Child-Directed Language Does Not Consistently Boost Syntax Learning in Language Models

Anonymous ACL submission

Abstract

Seminal work by Huebner et al. (2021) showed that language models (LMs) trained on English Child-Directed Language (CDL) can outperform LMs trained on an equal amount of adult-directed text like Wikipedia. However, it remains unclear whether these results generalize across languages, architectures, and evaluation settings. We test this by comparing models trained on CDL vs. Wikipedia across two LM objectives (masked and causal), three languages (English, French, German), and three syntactic minimal pair benchmarks. Our results on these benchmarks show inconsistent benefits of CDL, which in most cases is outperformed by Wikipedia models. We then identify various shortcomings in previous benchmarks, and introduce a novel testing methodology, FIT-CLAMS, which uses a frequency-controlled design to enable balanced comparisons across training corpora. Through minimal pair evaluations and regression analysis we show that training on CDL does *not* yield stronger generalizations for acquiring syntax and highlight the importance of controlling for frequency effects when evaluating syntactic ability.[1]

1 Introduction

The prevailing view in language acquisition research has long held that child-directed language (CDL) is inherently more effective than adult-directed language (ADL) for supporting first language development (Ferguson, 1964; Schick et al., 2022). This has led to the assumption that the way caregivers speak to children is tailored to their developmental needs and functional for language learning.

Motivated by this long-standing assumption, recent computational modeling research has used neural network-based language models (LMs) to test how CDL vs. ADL affect learning in such models (Feng et al., 2024; Mueller and Linzen, 2023; Yedetore et al., 2023). Notably, Huebner et al. (2021) showed that BabyBERTa, a masked LM trained on 5M tokens of child-directed speech transcripts[2], achieves the level of syntactic ability similar to that of a much larger RoBERTa model trained on 30B tokens of ADL (Zhuang et al., 2021).

Despite these encouraging findings, several issues complicate direct comparisons between CDL and ADL in LM training, including the variability in training setups (Cheng et al., 2023; Feng et al., 2024; Qin et al., 2024) and evaluation benchmarks across studies (Warstadt et al., 2020; Huebner et al., 2021; Mueller et al., 2020), as well as the frequent focus on the overall accuracy scores averaged over many syntactic paradigms. Moreover, recent work by Kempe et al. (2024) reveals that the evidence for the facilitatory role of CDL in child language acquisition is scarce and specific to narrow domains, such as prosody and register discrimination, raising concerns about its generalizability. In this light, we believe it is crucial to carefully re-evaluate the benefits of CDL for LM training.

To better understand the specific effects of CDL as training input, we systematically compare LMs trained on CHILDES vs. Wikipedia across two architectures (RoBERTa and GPT-2) and three languages (English, French, and German) on four different benchmarks of minimal pairs. Crucially, we control for lexical frequency effects by introducing **FIT-CLAMS**, a Frequency-Informed Testing (FIT) methodology, which we apply to the CLAMS benchmark (Mueller et al., 2020). The resulting benchmark consists of minimal pairs balanced for subject and verb frequency in the training data, to disentangle true syntactic generalization from mere reliance on high-frequency lexical items present in the training data. We also perform a regression analysis to assess the impact of the distributional

[1] Code and data will be released at `anonymized`.

[2] Throughout this paper, we use the term CDL specifically to refer to transcripts of child-directed speech.

properties of CDL and ADL on the models' confidence in predicting grammaticality.

Our findings challenge the presumed advantage of CDL for syntax learning in LMs, showing that CDL does not consistently yield better performance than ADL. These results underscore the need for a more nuanced understanding of when and how CDL may be beneficial, whether as a source of insights to improve training regimes in large-scale LMs (e.g., data augmentation with variation sets (Haga et al., 2024) or context variation (Xiao et al., 2023)), or as a foundation to explore alternative learning paradigms that more closely mirror the interactive, contextual, and multimodal nature of human language acquisition (Beuls and Van Eecke, 2024; Stöpler et al., 2025).

2 Related Work

An ongoing debate in computational linguistics literature concerns whether CDL offers a measurable advantage over ADL in supporting the acquisition of formal linguistic knowledge in language models, with studies reporting conflicting results; some highlight CDL's benefits for grammatical learning and inductive bias, others find little or no advantage. Among the studies that support the benefits of CDL, a prominent example is Huebner et al. (2021). Their study shows that LMs trained on CHILDES (MacWhinney and Erlbaum, 2000), a database containing transcripts of child–adult conversations, show higher average accuracy on Zorro, a minimal pair benchmark designed by Huebner et al. (2021), compared to models trained on Wikipedia, when strictly controlling for dataset size and model configuration. An even better accuracy is achieved by LMs trained on written language adapted for children, such as AO-Newsela (Xu et al., 2015). Salhan et al. (2024) report similar results in a cross-linguistic setting involving French, German, Chinese, and Japanese. Across all four languages, their baseline RoBERTa-small model trained on CHILDES outperforms models trained on a size-matched Wikipedia corpus when evaluated on minimal-pair benchmarks available for each language. Mueller and Linzen (2023) further support the benefits of CDL by showing that pretraining LMs on simpler input promotes hierarchical generalization in question formation and passivization tasks, even with significantly less data than required by models trained on more complex sources like Wikipedia. Finally, You et al. (2021) leverage a non-contextualized word embedding model (Word2Vec by Mikolov, Chen, Corrado, and Dean, 2013) to show that CDL is optimized for enabling semantic inference through lexical co-occurrence even in the absence of syntactic cues, suggesting that early meaning extraction in humans may be supported by surface-level regularities.

In contrast to studies highlighting the advantages of CDL, Feng et al. (2024) report that models trained only on CDL consistently perform worse than those trained on ADL datasets with higher structural variability and complexity (e.g., Wikipedia, OpenSubtitles, BabyLM Challenge dataset) in both syntactic tasks (like the ones in Zorro) and semantic tasks measuring word similarity. A similar result is reported by Bunzeck et al. (2025), who focus on German language models: while lexical learning tends to improve with the simpler, fragmentary language constructions typical of CDL, syntactic learning benefits from more structurally complex input. Yedetore et al. (2023) further challenge the benefits of CDL by demonstrating that both LSTMs and Transformers trained on CDL input fail to acquire hierarchical rules in yes/no question formation, instead relying on shallow linear generalizations. Finally, beyond text-based models, Gelderloos et al. (2020) train their models on unsegmented speech data using a semantic grounding task and find that whilst child-directed speech may facilitate early learning, models trained on adult-directed speech ultimately generalize more effectively.

In light of conflicting findings on the effectiveness of CDL, our study offers a systematic reassessment of its impact on syntax learning across two model architectures, three languages and multiple benchmarks, including a frequency-controlled one.

3 Method

We train RoBERTa- and GPT-2-style language models from scratch on size-matched corpora of CHILDES and Wikipedia text in English, French and German. To assess their syntactic performance, we evaluate them on a set of existing minimal-pair benchmarks, enabling cross-linguistic and cross-architectural comparisons. Additionally, we propose FIT-CLAMS, a novel evaluation methodology inspired by Mueller et al. (2020), which controls for lexical frequency effects and facilitates more reliable comparisons across datasets.

		Tokens	Avg Sent. Length	Type/Token Ratio (TTR) 1-grams	2-grams	3-grams
EN	CHILDES	4.3 M	6.13	0.005	0.073	0.275
	Wiki		24.13	0.026	0.286	0.680
FR	CHILDES	2.3 M	6.48	0.009	0.089	0.310
	Wiki		37.19	0.019	0.159	0.386
DE	CHILDES	3.8 M	5.61	0.012	0.129	0.424
	Wiki		21.34	0.055	0.379	0.752

Table 1: Descriptive statistics of our training datasets.

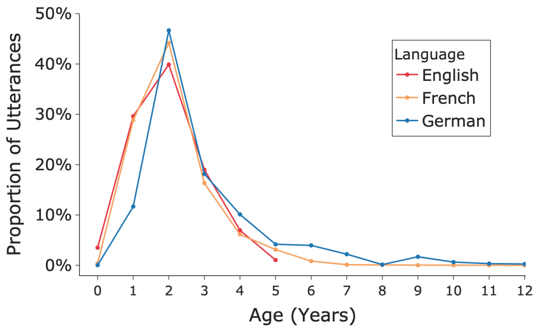


Figure 1: CHILDES age distribution across languages.

3.1 Training Datasets

We choose English for comparability to previous results, and French and German because they are included in the existing CLAMS benchmark (Mueller et al., 2020), enabling consistent cross-linguistic evaluation.[3] While typologically related, these languages provide sufficient variation, particularly in subject–verb agreement, to test the robustness of our findings.

We train models on two data types: CHILDES transcripts and Wikipedia. For English, we use the same data split as Huebner et al. (2021), comprising approximately 5M words of American English CDL, which in their work is referred to as AO-CHILDES (Age-Ordered CHILDES; Huebner and Willits, 2021). The French and German portions are extracted from CHILDES using the `childesr` library in R through the childes-db interface (Sanchez et al., 2019). We keep only adult-to-child utterances, excluding those produced by children. To enable fair comparisons, we sample Wikipedia corpora of matching sizes (measured in terms of whitespace-separated tokens). We exclusively use curated, small-scale corpora from the CHILDES database to ensure a controlled and comparable experimental setup across languages, enabling a targeted examination of CDL—as interactive, infant-oriented register—versus the formal, written style of Wikipedia, despite the availability of larger datasets such as the BabyLM Challenge (Warstadt et al., 2023) and the German BabyLM corpus (Bunzeck et al., 2025).

Summary statistics for our CHILDES and Wikipedia datasets are provided in Table 1. A consistent pattern across languages is that Wikipedia has substantially higher average sentence lengths compared to CHILDES. Additionally, notable disparities in type–token ratios reflect the highly repetitive nature of CDL, at both lexical and phrase level. Further comparisons of the two corpora are presented in Appendix A. As for CHILDES-specific properties, Figure 1 shows that the data is heavily skewed toward the first 2–3 years of life, in all three languages.

[3]CLAMS also includes Hebrew and Russian, but these languages were not selected due to the limited amount of available CHILDES data.

