicative context that LMs lack at the pretraining stage, and it is typically paired with cross-modal data, which is not part of the `strict-small` track we choose for the BabyLM Challenge.

At a very high level, taking this second direction means that we look beyond human linguistic and cognitive development for optimization strategies – or at least, we do not need to expect that those will be the ones that necessarily work best.

We sharpen this point and contrast language learning in an acquisitionally realistic setting (first-language, or L1, acquisition) – and language learning in a more artificial setting – learning a second language, L2; a human activity that also leads to (different levels of) linguistic proficiency but contrasts dramatically with L1 acquisition by children. Almost everything is different: the set-up, the data, typical tasks the learner faces, and very often modality and their combinations.

We choose this particular direction mainly be-

cause in the current, second, edition, of the BabyLM Challenge participants are allowed to construct their own datasets within the track word budget. A lot of submissions last year, including ours, experimented with curriculum learning – different ways to order the same data (see our submission Edman and Bylinina (2023) as well as the BabyLM 2023 findings (Warstadt et al., 2023)). These attempts gave only limited results.

This year we instead focus on the effect of choosing different data on LM pretraining. In particular, roughly in line with how people learn foreign languages through explicit linguistic instruction, we divide training data into blocks roughly corresponding to types of linguistic information commonly found in English-as-a-foreign-language courses. We participate in the `strict-small` track allowing for only 10M words and experiment with four different types of linguistic information:

- **Lexical information** (information about word meaning and use), parallel to word learning in L2 acquisition. We use Wiktionary data as a source of this knowledge.
- **Grammatical information**, parallel to grammar learning for L2. We try two ways of constructing grammar data: a set of sentences marked with grammar phenomena, and texts of grammar books for L2 English learners.
- **Paraphrasing** has perhaps fewer obvious parallels in L2 learning practice, but is related to the explicit focus on sentential semantics ('different ways to say the same thing') and how different modifications in syntax and vocabulary can preserve and alter the meaning of a sentence, which is a common focus in L2 class discussions and exercises. For this data, we use one of the two SynSCE corpora from Zhang et al. (2021).
- A mix of **unconstrained textual data** that corresponds to various input during language acquisition of any kind, be it L1 or L2 acquisition. For this, we use portions of the BabyLM data provided by the challenge organizers.

We find that data on paraphrasing brings in the most significant improvements. Grammatical information is only marginally useful, even though it does come with some improvement, depending on the training set-up. Finally, lexical information does not seem useful for LM pretraining. One cannot be sure what to attribute these results to: the usefulness or lack thereof of particular types of data; the quality of the actual various datasets that we use; or the properties of evaluation used to judge whether a particular type of data is useful. One way or another, our answer to the question of whether BabyLMs are L2 learners is 'only when it comes to certain types of data'.

2 Data

2.1 BabyLM data

We make use of data provided by BabyLM organizers for our experiments. One of our two submitted models (Contr.) doesn't use BabyLM data at all, while the other one (Half/Half) uses a subset of BabyLM data. In the Half/Half model, we use the following parts of the BabyLM dataset:

Dataset	Words
Simple Wikipedia	145K
Gutenberg	254K
Switchboard	147K

Table 1: BabyLM data used for the Half/Half model.

We think BabyLM data roughly corresponds to unconstrained linguistic input in a language learner's experience (reading materials and practice conversations with language teachers and peers).

The rest of the data in the Half/Half model comes from the dataset we discuss next.

2.2 Contrastive dataset

An important part of the language acquisition experience is finding out how changes in phrasing and syntactic structure can alter or preserve meaning. This is seen in typical L2 learning tasks such as paraphrasing, which highlight the semantics of the sentence and the ways syntactic manipulation can affect its meaning.

