Do Syntactic Categories Help in Developmentally Motivated Curriculum Learning for Language Models?

Arzu Burcu Güven **Anna Rogers** **Rob van der Goot**
IT University of Copenhagen, Denmark
{argy, arog, robv}@itu.dk

Abstract

We examine the syntactic properties of BabyLM corpus, and age-groups within CHILDES. While we find that CHILDES does not exhibit strong syntactic differentiation by age, we show that the syntactic knowledge about the training data can be helpful in interpreting model performance on linguistic tasks. For curriculum learning, we explore developmental and several alternative cognitively inspired curriculum approaches. We find that some curricula help with reading tasks, but the main performance improvement come from using the subset of syntactically categorizable data, rather than the full noisy corpus. [1]

1 Introduction

Curriculum Learning (CL), a training regimen where the input is ordered from easier to more difficult, has been shown to improve performance of the machine learning algorithms in various scenarios (Soviany et al., 2022). In NLP, the BabyLM challenge (Warstadt et al., 2023), inspired by human efficiency in acquiring language from a small amount of data, has sparked interest in applying CL to small-scale training setups. Most studies in this research area base their curricula on language or syntactic complexity. However, to quantify these complexities they rely on coarse proxies, such as ordering different corpora (Martinez et al., 2023), mean length of utterance (MLU) (Oba et al., 2023) or the average number of syntactic dependents (Mi, 2023). Despite being a popular approach, CL has not consistently led to performance gains in these settings (Hu et al., 2024).

One of the core corpora in CL studies in NLP is CHILDES (MacWhinney, 2000), which consists mostly of interactions between children and adults. It is currently the primary resource for Child Directed Speech (CDS), which is known to exhibit distinct topical, lexical and morphosyntactic features (Gallaway, 1999; Huttenlocher et al., 2002; Soderstrom, 2007). Several studies use CHILDES as a stand-in for developmentally grounded training (Feng et al., 2024; Huebner, 2018; Huebner et al., 2021; Martinez et al., 2023). Surprisingly, although there are many CL studies relying on CHILDES (based on CDS (Huebner, 2018; Huebner et al., 2021), syntactic complexity (Oba et al., 2023; Mi, 2023), or language complexity (Martinez et al., 2023)), its syntactic properties have not been explored in a fine-grained manner in this line of work.

To address the gaps in the literature, namely the lack of concrete curriculum quantification and the limited analysis of CHILDES both in itself and in comparison to other corpora as training data, we propose a syntax-based approach. Our contributions are as follows:

1. We introduce a toolkit[1] to analyze, label, and order training data based on the syntactic properties of each sentence, based on approximately 300 expert-designed tregexes capturing 71% of sentences in CHILDES.

2. We contribute a detailed analysis of the BabyLM corpora for syntactic properties, and we present the analysis of developmentally motivated marco-categories across each subcorpus.

3. For CHILDES, we examine distributions by age group. We find no clear differences that align with the developmental syntactic stages proposed in language acquisition research, and we propose hypotheses for why this might be the case.

4. We train language models on syntactically and developmentally motivated curricula and compare them against baselines. We find that the primary performance gain stems not from CL

[1] https://github.com/arzuburcuguven/syntactic-categorization

Proceedings of the First BabyLM Workshop, pages 288–300

Macro-category	Syntactic Category	Examples
Simple	Subject-Verb	*She runs. She opens the bottle.*
	Adverbs & Possessives	*Try again. She runs fast. She opens your bottle.*
	Prepositions	*Good for you. She runs with her friend.*
	Particle verbs	*Cut it off. She opens up to you.*
	Auxiliaries	*She can run fast. She should open up to you.*
	Negation	*Don't run fast. She should not open up to you.*
	Tense	*You are running fast. She has been opening up to you.*
Complex	Embedded clauses	*Let's go. I know what I need.*
	To-infinitives	*I want to run. I'm going to call you.*
	Linked clauses	*I want to run and smell flowers. I run because I like it.*
	Relative clauses	*The tooth fairy who loves good children*
	Fragments	*Uh, ah yes, umm, not into that*
Interrogatives	Interrogatives	*What? Is that a hat? Does she know what the moon is?*

Table 1: Developmental macro-categories, associated syntactic categories, and example utterances.