3.2 Models

We evaluate two model architectures: **RoBERTa**, a masked language model (MLM) chosen for consistency with prior CDL vs. ADL studies (Huebner et al., 2021; Salhan et al., 2024) and **GPT-2**, a causal language model (CLM), whose autoregressive objective more closely approximates the incremental nature of human language processing (Goldstein et al., 2022). To ensure a fair comparison, both models share the same architecture: 8 transformer layers, 8 attention heads, an embedding size of 512, and an intermediate feedforward size of 2048.

3.2.1 Training Setup

We train all models from scratch for 100,000 steps using the AdamW optimizer (Loshchilov and Hutter, 2017) with linear scheduling and a warm-up phase of 40,000 steps (during our hyperparameter search, we experimented with various warm-up durations and found that shorter warm-up phases lead to early overfitting, particularly in the case of GPT-2). A learning rate of 0.0001, chosen for producing more stable learning curves, is applied consistently across all experiments. We use 95/5% train/validation split.

4 Evaluation on Existing Benchmarks

Following Huebner et al. (2021) and Salhan et al. (2024), we train separate Byte-Pair Encoding (BPE) tokenizers (Sennrich et al., 2016) for each language and dataset, resulting in distinct vocabular-

	Minimal Pair	#(noun;C)	#(noun;W)	#(verb;C)	#(verb;W)
CLAMS	*the pilot [smiles/*smile]*	75	271	196	21
	*the author next to the guard [laughs/*laugh]*	4	724	161	17
	*the surgeon that admires the guard [is/*are] young*	3	63	86,217	65,369
FIT-CLAMS-C	*the [resident/*residents] awaits*	6	343	2	15
	*the [farmer/*farmers] next to the guards arrives*	247	185	18	117
	*the [daddy/*daddies] that hates the friends thinks*	6,184	5	16,056	254
FIT-CLAMS-W	*the [picker/*pickers] exaggerates*	17	2	2	6
	*the [painter/*painters] in front of the waiter enjoys*	8	177	64	99
	*the [president/*presidents] that admires the speakers works*	46	1,599	2,221	3,923

Table 2: Minimal pair examples for CLAMS and FIT-CLAMS, and the noun and verb frequency across CHILDES (C) and Wikipedia (W) in each dataset.

ies.[4] A vocabulary size of 8,192 tokens is used throughout, following prior developmental studies (Biemiller, 2003) and consistent with earlier work in this area (Salhan et al., 2024). Specifically, research has estimated that the average English-speaking 6-year-old has acquired approximately 5,000–6,000 words (Biemiller, 2003). Although our CHILDES datasets for German and French are not strictly limited to children up to age six, the majority of the data comes from younger children, with relatively fewer samples from older age groups.

4.1 Evaluation Procedure

For evaluation, we report metrics averaged over three random seeds per model configuration. For MLMs, where no overfitting is observed, we use the final checkpoint at step 100,000. For CLMs, where overfitting *is* observed, we select the checkpoint that yields the lowest validation perplexity for each language and training dataset. Full validation perplexities trajectories are provided in Appendix B, along with the list of selected checkpoints for each model and dataset (Table 7).

We assess the syntactic performance of a model by testing whether it assigns a higher probability to the grammatical version in a minimal sentence pair, a well-established paradigm in LM evaluation (Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018). Sentence probabilities are computed using the `minicons` library (Misra, 2022). For CLMs, we use the summed sequence log-probability with BOW correction.[5] For MLMs, we use the likelihood score with a within-word left-to-right masking strategy, which mitigates overestimation of token probabilities in multi-token words (Kauf and Ivanova, 2023).

4.2 Benchmark Description

Several minimal-pair benchmarks have been proposed in the literature to evaluate grammatical learning in models trained on CDL and ADL, most notably BLiMP (Warstadt et al., 2020), Zorro (Huebner et al., 2021), and CLAMS (Mueller et al., 2020).

BLiMP has become the standard benchmark for English, consisting of 67 paradigms representing 12 different linguistic phenomena. It is generated through a semi-automated process where lexical items are systematically varied within manually crafted sentence templates. While carefully controlled, this approach still produces semantically odd or implausible sentences (Vázquez Martínez et al., 2023). Moreover, this benchmark does not account for the vocabulary typical of CDL.

To address this lexical mismatch, Huebner et al. (2021) introduce **Zorro**, a benchmark comprising 23 grammatical paradigms that represent 13 phenomena. Lexical items in Zorro's minimal pairs are selected by manually identifying entire words (never words split into multiple subwords) from the BabyBERTa tokenizer's vocabulary[6] and by counterbalancing word frequency distributions across the five training corpora. While this design enhances lexical compatibility across CDL and ADL training domains, evaluating only on whole words overlooks the fact that models with robust syntactic

[4] Bunzeck and Zarrieß (2025) propose character-level models as a viable alternative for syntax learning, which could be tested in future CDL vs. ADL settings.

[5] Beginning-of-word (BOW) correction adjusts LM scoring by shifting the probability mass of 'ending' a word from the BOW of the next token to the current one (Pimentel and Meister, 2024; Oh and Schuler, 2024).

[6] The BabyBERTa tokenizer is jointly trained on AO-CHILDES, AO-Newsela, and an equally sized portion of Wikipedia-1.

Model	Training Data	BLiMP	Zorro	CLAMS		
				English	French	German
CLM	CHILDES	0.61 ± 0.02	**0.76 ± 0.04**	0.60 ± 0.01	0.64 ± 0.01	0.69 ± 0.03
	Wiki	0.60 ± 0.02	0.68 ± 0.02	**0.71 ± 0.02**	**0.82 ± 0.01**	**0.81 ± 0.01**
MLM	CHILDES	0.59 ± 0.03	0.66 ± 0.05	0.57 ± 0.02	0.59 ± 0.01	0.70 ± 0.01
	Wiki	0.59 ± 0.02	0.65 ± 0.02	**0.64 ± 0.01**	**0.69 ± 0.02**	**0.74 ± 0.01**

Table 3: Model accuracies on BLiMP, Zorro, and CLAMS, averaged across paradigms and model seeds.

understanding should be able to handle structure even when key items are split into subword units, raising concerns about the fairness and broader applicability of this benchmark.

Finally, **CLAMS** extends minimal pair coverage to five languages to enable a comparable cross-lingual evaluation of syntactic learning, but only focuses on the phenomenon of subject–verb agreement (across 7 paradigms). Despite its more limited syntactic scope compared to BLiMP and Zorro, we select CLAMS for our extended analyses, as subject–verb agreement represents a foundational aspect of grammatical ability which is typically acquired early in child language development (Bock and Miller, 1991; Phillips et al., 2011). CLAMS is based on translations of minimal pairs originally created for English by Marvin and Linzen (2018). As stated by these authors, their models showed varied accuracy across specific verbs in the minimal pairs, with frequent ones like *is* reaching 100% accuracy and rarer ones like *swims* only around 60%, likely reflecting frequency effects. To account for such effects as well as ensure cross-linguistic consistency, we introduce in Section 5 a new methodology for constructing minimal pairs inspired by CLAMS, explicitly controlling for both verb and noun frequency across all language conditions.

4.3 Results

The results obtained with our two model architectures, presented in Table 3, are partially consistent with prior findings reported for English (Huebner et al., 2021). In our experiments, both the causal and masked language models trained on CHILDES and Wikipedia perform comparably, with no significant differences in accuracy when tested on BLiMP. On Zorro, the CLM trained on CHILDES outperforms its Wikipedia-trained counterpart, replicating the findings of Feng et al. (2024). For the MLM architecture, the CHILDES-trained model shows only a modest accuracy advantage, smaller than that reported by Huebner et al. (2021).

A more fine-grained analysis of the paradigms is provided in Appendix C, where Table 8–9 indicate that the advantage of CDL is partly driven by grammatical phenomena involving questions. We find that this effect is even more pronounced in CLMs than in MLMs. The trend reflects the prevalence of interrogatives in the CDL data (40% in English, see Appendix A), which may bias models toward better handling of question-related constructions, as already noted by Huebner et al. (2021). To validate our evaluation pipeline and contextualize our results, we also applied our evaluation methodology to the models trained and released by Huebner et al. (2021); details of this analysis are provided in Appendix C.

Focusing on subject–verb agreement, CLAMS results for the three languages are overall consistent across the two model types, but contradict the findings reported in previous work (Salhan et al., 2024). For English, French and German, neither CLM nor MLM demonstrates an advantage when trained on CHILDES compared to Wikipedia, as shown in Table 3.

In summary, results are mixed: models trained on CDL sometimes perform better and sometimes worse than those trained on ADL, depending on the evaluation benchmark. We hypothesize that lexical frequencies may be an important confounder in this type of evaluation, and in the next section we set out to design new minimal pairs that balance the distribution of nouns and verbs representative of each training corpus. As CLMs generally demonstrate higher performance than MLMs, we only focus on CLMs in our subsequent analyses.

5 FIT-CLAMS

When comparing two models (trained on different data sets) on a syntactic evaluation task, we must ensure that any differences in their performance do not stem from the evaluation data being more ‘aligned’ with the training distribution of one model over the other. To that end, we propose a new Frequency-Informed Testing (FIT) evaluation methodology based on CLAMS, through which we generate *two* sets of minimal pairs guided by the distribution of the two training corpora, ensuring a spread of high- and low-frequency items across each corpus.

5.1 Data Creation

Our data creation follows the following four steps:

1. **Vocabulary selection**: We compute the intersection of the vocabularies (*before* applying subword tokenization) from Wikipedia and CHILDES, then select lexical items ensuring that all word forms have been encountered by *both* models during training.

2. **Candidate selection**: Using SpaCy (Honnibal, 2017), we select candidate subjects and verbs by ensuring they have the right part-of-speech and grammatical features. Specifically, we select only animate nouns and limit verbs to present-tense third-person forms in indicative mood. We only keep nouns and verbs that occur in the corpora in both singular and plural form.

3. **Controlling frequency**: To control for lexical frequency effects, nouns and verbs are grouped into 10 frequency bins based on their occurrence in the training data. Frequency binning is done using uniformly spaced bins at a logarithmic scale, to account for the Zipfian distribution of word frequencies. The distribution of nouns and verbs across bins is shown in Appendix D. From each frequency bin, one noun and one verb are manually selected from both the CHILDES and Wikipedia distributions, ensuring semantic compatibility.