As data approximating this type of information, we use a dataset by Zhang et al. (2021). They release two datasets as part of SynCSE, a contrastive learning framework for training sentence embeddings. The data in both datasets (SynCSE-partial and SynCSE-scratch) is synthetic: synthesized by LLMs. The two different datasets are results of different prompting set-ups (for the dataset construction and prompting details, we refer the reader to the original paper). We use one of these two

datasets, SynSCE-partial[1].

The dataset is structured as follows: each datapoint comes as a triple consisting of 1) a sentence; 2) its paraphrase, and 3) a hard negative (a sentence that is similar to the original one lexically and/or structurally but has a different meaning). Here is an example of a triplet from the dataset:

> **sent0:** `One of our number will carry out your instructions minutely.`
>
> **sent1:** `One person from our group will execute your instructions with great attention to detail.`
>
> **hard_neg:** `Each member of our group will carry out your instructions differently.`

We use all three elements of the triplet in our experiments.

2.3 Grammar data

To mimic explicit grammar instruction in the typical L2 learning setting, we look for ways to expose the model to targeted grammatical information. We explore two strategies and corresponding datasets, which we call `Gram Gen` and `Gram Books`.

For **Gram Gen**[2], we first compile a list of grammatical notions that a sentence can contain. This list is inspired by the typical structure of reference and learners' grammars and the topics covered by those. We then pass these notions to GPT 4o-mini[3] to generate examples, using the prompt in Figure 1. To ensure that we generate a diverse set of sentences, we prompted the model to generate sentences about specific topics.[4]

After this, we additionally tag each sentence with the grammatical notions as a sentence can contain more than one. This again is done with GPT 4o-mini, using the prompt in Figure 2. Due to pricing restrictions, we generate 500 sentences per notion, and tag 100 of these sentences for 50 different notions. We include an example of a sentence tagged, where we verify the correctness of the given tags in Table 2.

In the table we can see that GPT 4o-mini appears only partially capable of recognizing grammatical notions. For the simpler, very well-known notions such as common nouns, verb person, tense, and number, GPT performs well. For less commonly-known phenomena, such as ellipsis, it seems to have no understanding. For ellipsis specifically, GPT often has false positives with sentences of this 2-clause structure, likely because that is a necessary component for an ellipsis to occur, but not what defines an ellipsis. GPT also appears to occasionally hallucinate, with "it" not appearing in the sentence despite it being tagged as an object pronoun. Overall, given the accuracy of GPT in tagging, it is not surprising that our model would struggle to grasp grammatical notions.

"The engineers proposed a new design for the bridge, while the architects focused on the aesthetic elements, emphasizing sustainability instead."

Notion	Tag	Correct?
common noun	engineers, design, bridge, architects, elements, sustainability	✓
collective noun	engineers, architects	✓
singular noun	design	✓
plural noun	engineers, architects, elements	✓
nominative case	The engineers	✓
simple past tense	proposed, focused, emphasized	✓
third person	engineers, architects	✓
plural verb	proposed, focused, emphasizing	✓
indicative mood	proposed, focused, emphasizing	✓
non-gradable adjective	sustainable	✓
positive adjective	sustainable	✗
aspectual adverb	emphasizing	✗
comparative adverb	instead	✗
object pronoun	it	✗
case preposition	for, on, instead	✓
coordinating	while	✓
indefinite determiner	a new design	✓
noun phrase	The engineers, a new design, the bridge, the architects, the aesthetic elements, sustainability	✓
adjectival modification	aesthetic, sustainability	✓
verb phrase	proposed, focused, emphasizing	✗
transitive verb phrase	proposed a new design, focused on the aesthetic elements, emphasizing sustainability	✓
direct object	design, elements	✓
adjunct clause	Yes	✓
ellipsis gapping	Yes	✗
ellipsis pseudo-gapping	Yes	✗

Table 2: Tags produced for the sentence above. Only positive tags are shown for brevity. ✓ indicates the tag is completely correct, ✓ partially correct, ✗ incorrect.