4. **Minimal pair creation**: The final minimal pairs are generated adhering to the syntactic templates used by Mueller et al. (2020). We adopt a minimal-pair design in which the critical region—the verb—is held constant across grammatical and ungrammatical conditions, as is also done in BLiMP-NL (Suijkerbuijk et al., 2025). Evaluating model probabilities only at this critical region (while changing the context) avoids confounding effects from differences in subword tokenization.

Following this pipeline, we generate two sets of minimal pairs, one from CHILDES distribution (FIT-CLAMS-C) and one from Wikipedia distribution (FIT-CLAMS-W), forming together the FIT-CLAMS benchmark. In total, we generated 16,400 minimal pairs for English, 4,914 for French, and 10,800 for German across the various paradigms (see Table 14 for detailed counts per paradigm). Examples of our minimal pairs, together with the corresponding noun and verb frequencies in both CHILDES and Wikipedia data, are provided in Table 2.

Model	FIT-CLAMS	EN	FR	DE
CHILDES	FIT-CLAMS-C	0.63	0.78	0.73
	FIT-CLAMS-W	0.63	0.67	0.69
Wiki	FIT-CLAMS-C	**0.76**	**0.84**	**0.82**
	FIT-CLAMS-W	**0.77**	**0.89**	**0.83**

Table 4: Average FIT-CLAMS accuracy on the CLM models. Best scores per dataset are shown in boldface.

5.2 Results

Since our minimal pairs are constructed such that the verb remains constant across the pair, we compute the model's probability for the verb alone, rather than over the entire sentence, as was done for the three previous benchmarks. This approach more directly probes the model's syntactic ability by correctly solving subject–verb agreement, assigning higher probability to the verb form that matches the sentence-initial subject. Each CLM (whether trained on CHILDES or Wikipedia) is evaluated on both sets. Depending on the lexical source, each evaluation set may be considered either in-distribution or out-of-distribution relative to the model's training data.

Table 4 presents average accuracy scores across the seven syntactic paradigms. Overall, we observe that average accuracy on FIT-CLAMS increases for both models (trained on CHILDES and Wikipedia) compared to their performance on CLAMS (see Table 3). This increase can be explained by the fact that in the original CLAMS dataset, some minimal pairs contain tokens that are not observed at training time, which is not the case for FIT-CLAMS. These results also reveal that, as expected, models generally perform better on minimal pairs constructed with in-distribution lexical items than with out-of-distribution ones.

Importantly, the most pronounced contrast we see is still the one between the models trained on CHILDES (first two rows) vs. Wikipedia (last two

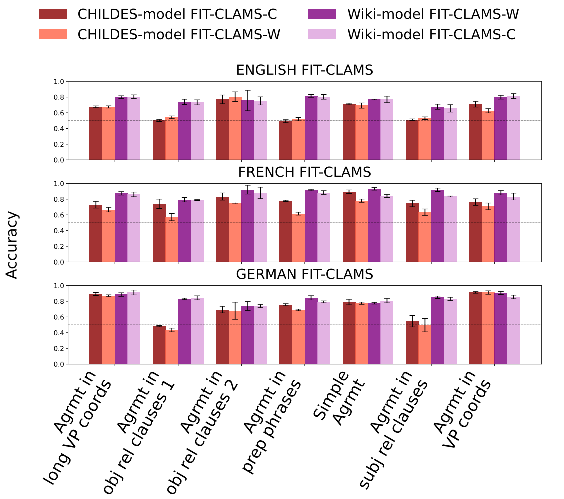


Figure 2: Accuracy of our models on the individual paradigms in the new set of minimal pairs, FIT-CLAMS.

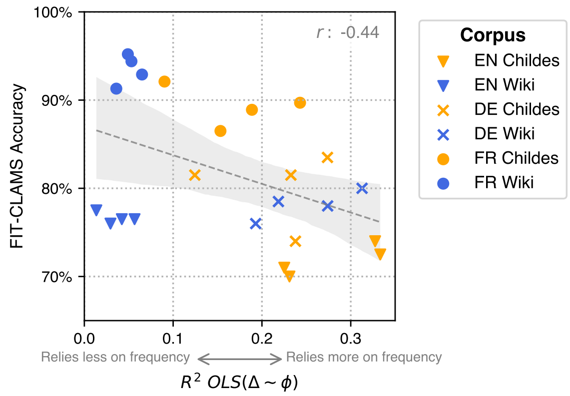


Figure 3: Relation between LM accuracy on FIT-CLAMS and proportion of variance (R^2) explained by the OLS regression fitted on lexical frequency factors. The lower the R^2 is, the less the LM's behavior is driven by lexical frequency. Each LM configuration is represented by four data points: three individual LMs (random seeds) and the average of the three.

rows): the latter consistently outperform the former across all languages on the subject–verb agreement task. Thus, even when strictly controlling for lexical frequency, models trained on Wikipedia continue to show a systematic advantage, underscoring the benefits of training on larger and more diverse textual resources for developing robust syntactic ability.

6 Regression Analysis

To further investigate how training data shapes model behavior, we conduct a linear regression analysis examining whether and how the presence of specific lexical items in the training data influences model performance. A model that builds up a robust representation of number agreement will be better able to generalize to infrequent constructions, without relying on memorization (Lakretz et al., 2019; Patil et al., 2024). We focus on the Simple Agreement paradigm, as it is the most straightforward paradigm to connect the frequency of occurrence of individual words in the training data to subsequent model performance. Specifically, we assess how unigram frequency of critical lexical items—the subject and the verb—affects the model's preference for grammatical over ungrammatical sentences. This controlled setup allows us to isolate frequency effects and compare the degree to which models trained on CDL and ADL generalize beyond lexical co-occurrence patterns.

We fit ordinary least squares (OLS) regressions to the training data (CHILDES or Wikipedia in three different languages) and the probabilities generated by LMs. Specifically, the *dependent variable* used in the regression analysis is the ΔP-*score*, defined as the difference between the probability assigned by the model to the verb in a grammatical and ungrammatical context:

$$\Delta P(v|c) = \log P(v|c^+) - \log P(v|c^-)$$

for verb v in a grammatical (c^+) and ungrammatical context (c^-): e.g., *The boy walks* vs. **The boys walks*. As *independent variables*, we use the log-frequencies of (1) the verb, (2) the grammatical subject noun, and (3) the ungrammatical subject noun, fitting a single multivariate model on all variables. Lexical frequency values are extracted from the corpus used to train the respective model (either CHILDES or Wikipedia). All predictor variables are standardized using z-score normalization.

To investigate the impact of lexical frequency, we examine the relationship between the fit (i.e., R^2) of the OLS regression and the LMs' accuracy on the FIT-CLAMS data. Our hypothesis is that the predictions of an LM will be less driven by frequency if it generalizes well beyond the sentences it saw during training, and as such the OLS will lead to a *lower* R^2 score. The results in Figure 3 reveal a strong negative correlation between R^2 and accuracy ($r = -0.44$, $p = 0.03$): the best-performing LM (trained on French Wikipedia) yields the lowest R^2, whereas the worst-performing LM (trained on English CHILDES) yields the highest.

For both French and English, models trained on Wikipedia obtain a higher accuracy and lower R^2

than those trained on CHILDES; for German this pattern is less clear. We hypothesize this is partly driven by the Type/Token Ratio (TTR), which is the lowest for English and French CHILDES (see Table 1). Generalization in LMs is driven by *compression*: by being forced to build up representations for a wide range of inputs in a bounded representation space, models have to form abstractions that have been shown to align with linguistic concepts (Tishby and Zaslavsky, 2015; Tenney et al., 2019; Wei et al., 2021). Training models on low-TTR data, therefore, leads to a weaker generalization than on high-TTR data, since a model trained on low-TTR data can rely more on memorization. We leave a more detailed exploration of such factors open for future work.

7 Discussion and Conclusions

This study examined whether language models trained on CDL can match or surpass the syntactic ability of models trained on size-matched ADL data. We trained RoBERTa- and GPT-2-based models in English, French, and German and evaluated them on multiple minimal pair benchmarks as well as on the newly introduced, frequency-controlled FIT-CLAMS. The results show that models trained on CHILDES perform consistently worse than models trained on Wikipedia, across languages, architectures, and evaluation settings.

Our regression analysis further revealed a modest negative correlation between model accuracy and variables based on lexical frequency, indicating that stronger models rely less on surface-level patterns of lexical co-occurrence. This trend holds for English and French models, but not clearly for German, pointing to possible language-specific effects.

When interpreting these findings through the lens of language acquisition, it is important to consider the limitations of the training paradigm we used. Our models are trained in artificial conditions that diverge substantially from the way humans acquire language. Unlike children, these models are exposed to static datasets without any form of interaction, feedback, or communicative pressure. Additionally, the learning process is not incremental or developmentally grounded, the vocabulary is extracted from the entire corpus at once when the tokenizer is trained, and the models operate without cognitive constraints or working memory limitations. These discrepancies highlight an important gap between current computational learning frameworks and the dynamics of natural language acquisition.

Rather than completely dismissing CDL, we contend that it should be recontextualized and rigorously tested within frameworks that better resemble human language learning processes. CDL might hold particular promise when integrated into models that simulate interactive, situated communication (Beuls and Van Eecke, 2024; Stöpler et al., 2025), shifting the focus toward the communicative and contextual factors essential to language acquisition, which are absent in static text-based training regimes. Moreover, LM experiments can still contribute significantly to the study of human language acquisition (Warstadt and Bowman, 2022; Pannitto and Herbelot, 2022; Portelance and Jasbi, 2024), where the benefits of CDL remain poorly understood (Kempe et al., 2024), by helping to uncover specific properties of CDL that make it particularly suitable for specific kinds of learning outcomes. For instance, scaling up experiments like those of You et al. (2021) could provide valuable insights into various aspects of language acquisition, such as morphological, syntactic, and semantic development.

Finally, rather than serving solely as pretraining data, CDL, together with insights from the language acquisition literature (Kempe et al., 2024), can inspire the design of inductive biases and data augmentation strategies, such as context variation (Xiao et al., 2023) or variation sets (Haga et al., 2024), with the practical aim of improving generalization or enabling more data-efficient learning in models trained on the standard adult-directed text corpora that are used in NLP applications.