We construct the second grammar dataset, **Gram Books**, as an alternative to grammatical instruction via examples. This dataset contains grammar books that overtly discuss the rules of English grammar and are intended mainly for second language learners of English. Here is the full list

[1] https://huggingface.co/datasets/hkust-nlp/SynCSE-partial-NLI

[2] We release this dataset on HF: `link placeholder`.

[3] We changed from GPT 3.5 to 4o-mini due to pricing changes.

[4] The possible topics are: accounting, anthropology, archaeology, architecture, art, artificial intelligence, astronomy, biology, botany, business, chemistry, computer science, cosmology, criminology, design, economics, education, environmental science, engineering, geography, geology, government, history, humanities, international relations, journalism, law, literature, linguistics, math, medicine, music, philosophy, physics, poetry, politics, psychology, religion, sports, and theater.

You are an expert in grammar. Write 500 detailed sentences containing <notion> (as opposed to <alternate notion>). Make sure to write 500 detailed sentences that are all different from each other. Try to make the sentences sufficiently different, for example, don't start every sentence with "the", make both short and long sentences, and write about the topic of <topic>. Don't write anything else.

Figure 1: The prompt used to generate example sentences of a grammatical notion. The <alternate notion> is not always used, but corresponds to notions with clear alternatives, such as telic vs. atelic verbs.

Consider the sentence: <sentence> Does the sentence contain the notion of <notion>? If so, write which word or words correspond to the notion. If not, write "N/A". Only write the word or words that correspond, or N/A otherwise.

Figure 2: The prompt used to tag sentences with their grammatical notion. The prompt for sentential notions only contained the initial question, along with: "Answer with yes or no. Only write 'yes' or 'no', nothing else."

of the grammar books we used: Newson (2006); Greenbaum and Nelson (2009); Roth and Aberson (2010); Thomson and Martinet (2015); Brutjan and Brutjan (2022); Wright (2024). We do not release this dataset due to copyright constraints.

We use both grammar datasets for two types of experiments: 1) regular MLM training (described in Section 3.2); 2) more elaborate training schemes involving a combination of an encoder and a decoder (discussed in Section 3.3).

2.4 Wiktionary

For lexical instruction, we make use of a segment of data from Wiktionary[5], the largest available collaborative source of lexical knowledge. We constrain ourselves to the English segment of Wiktionary, and extract the lemma together with parts of speech and the definitions of each of its senses and examples that illustrate the senses.

We parse the Wiktionary data into CSV, where each row contains a word, part of speech, a definition, and up to 13 examples, though many contained no examples.

[5]http://www.wiktionary.org/

Give 3 examples of the word <word> as a(n) <part of speech>, where it means <definition>. List the 3 examples in a numbered list, they should be full sentences. Don't say anything else. The format should look like:
1. Example 1
2. Example 2
3. Example 3

Figure 3: The prompt used to generate example sentences of a word sense.

For words without an example, we attempted two things: we generated examples with GPT 3.5, and we fed the word in as is. The examples generated were of notably high quality, with GPT even able to generate sentences for rare word senses. The prompt we used is shown in 3.

As with other types of linguistic knowledge, with this data we are looking for a way to mimic typical L2 learning. Wiktionary comes pretty close to word learning in this setting, as it contains explicit information about different senses of the word, its morphological and syntactic profile, defines its lexical semantics and illustrates all of this information with sentences where the word is used in its different senses.

Again, as with grammar data, we use the resulting Wiktionary dataset[6] both in experiments with simple MLM pretraining and in experiments with more complicated training set-ups, which are described in more detail in Sections 3.2 and 3.3, respectively.

3 Method

3.1 Model Choice

We opted to use encoder-only models for our final submission. This is based on our observation from last year's competition, where encoder-only models generally outperformed decoder-only or encoder-decoder models. We chose the DeBERTa-base (He et al., 2021) architecture as it is considered state-of-the-art for encoder-only models. Unlike in last year's competition where we saw improvements

[6]The dataset we construct is available on HF: `link placeholder`.

from using DeBERTa-large, we saw no improvement this year in initial testing and thus only used the base model size.