In conclusion, although conventional training on CDL does not currently improve syntactic learning in LMs, we maintain that it remains a valuable resource deserving further investigation. Future work should prioritize CDL's integration within cognitively and interactively grounded frameworks, while also exploring how its distinctive characteristics can inform the development of more effective model architectures and training methodologies.

Limitations

This work does not explicitly account for certain grammatical inconsistencies characteristic of child-directed language, such as the frequent use of infinitive verb forms in contexts where a third- or first-person singular subject is intended, resulting in subject–verb agreement violations. Such errors, which have been systematically mapped for English in an extensive taxonomy by Nikolaus et al. (2024), may introduce noise into the expression of subject–verb agreement. We hypothesize that these properties of CDL could affect the grammatical learning of this syntactic phenomenon. Future experiments could explore whether removing or correcting these occurrences in the training data improves model performance on subject–verb agreement tasks. Another limitation concerns the restricted syntactic scope of our regression analysis, which is limited to simple cases of subject–verb agreement. More structurally complex agreement configurations, such as those involving long-distance dependencies, coordination structures, or prepositional phrases, are not included in the current regressions. In future work, we plan to broaden the analysis to these more challenging constructions to examine whether surface-level factors like lexical frequency continue to influence model performance, and how these effects may differ between CHILDES- and Wikipedia-trained models.

The lexical diversity of our regression setup also imposes constraints on the generalizability of our findings. Each grammatical item (verb or noun) is represented by at most 10 lexical instances per language, with as few as 7 for French verbs. Expanding this set to include a broader and more representative frequency distribution would allow for more robust and precise estimates of how lexical frequency relates to syntactic generalization.

Finally, the construction of our FIT-CLAMS benchmark involved manual selection of animate nouns and semantically compatible verbs shared across CDL and Wikipedia corpora. While this ensured controlled and interpretable comparisons, it limits scalability. In future work, automating this process, could facilitate broader and more flexible evaluations.

References

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, 50(3):1277–1311.

Andrew Biemiller. 2003. Vocabulary: Needed if more children are to read well. *Reading Psychology*, 24(3-4):323–335.

Kathryn Bock and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in german babylms? *arXiv preprint arXiv:2503.11593*.

Bastian Bunzeck and Sina Zarrieß. 2025. Subword models struggle with word learning, but surprisal hides it. *arXiv preprint arXiv:2502.12835*.

Ziling Cheng, Rahul Aralikatte, Ian Porada, Cesare Spinoso-Di Piano, and Jackie CK Cheung. 2023. Mcgill babylm shared task submission: The effects of data formatting and structural biases. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 207–220.

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 A. Ferguson. 1964. Baby talk in six languages. *American Anthropologist*, 66(6_PART2):103–114.

Lieke Gelderloos, Grzegorz Chrupała, and Afra Alishahi. 2020. Learning to understand child-directed and adult-directed speech. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1–6, Online. Association for Computational Linguistics.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. BabyLM challenge: Exploring the effect of variation sets on language model training efficiency. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 252–261, Miami, FL, USA. Association for Computational Linguistics.

Matthew Honnibal. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *(No Title)*.

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.

Philip A Huebner and Jon A Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In *Psychology of learning and motivation*, volume 75, pages 279–331. Elsevier.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Vera Kempe, Mitsuhiko Ota, and Sonja Schaeffler. 2024. Does child-directed speech facilitate language development in all domains? a study space analysis of the existing evidence.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. 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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Brian MacWhinney and Lawrence Erlbaum. 2000. *The CHILDES project . Volume I: Tools for analyzing talk: Transcription format and programs*. Mahwah, NJ: Lawrence Erlbaum.

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.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *CoRR*, abs/2203.13112.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Annual Meeting of the 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.

Mitja Nikolaus, Abhishek Agrawal, Petros Kaklamanis, Alex Warstadt, and Abdellzhah Fourtassi. 2024. Automatic annotation of grammaticality in child-caregiver conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1832–1844, Torino, Italia. ELRA and ICCL.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

Ludovica Pannitto and Aurelie Herbelot. 2022. Can recurrent neural networks validate usage-based theories of grammar acquisition? *Frontiers in Psychology*, Volume 13 - 2022.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Colin Phillips, Matthew W. Wagers, and Ellen F. Lau. 2011. *5: Grammatical Illusions and Selective Fallibility in Real-Time Language Comprehension*, pages 147 – 180. Brill, Leiden, The Netherlands.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Eva Portelance and Masoud Jasbi. 2024. The roles of neural networks in language acquisition. *Language and Linguistics Compass*, 18(6):e70001.

Yulu Qin, Wentao Wang, and Brenden M Lake. 2024. A systematic investigation of learnability from single child linguistic input. *arXiv preprint arXiv:2402.07899*.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 174–188, Miami, FL, USA. Association for Computational Linguistics.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51:1928–1941.

Johanna Schick, Caroline Fryns, Franziska Wegdell, Marion Laporte, Klaus Zuberbühler, Carel van Schaik, Simon Townsend, and Sabine Stoll. 2022. The function and evolution of child-directed communication. *PLoS Biology*, 20(5):e3001630.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Lennart Stöpler, Rufat Asadli, Mitja Nikolaus, Ryan Cotterell, and Alex Warstadt. 2025. Towards developmentally plausible rewards: Communicative success as a learning signal for interactive language models. *arXiv preprint arXiv:2505.05970*.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–39.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5.

Héctor Javier Vázquez Martínez, Annika Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. 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, et al. 2023. Proceedings of the babylm challenge at the 27th conference on computational natural language learning. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

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.

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.

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.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2023. Towards more human-like language models based on contextualizer pretraining strategy. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 317–326, Singapore. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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.

Guanghao You, Balthasar Bickel, Moritz M Daum, and Sabine Stoll. 2021. Child-directed speech is optimized for syntax-free semantic inference. *Scientific Reports*, 11(1):16527.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Training Corpora

As detailed in the main text, a subset of Wikipedia is selected for each language to closely match the token count of the corresponding CHILDES data. For English, we follow Huebner et al. (2021) and used `wikipedia1.txt` from their repository; for German, the corpus `gwlms/dewiki-20230701-nltk-corpus` was employed; and for French, we rely on `asi/wikitext_fr`. We report the differences between the two data types in the three target languages in terms of word frequency in Figure 4 and sentence length in Figure 5. Additionally, as mentioned in the main text, since the age range covered by the CHILDES corpus varies across languages, in Figure 6 we display the total number of utterances directed at children of different ages for each CHILDES split. To generate the bins shown in Figure 4, we use the same strategy adopted for the new evaluation methodology described in Section 5, where the binning is done using uniformly spaced bins at a logarithmic scale, to account for the Zipfian nature of word frequencies.

Moreover, Table 5 provides a quantitative summary of the proportion of sentences classified as interrogatives in the two datasets. It clearly shows that interrogative sentences are substantially more frequent in CDL compared to Wikipedia.

Language	CHILDES	Wikipedia
English	39.84%	0.07%
French	31.28%	0.28%
German	28.93%	0.09%

Table 5: Comparison of interrogative clauses in CHILDES and Wikipedia datasets across languages.

B Models Details

Table 6 summarizes the configuration of the MLM and CLM models used in our experiments. We align the hyperparameters as closely as possible between the two architectures.

Figure 7 presents the validation perplexity curves for MLM and CLM models trained on CHILDES and Wikipedia corpora across English, French and German. A clear pattern of earlier overfitting emerges for CLMs trained on CHILDES, with validation perplexity increasing after fewer training steps compared to their Wikipedia-trained counterparts.

Table 7 reports the CLM checkpoints selected for each language and dataset based on validation perplexity before overfitting, which we used for evaluation on both the existing benchmarks and our FIT-CLAMS dataset.

Hyperparameter	MLM	CLM
Architecture	RoBERTa	GPT-2
Layers	8	8
Attention heads	8	8
Intermediate size	2048	2048
Max seq. length	512	512
Objective	Masked LM	Causal LM
Total parameters	12.7M	14.8M
Learning Rate	0.0001	0.0001
lr_scheduler_type	linear	linear
Training Batch Size	16	16
Evaluation Batch Size	16	16
Gradient Accumulation Step	2	2

Table 6: Comparison of CLM and MLM model configurations.

Language	CHILDES	Wikipedia
EN	ckpt-48000	ckpt-64000
FR	ckpt-36000	ckpt-44000
DE	ckpt-48000	ckpt-64000

Table 7: Best-performing CLM checkpoints per language and dataset that we used for evaluation on existing benchmarks and on FIT-CLAMS.

C Accuracy Results of CLMs and MLMs on existing benchmarks

Models are evaluated on each benchmark across 19 training checkpoints: 10 selected from the first 10% of training steps and 9 from the remaining 90%. This selection strategy allows for a more detailed examination of the early-stage learning trajectories. Figure 8 refers to Zorro accuracy learning curves across steps for our CHILDES- and Wikipedia-models trained on English. Instead, Figure 9- 10- 11 show the accuracy learning curves on CLAMS of our CLMs trained on the three language of interest.

Tables 8- 9 report the accuracies of the two models types (CLM and MLM) trained on the two different datasets (CHILDES vs Wikipedia for each paradigm targeted in Zorro. Paradigms involving questions are highlighted in bold and with color. For CLMs, 7 out of 10 paradigms where CHILDES outperforms Wikipedia include questions, whereas the advantage for question-related paradigms is less pronounced in the case of the MLMs (only 6 out of 13).

Figure 12 illustrates the performance of our two model architectures—CLM and MLM—when evaluated on the CLAMS benchmark, providing

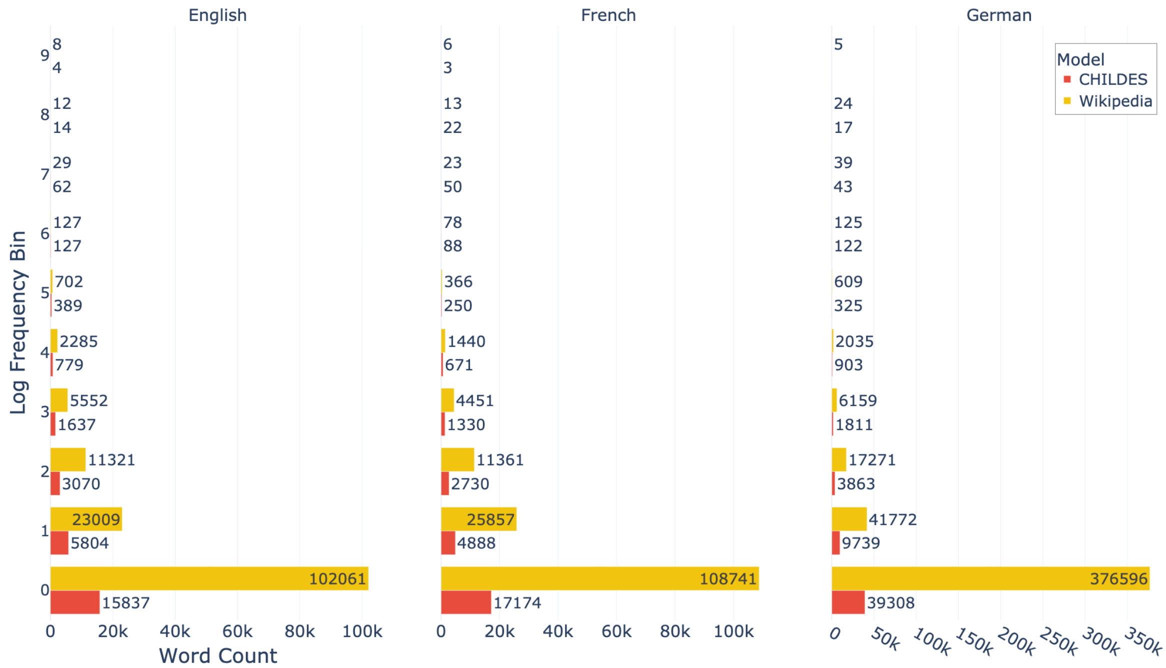


Figure 4: Word Frequency Distribution (CHILDES vs Wikipedia) ascross languages.

a visual summary of their accuracy across languages and conditions. The highest accuracy is observed in the simple agreement paradigm— the least complex, involving only an article, a noun and a verb. Although one might expect models trained on CHILDES, with its simpler syntactic structures, to perform better here, those trained on Wikipedia achieve higher accuracy across all languages. As agreement complexity increases, overall performance declines, yet the Wikipedia-trained models continue to hold an advantage, except in the agreement within relative clauses.

C.1 Testing Previous Models from the Literature

To validate our evaluation pipeline, we applied it to models released by Huebner et al. (2021), allowing for a direct comparison with existing findings in the literature. Specifically, we evaluated two versions of their model trained on AO-CHILDES: one that uses no unmasking probability, and the other that uses the standard unmasking probability value typically employed in MLM training. In both cases, the average accuracies across all paradigms on the Zorro benchmark diverge from the results reported in their original paper. The first model (unmasking probability = 0) achieves an average accuracy of 66%, while the second (standard unmasking probability) scores 68%.

D FIT-CLAMS Minimal Pairs Generation

Curation pipeline details:

- The shared vocabularies extracted for lexical selection include 15,502 tokens in English, 9,354 in French, and 20,366 in German. Frequency distributions are calculated using SpaCy's `en_core_web_sm`, `fr_core_web_sm`, and `de_core_web_sm` pipelines. For noun frequency analysis, singular and plural forms are counted separately. In German, case information is explicitly used to retain only nouns marked as nominative. For verbs, English selection includes forms tagged as `VB`, `VBP`, and `VBZ`, recognizing that these categories respectively capture infinitives, non-third-person present forms, and third-person singular present forms. In contrast, French and German verb selection involves additional morphological constraints: verbs must be third person and present tense, and forms in the subjunctive or conditional moods are excluded.

- As mentioned in the main body, the binning of candidate nouns and verbs into ten frequency categories was performed using a logarithmic scale. The bin edges were defined using log-spaced intervals between the minimum and

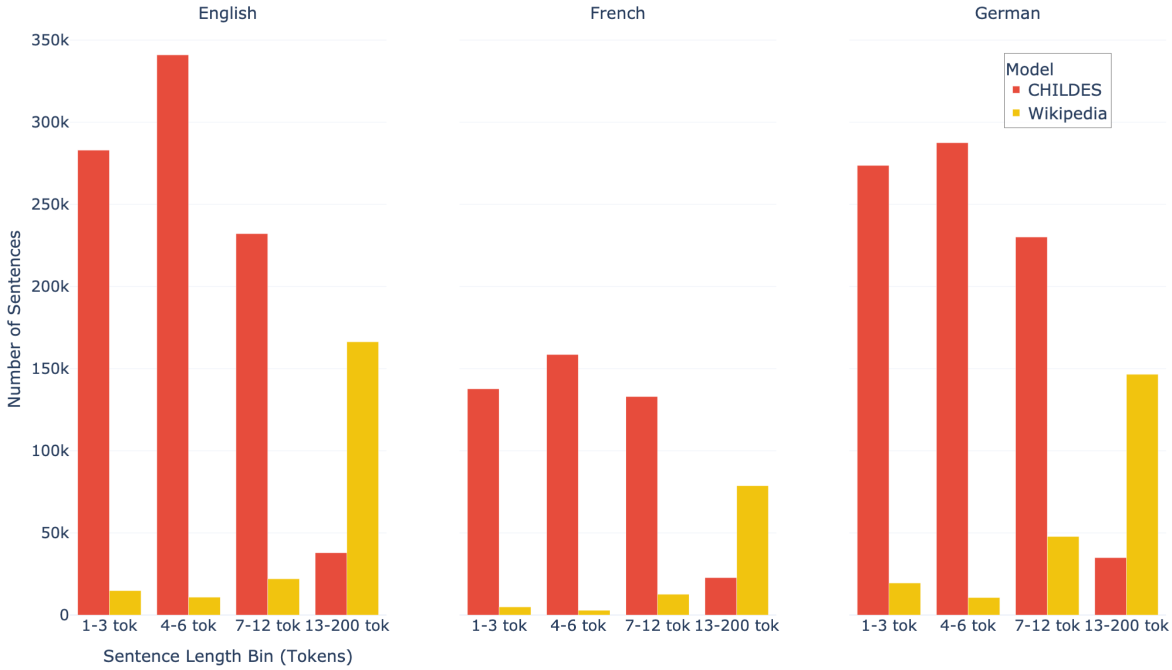


Figure 5: Sentence Length Distribution (CHILDES vs Wikipedia) across languages and data types

maximum frequencies observed across the dataset. To visualize the distribution of nouns and verbs across the bins, reference can be made to the histograms in Figure 13.

- For English and German, a total of 10 nouns and 10 verbs per dataset are retained. For French, due to constraints in the shared vocabulary, the final selection includes 9 nouns and 7 verbs. Additionally, two extra nouns per dataset are selected to serve as objects in the relative clause paradigms. The complete lists of selected lexical items for each language, along with their corresponding frequencies, are reported in Table 10. It is also important to note that the relative clause paradigms include the verb within the relative clause itself. These verbs are adapted from CLAMS, with exclusions made for items not present in the shared vocabulary between CHILDES and Wikipedia. The final set of relative clause verbs used in our study is provided in Table 13. Furthermore, the prepositions used in the prepositional phrase paradigms are also drawn from the multilingual CLAMS version (Table 12). For paradigms involving long-distance dependencies within verb phrase coordination, the CLAMS minimal pairs include attractor nouns following both verbs in the coordinated structure. In our adaptation, we manually construct semantically appropriate fillers for each verb, without explicitly controlling for the frequency of the inserted lexical items.

All generated sentences in English and French have been manually reviewed by the authors, who have linguistic expertise in these two languages, while the German sentences were validated by a native speaker to ensure grammaticality and naturalness.