3.2 Training and Evaluation

Our pretraining uses the standard MLM scheme (Liu et al., 2019), which we used last year to great effect. Table 3 shows the hyperparameters we used for our pretraining experiments. For fine-tuning, we use the default hyperparameters provided by the organizers.

Hyperparameter	Value
Vocabulary size	40000
Context size	64
Learning rate	2e-4
Decay	0.01
Warmup steps	4000
Optimizer	AdamW
Batch size	64, 256
Epochs	50

Table 3: Hyperparameters used.

The hyperparameters chosen are largely the same as what we used in last year's competition (Edman and Bylinina, 2023), with some minor changes to the learning rate (2e-4 vs. 1e-4) and warmup steps (4000 vs. 10000), as well as using both a batch size of 64 and 256. We found that, in some circumstances, a batch size of 64 would result in a more performant model, but this phenomenon was inconsistent. As such, we report the best performing batch size for each model. We note that "context size" refers to the number of tokens in a given example. This is constant, so each example may contain multiple sentences or fragments.

We evaluate our models with the tasks included in this year's shared task: BLiMP (Warstadt et al., 2020), BLiMP supplement, (Super-)GLUE (Wang et al., 2018, 2019), and EWoK (Ivanova et al., 2024).

3.3 Additional Training Schemes

In addition to using encoder-only MLM training, we experimented with other objectives to train using our Wiktionary and grammar data, but ultimately found no discernible difference in performance. For these experiments, we use an encoder-decoder model, where the decoder is later removed after training. The encoder part is simultaneously trained on MLM as well as the additional objectives, which we now describe.

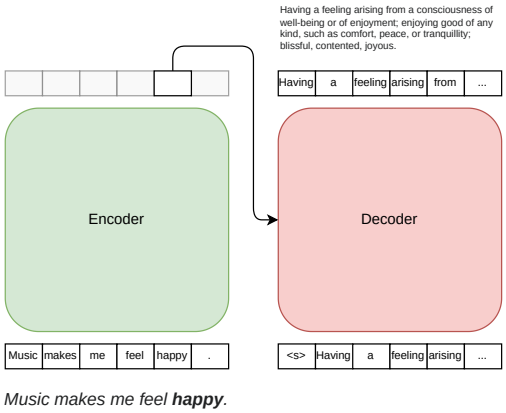

Figure 4: The model layout for training wiktionary.

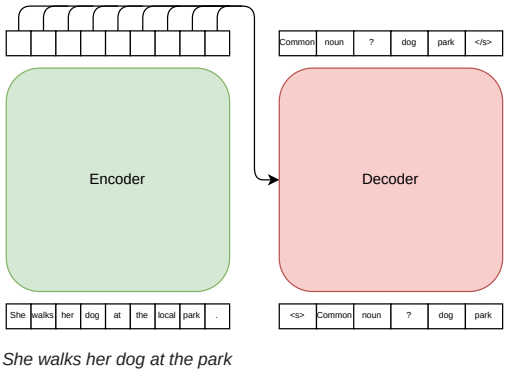

Figure 5: The model layout for training with grammar examples.

Wiktionary Training For each Wiktionary entry, we feed the example as input to the encoder and mark the specific token that corresponded to the target word. For the marked position, we pass this to a separate decoder, which is tasked with generating the definition. This process can be seen in Figure 4.

Grammar Training For the `Gram Gen` data, we feed in a sentence to the encoder, passing its hidden states to the decoder, and prompt the model to answer whether it contains a particular notion, and if that notion corresponds to a particular word or words, which word(s) does it correspond to. The scheme for training is shown in Figure 5.