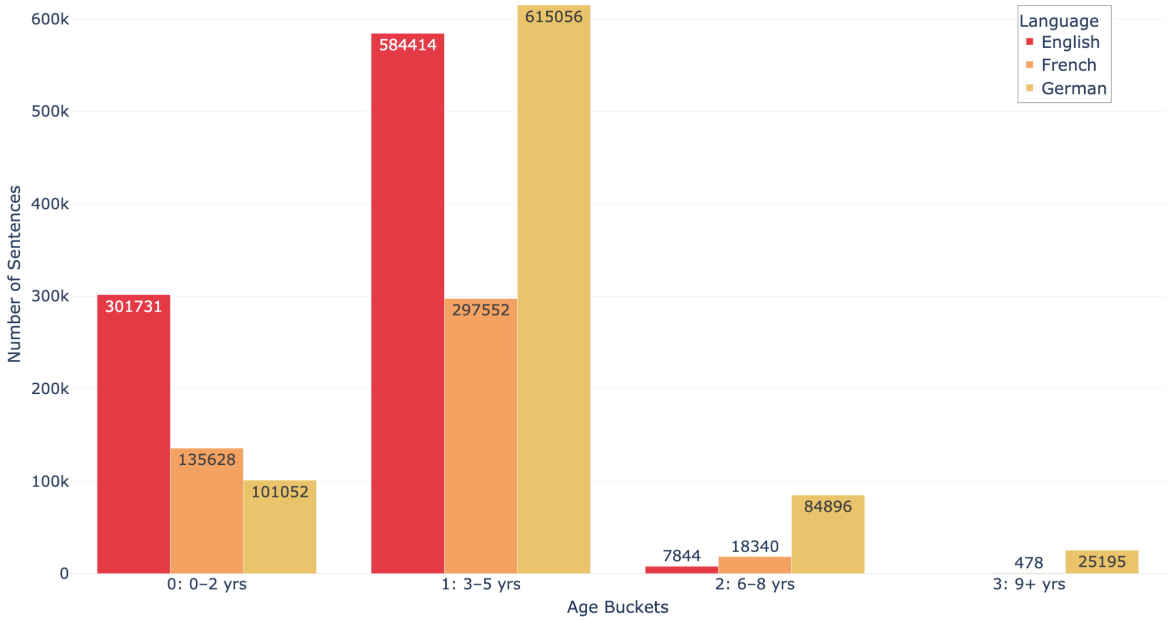


Figure 6: Sentence Count per Age Group in the three CHILDES datasets

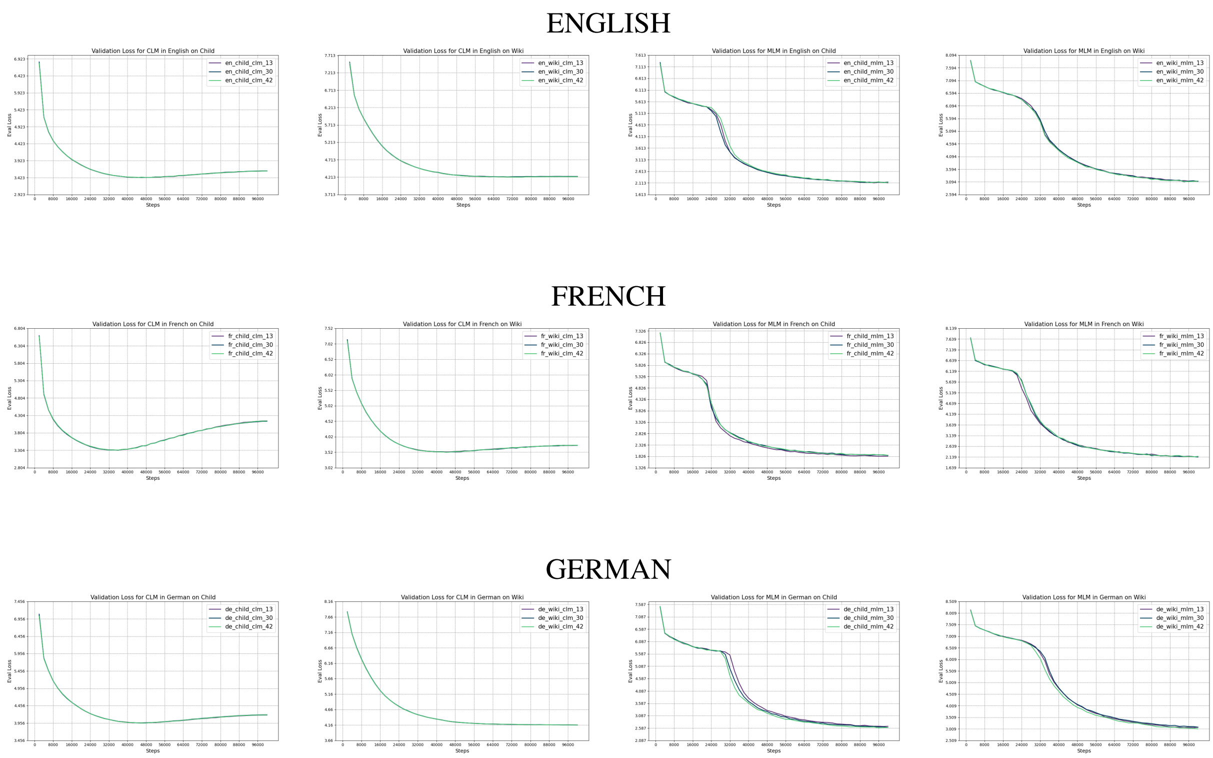


Figure 7: Validation perplexity curves for CLM and MLM models trained on CHILDES and Wikipedia corpora across English, German, and French.

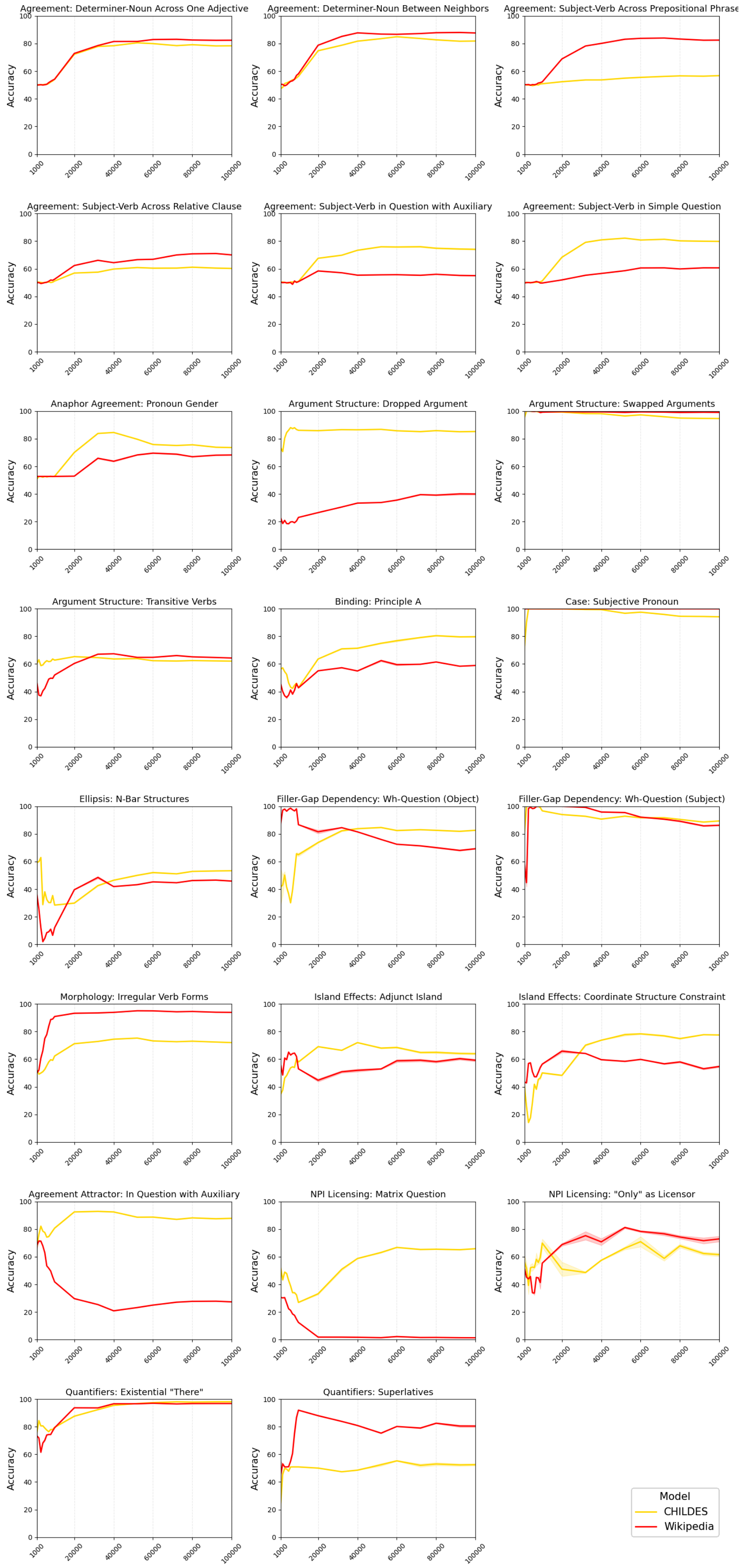


Figure 8: CLM models' accuracy curves on Zorro.

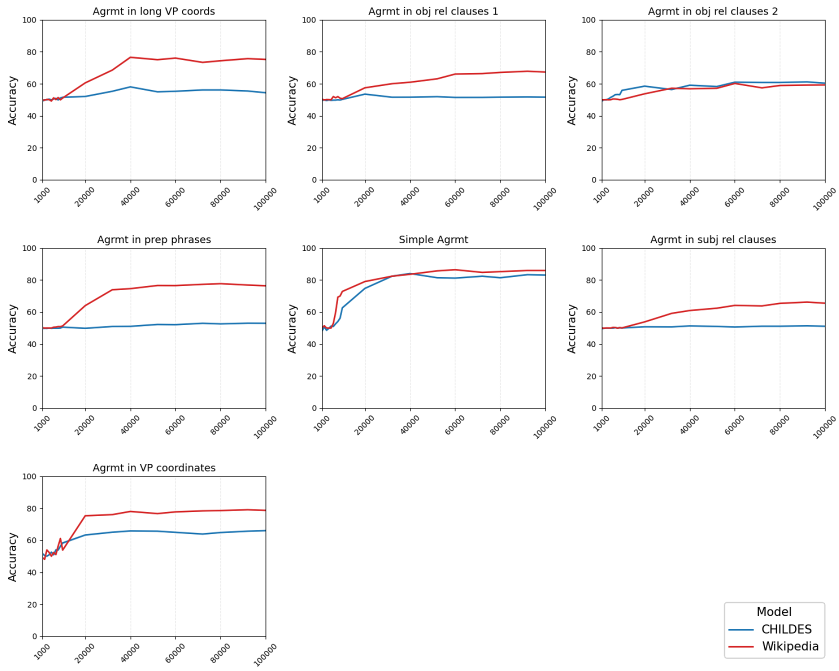


Figure 9: English CLM models' accuracy curves on CLAMS.

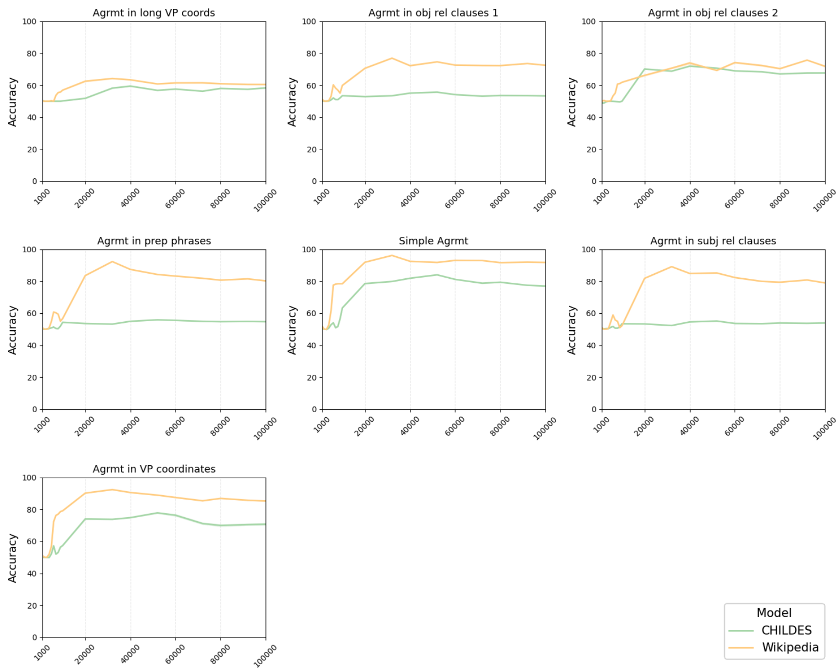


Figure 10: French CLM models' accuracy curves on CLAMS.