4 Results

We first discuss the results of our experiments with MLM-only models trained on grammar and lexical data, then we move on to discuss the results of the models with additional training schemes. Finally, we cover the results of our best-performing models that we submitted to the challenge.

4.1 Grammar Results

The results for our best models using grammar data are shown in Table 4. As we can see, adding grammar data appears to help with BLiMP to a limited extent, but hurts performance on all other metrics. The increase in BLiMP is expected, as the BLiMP evaluation necessitates that grammatical sentences are given a lower perplexity than ungrammatical sentences. A lot of the sentences in BLiMP are grammatical, but are very unnatural for a native speaker to read. As such, an excellent source for unnatural sounding yet grammatically correct sentences is a grammar book. This is likely why we see the most improvement from training on those.

The generated data, seeing as it is generated by GPT 3.5, is likely going to reflect the data that GPT itself was trained on. Although we do not know specifically the data that GPT is trained on, it is likely much more representative of "natural" data, rather than these unnaturally constructed sentences that are ubiquitous in BLiMP.

	Half / Half	+ Gram Gen	+ Gram Books
BLiMP	74.2	74.7	**75.4**
Supplement	**63.7**	63.3	61.1
GLUE	**77.1**	75.9	74.7
EWoK	**54.3**	53.0	50.3
Average	**67.3**	66.7	65.4

Table 4: Results of our grammar-informed models.

To further improve BLiMP scores, we expect that including more grammar books or perhaps explicitly prompting an LLM to produce unnatural sounding sentences may be the key. However, we also expect that such data would have a negative impact on GLUE and EWoK. This may simply be an immutable trade-off for low-resource pretrained models.

4.2 Wiktionary Results

We show the results of adding Wiktionary data in Table 5. Unfortunately, adding Wiktionary definitions and examples appears to only hurt performance. We speculate that it might have to do with the structure of Wiktionary entries and how the structure of lexical information is drastically different from other types of training and evaluation data.

	Half / Half	+ Wikt
BLiMP	**74.2**	72.9
Supplement	**63.7**	62.8
GLUE	**77.1**	75.7
EWoK	**54.3**	50.1
Average	**67.3**	65.4

Table 5: Results of adding Wiktionary data.

4.3 Additional Training Schemes Results

	MLM	MLM + Gram	MLM + Wikt
BLiMP	74.2	71.5	**75.7**
Supplement	**63.7**	61.0	59.3
GLUE	**77.1**	75.9	73.4
EWoK	**54.3**	51.1	50.8
Average	**67.3**	64.9	64.8

Table 6: Our models with additional objectives, compared to the MLM-only baseline (i.e. our half/half model).

We show the results of our models with added objectives for Wiktionary definition learning and grammatical notion identification in Table 6. Concerning the grammar objective, we see slightly worse performance overall. Notably, despite BLiMP being an evaluation aimed at gauging understanding of grammaticality, we still see a decrease in the performance.

Ironically, our Wiktionary-based objective increases BLiMP scores. It is unclear why our method for improving semantic understanding increased performance on the grammar benchmark, but there is of course information that can be extracted from word definitions that is useful for parsing grammaticality, such as part of speech information, and even quite literal information about the usage of words (e.g. the definition of "the" starts with "used before a noun phrase...").

Though it does not explain the improvement on BLiMP from our model trained with the Wiktionary objective, we believe that adding an additional objective is the main source of the loss in performance for our additional models. BLiMP (as well as EWoK) is designed such that a model's zero-shot default behavior is to provide a perplexity for a sentence. This is achieved trivially with a model trained on MLM or CLM, but adding another objective means that the hidden states are forced to

	BabyLlama		LTG-BERT		BabyLM	Half / Half	Contr.
	10M	100M	10M	100M	10M	10M	10M
BLiMP	69.8	73.1	60.6	69.2	**74.2**	**74.2**	65.5
Supplement	59.5	60.6	60.8	**66.5**	66.2	63.7	60.3
GLUE	50.7	52.1	48.9	51.9	69.0	**77.1**	76.6
EWoK	50.7	52.1	47.4	51.9	51.8	**54.3**	51.6
Average	57.7	59.5	54.4	59.9	65.3	**67.3**	63.5