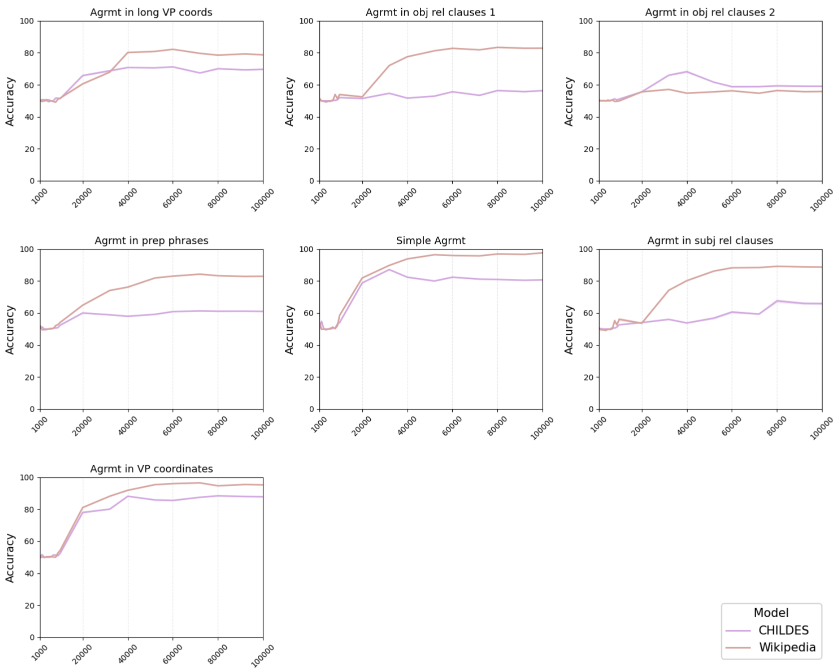


Figure 11: German CLM models' accuracy curves on CLAMS.

Paradigm	CHILDES	Wiki
agr_det_noun_across_1_adj	0.813	**0.836**
agr_det_noun_between_neighbors	0.846	**0.888**
agreement_subj_verb_in_q_with_aux	**0.743**	0.556
agr_subj_verb_across_prep_phr	0.546	**0.836**
agr_subj_verb_across_relclause	0.606	**0.667**
agr_subj_verb_in_simple_q	**0.799**	0.594
anaphor_agreement_pronoun_gender	**0.823**	0.681
arg_structure_dropped_arg	**0.861**	0.353
arg_structure_swapped_args	0.973	**0.991**
arg_structure_transitive	0.626	**0.658**
binding_principle_a	**0.771**	0.608
case_subj_pronoun	0.983	**1.000**
ellipsis_n_bar	0.475	**0.477**
filler_gap_wh_q_object	**0.834**	0.750
filler_gap_wh_q_subject	0.916	**0.931**
irregular_verb	0.735	**0.944**
island_effects_adjunct_island	**0.665**	0.569
island_effects_coord_constraint	**0.765**	0.579
local_attractor_in_q_aux	**0.912**	0.249
npi_licensing_matrix_question	**0.648**	0.021
npi_licensing_only_npi_lic	0.719	**0.721**
quantifiers_existential_there	0.957	**0.975**
quantifiers_superlative	0.496	**0.829**

Table 8: CLM scores on Zorro subparadigms. Question-related paradigms are emphasized with boldface and deeper highlighting.

Paradigm	CHILDES	Wiki
agr_det_noun_across_1_adj	0.664	0.827
agr_det_noun_between_neighbors	0.726	0.900
agreement_subj_verb_in_q_with_aux	**0.603**	0.579
agr_subj_verb_across_prep_phr	0.548	**0.717**
agr_subj_verb_across_relclause	0.559	0.643
agr_subj_verb_in_simple_q	**0.654**	0.556
anaphor_agreement_pronoun_gender	**0.864**	0.667
arg_structure_dropped_arg	**0.550**	0.362
arg_structure_swapped_args	**0.705**	0.578
arg_structure_transitive	0.527	**0.542**
binding_principle_a	**0.805**	0.771
case_subj_pronoun	**0.862**	0.817
ellipsis_n_bar	**0.671**	0.392
filler_gap_wh_q_object	**0.852**	0.784
filler_gap_wh_q_subject	0.828	**0.980**
irregular_verb	0.634	0.921
island_effects_adjunct_island	**0.612**	0.553
island_effects_coord_constraint	0.572	**0.833**
local_attractor_in_q_aux	**0.727**	0.250
npi_licensing_matrix_question	**0.260**	0.026
npi_licensing_only_npi_lic	0.707	**0.802**
quantifiers_existential_there	**0.970**	0.937
quantifiers_superlative	0.376	**0.628**

Table 9: MLM scores on Zorro subparadigms. Question-related paradigms are emphasized with boldface and deeper highlighting.

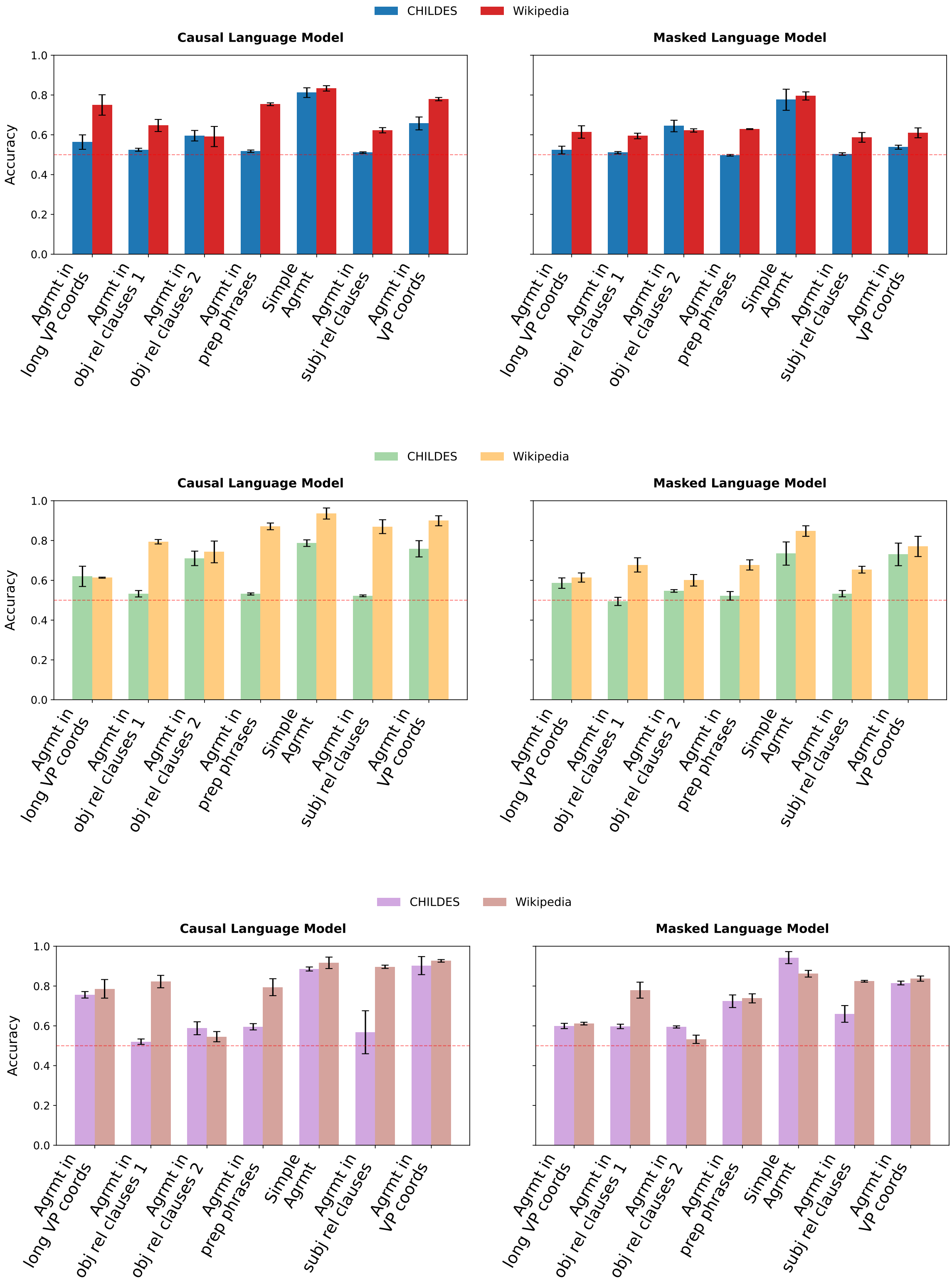


Figure 12: Accuracy scores per paradigm for CLM and MLM across languages on CLAMS.

Table 10: Selected Nouns (used as Subjects) and Verbs in the three languages from CHILDES and Wikipedia Distributions.