Table 7: Final results compared to the baselines.

learn a representation that balances approximating the perplexity with optimizing for whatever the external objective requires. Thus, it is no surprise that the scores for BLiMP and EWoK are lower. This does not necessarily mean that this model is less capable of understanding grammaticality, but this could not be captured by BLiMP. We are not aware of another benchmark that would resolve this issue.

4.4 Submission

In Table 7, we show the overall results for our best models, compared to the baselines. The results from BabyLlama and LTG-BERT are taken from the reported scores from the organizers. The "BabyLM" model is our internal baseline, using the same parameters and training as our other models, but trained on the data provided by the organizers. "Half / Half " is a model trained on a mixture of the provided data and contrastive data, and "Contr." is trained on exclusively contrastive data.

As we can see, our models outperform even the provided models trained on 100M overall. We suspect this is for the same reason as we found last year in Edman and Bylinina (2023), where the models trained on too large of a context size have trouble converging. In terms of the data used, we see that using the contrastive dataset hurts BLiMP performance, but raises GLUE performance. Using a mix is able to capture a best of both worlds, retaining performance on BLiMP while even improving performance on GLUE and EWoK.

5 Conclusion

In this year's BabyLM Challenge, we attempted to buck the trend of administering strategies based on L1 acquisition, having seen little success from such strategies in last year's Challenge. Instead, we hypothesized that L2 acquisition, with more explicit information regarding semantics and syntax, might be what a language model needs. To that end, we also saw limited success. Our strategy of using Wiktionary data did not show any indication of improved output quality. Using grammar information did have a small positive effect on BLiMP scores, though it is unclear whether the grammar itself helped or simply the more diverse data domain.

Nevertheless, our strategy of reducing context size from the previous year was yet again successful at outperforming the baselines, even those with $10\times$ more data used in training. Additionally, using data that includes paraphrases and contrastive pairs helped improve the GLUE scores by a remarkable 8 points. This goes to show that the data chosen for low-resource pretraining can have a profound impact. The study of the exact structure of data that LMs efficiently learn from is a productive future direction, as tentatively shown by our results.

Acknowledgements

We thank BabyLM anonymous reviewers for useful comments. We also thank Oleg Serikov for helpful informal discussions. The work was supported by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101113091) and by the German Research Foundation (DFG; grant FR 2829/7-1).

References

Asmik Brutjan and Karine Brutjan. 2022. *Learn English with short stories. A Textbook with Grammar References for Pre-intermediate and Intermediate Learners.* A. Brutjan.

Lukas Edman and Lisa Bylinina. 2023. Too much information: Keeping training simple for BabyLMs. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 89–97, Singapore. Association for Computational Linguistics.

Sidney Greenbaum and Gerald Nelson. 2009. *An Introduction to English Grammar*. Pearson Education.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Anna Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, Pramod RT, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyurek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Josh Tenenbaum, and Jacob Andreas. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv*.

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*.

Mark Newson. 2006. *Basic English Syntax with Exercises*. Bölcsész Konzorcium.

Eric H Roth and Toni Aberson. 2010. *Compelling Conversations: Questions and Quotations on Timeless Topics: An Engaging ESL Textbook for Advanced Students*. Chimayo Press.

Audrey Jean Thomson and Agnes V Martinet. 2015. *A Practical English Grammar*. New York: Oxford University Press.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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 Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

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.

Laura Wright. 2024. *English Grammar for Literature Students: How to Analyse Literary Texts*. De Gruyter Mouton.

Junlei Zhang, Zhenzhong Lan, and Junxian He. 2021. Contrastive learning of sentence embeddings from scratch. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.