EN Nouns	Bin	Freq	Df
roommate, roommates	0	2	CHI
resident, residents	1	6	CHI
librarian, librarians	2	15	CHI
officer, officers	3	40	CHI
toddler, toddlers	4	97	CHI
farmer, farmers	5	271	CHI
policeman, policemen	6	421	CHI
doctor, doctors	7	754	CHI
man, men	8	2373	CHI
daddy, daddies	9	7720	CHI

EN Nouns	Bin	Freq	Df
picker, pickers	0	2	Wiki
harvester, harvesters	0	3	Wiki
fireman, firemen	2	11	Wiki
superhero, superheroes	3	31	Wiki
explorer, explorers	4	80	Wiki
painter, painters	5	179	Wiki
parent, parents	6	394	Wiki
writer, writers	7	683	Wiki
president, presidents	8	1635	Wiki
group, groups	9	3419	Wiki

EN Verbs	Bin	Freq	Long VP	Df
awaits, await	0	2	*the guests*	CHI
complains, complain	1	9	*about the noise*	CHI
arrives, arrive	2	18	*at the station*	CHI
disappears, disappear	3	44	*from the scene*	CHI
bows, bow	4	267	*to the king*	CHI
hides, hide	5	442	*from the chicken*	CHI
leaves, leave	6	1968	*the room*	CHI
sits, sit	7	4651	*in the car*	CHI
thinks, think	8	16240	*about the trip*	CHI
goes, go	9	30425	*to the new store*	CHI

EN Verbs	Bin	Freq	Long VP	Df
grinds, grind	0	4	*the coffee beans*	Wiki
exaggerates, exaggerate	1	6	*with laughs*	Wiki
screams, scream	2	14	*very loudly*	Wiki
swims,swim	3	33	*in the pool*	Wiki
enjoys, enjoy	4	99	*the company of friends*	Wiki
draws, draw	5	230	*a nice picture*	Wiki
rests, rest	6	568	*on the couch*	Wiki
runs, run	7	1093	*at the park*	Wiki
plays,play	7	1375	*with the toys*	Wiki
works, work	8	3929	*on a new project*	Wiki

FR Nouns	Bin	Freq	Df
visiteur, visiteurs	0	3	CHI
joueur, joueurs	1	8	CHI
patient, patientes	2	12	CHI
capitaine, capitaines	3	40	CHI
homme, hommes	4	85	CHI
pompier, pompiers	5	191	CHI
dame, dames	6	351	CHI
enfant, enfants	7	753	CHI
lapin, lapins	8	1050	CHI

FR Nouns	Bin	Freq	Df
gamin, gamins	0	3	Wiki
vilaine, vilaines	3	18	Wiki
cuisinier, cuisiniers	3	19	Wiki
avocat, avocats	4	56	Wiki
pilote, pilotes	6	170	Wiki
lecteur, lecteurs	6	174	Wiki
prince, princes	7	469	Wiki
personnage, personnages	8	1045	Wiki
groupe, groupes	9	1906	Wiki

FR Verbs	Bin	Freq	Long VP	Df
poursuit, poursuivent	0	5	*une nouvelle missio*	CHI
grandit, grandissent	1	19	*très rapidement*	CHI
apprend, apprennent	3	72	*une nouvelle histoire*	CHI
descend, descendent	4	197	*les escaliers de la maison*	CHI
attend, attendent	4	296	*le repas chaud*	CHI
arrive, arrivent	6	1078	*au lieu de rendez-vous*	CHI
met, mettent	7	2207	*la nappe sur la table*	CHI

FR Verbs	Bin	Freq	Long VP	Df
casse, cassent	1	18	*le verre*	Wiki
rentre, rentrent	3	62	*dans la chambre*	Wiki
continue, continuent	5	223	*sur la route*	Wiki
suit, suivent	6	348	*le long chemin*	Wiki
rend, rendent	6	406	*le stylo à sa maman*	Wiki
va, vont	7	682	*au marché*	Wiki
permet, permettent	8	1149	*l'accès aux escaliers*	Wiki

DE Nouns	Bin	Freq	Df
feind, feinde	0	4	CHI
architekt, architekten	0	4	CHI
präsident, präsidenten	1	6	CHI
kollege, kollegen	2	20	CHI
ingenieur, ingenieure	3	26	CHI
sohn, söhne	4	106	CHI
arzt, ärzte	5	223	CHI
doktor, doktoren	6	341	CHI
mensch, menschen	7	1369	CHI
frau, frauen	8	2072	CHI

DE Nouns	Bin	Freq	Df
fahrgast, fahrgäste	1	9	Wiki
kleinkind, kleinkinder	2	16	Wiki
zwilling, zwillinge	2	25	Wiki
polizist, polizisten	3	42	Wiki
kunde, kunden	5	114	Wiki
schwester, schwestern	5	191	Wiki
bruder, brüder	6	428	Wiki
vater, väter	7	668	Wiki
mann, männer	7	713	Wiki
person, personen	8	1250	Wiki

DE Verbs	Bin	Freq	Long VP	Df
zweifelt, zweifeln	0	4	*am wetter*	CHI
konstruiert, konstruieren	1	5	*ein modell*	CHI
fürchtet, fürchten	3	35	*den starken sturm*	CHI
schält, schälen	3	46	*den reifen grünen apfel*	CHI
taucht, tauchen	4	84	*in das wasser des meeres*	CHI
kennt, kennen	5	282	*die antwort auf die frage*	CHI
schreibt, schreiben	7	918	*einen brief an verwandte*	CHI
erzählt, erzählen	7	1182	*eine geschichte über die ferie*	CHI
spielt, spielen	8	3411	*mit dem ball auf dem hof*	CHI
kommt, kommen	9	9843	*mit dem bus zum tennisplatz*	CHI

DE Verbs	Bin	Freq	Long VP	Df
schaukelt, schaukeln	0	2	*auf dem spielplatz*	Wiki
flüchtet, flüchten	2	13	*vor dem feuer*	Wiki
riecht, riechen	2	13	*den duft von frischem kaffee*	Wiki
wandert, wandern	4	39	*durch den wald*	Wiki
feiert, feiern	4	56	*den geburtstag des großvaters*	Wiki
verschwindet, verschwinden	5	76	*im nebel*	Wiki
denkt, denken	6	234	*an blumen im garten*	Wiki
spricht, sprechen	7	609	*über das abendessen*	Wiki
arbeitet, arbeiten	8	712	*an einem projekt*	Wiki
liegt, liegen	9	2549	*auf dem boden*	Wiki

Table 11: Chosen Nouns (used as Objects) in FIT-CLAMS for English, French and German.

English Nouns	Bin	Freq	Df
guard, guards	3	40	CHILDES
friend, friends	7	1525	CHILDES
waiter, waiters	2	11	Wiki
speaker, speakers	6	381	Wiki

French Nouns	Bin	Freq	Df	German Nouns	Bin	Freq	Df
femme, femmes	4	80	CHILDES	mitglied, mitglieder	1	8	CHILDES
adulte, adultes	3	39	CHILDES	bauer, bauern	6	357	CHILDES
constructeur, constructeurs	5	112	Wiki	matrose, matrosen	1	12	Wiki
docteur, docteurs	5	114	Wiki	familie, familien	7	1100	Wiki

Language	Prepositions
English	next to, behind, in front of, near, to the side of, across from
French	devant, derrière, en face, à côté, près
German	vor, hinter, neben, in der Nähe von, gegenüber

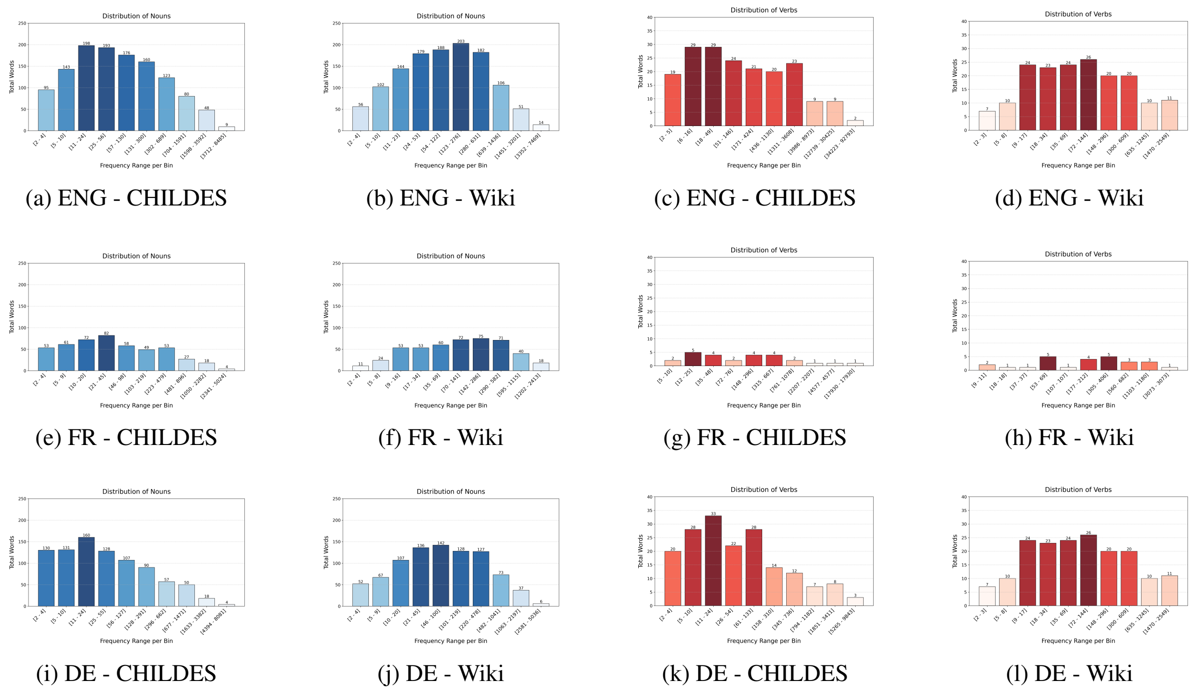
Table 12: Prepositions used FIT-CLAMS for English, French, and German.

Language	Verbs Used in Relative Clauses
English	like likes; hates hate; love loves; admires admire
French	aime aiment
German	mag mögen; vermeidet vermeiden

Table 13: Verbs used in FIT-CLAMS relative clauses for English, French, and German.

Paradigm	EN	FR	DE
Agreement in long VP coordinates	900	378	900
Agreement in object relative clauses (across)	3200	504	1600
Agreement in object relative clauses (within)	3200	504	1600
Agreement in prep phrases	4800	2520	4000
Simple agreement	200	126	200
Agreement in subject relative clauses	3200	504	1600
Agreement in VP coordinates	900	378	900

Table 14: Minimal pair counts of FIT-CLAMS (same for FIT-CLAMS-C and FIT-CLAMS-W) for each paradigm across three languages.



(a) ENG - CHILDES (b) ENG - Wiki (c) ENG - CHILDES (d) ENG - Wiki

(e) FR - CHILDES (f) FR - Wiki (g) FR - CHILDES (h) FR - Wiki

(i) DE - CHILDES (j) DE - Wiki (k) DE - CHILDES (l) DE - Wiki

Figure 13: Noun and Verb Distribution across bins, in the two datasets and in the three languages.