Corpus	Genre	Tokens
CHILDES	Child-directed speech	25.9M
BNC Spoken	Spoken English	9.2M
OpenSubtitles	Movie subtitles	25.8M
Switchboard	Telephone conversations	1.6M
Simple Wikipedia	Encyclopedia	17.3M
Gutenberg	Children stories	31.0M

Table 2: Overview of corpora used in this study, with genre and token count after clean-up.

itself, but from using syntactically categorizable data.

5. We utilize our syntactic classification framework to compile syntactically isolated subcorpora, and conduct a study on cross-construction generalization. We observe mixed results: simpler categories do not cross-generalize, whereas more complex categories can improve performance on other complex ones.

2 Methods

Our overall curriculum design is built upon classifying data by syntactic categories, and ordering the classified data according to curricula. We begin by describing the datasets used in this study, followed by the syntactic categories, the categorizing process, and the curriculum design.

2.1 Datasets

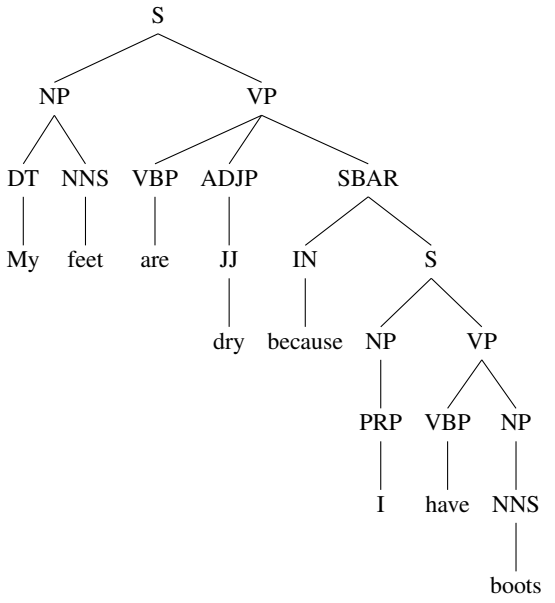
Both the training and the data analysis are conducted on the strict BabyLM dataset (Charpentier et al., 2025). The dataset comprises corpora with diverse properties, including CHILDES as CDS; Switchboard (Godfrey et al., 1992), the spoken portion of the British National Corpus (BNC) (Consortium, 2007), and OpenSubtitles (Lison and Tiedemann, 2016) as adult-directed speech (ADS); and Simple English Wikipedia and Project Gutenberg (children stories) (Gerlach and Font-Clos, 2018) as written text.

We remove speaker labels from all corpora, as the labels decrease the parser accuracy. For CHILDES, we additionally remove annotations and normalize nonstandard expressions. Sentence segmentation is applied to all corpora, and each resulting line is treated as a unit for parsing and extraction. We remove utterances shorter than two tokens. Table 2 summarizes the features and size of each corpus.

2.2 Syntactic Categorization

In order to design the syntactic categories, we examined various resources that classify syntactic phenomena into overarching groups, including typological databases such as Grambank (Lesage et al., 2022), language universals (Croft, 2002), and grammatical frameworks such as dependency relations (De Marneffe et al., 2021), and LinGO Grammar Matrix (Bender et al., 2010). Despite differences in terminology, underlying assumptions, and goals across the frameworks, we curated a set of categories that are at least represented twice among them. We found that the most comprehensive list was presented by Grambank, to which our 13 categories are most closely aligned. We restricted our final set to categories applicable to English. The resulting 13 categories are listed in Table 1. For a further discussion of these categories, see Appendix A.

For parsing the corpora we used Kitaev and Klein (2018)'s a constituency parser for its ease of use and high performance. Data was analyzed using Tregex (Levy and Andrew, 2006), for which



(a) Constituency parse of the sentence "*My feet are dry because I have boots.*"

```
% Subject-verb or intransitive sentence:
(S
  [ <1 (NP <: /NN|DT|PRP|CD|FW|VBG|EX|WP
       /)
  | <1 (NP <1 /NN|DT|PRP|CD|UH|FW|VBG|WP/
         <2 /^NN|DT|PRP|CD|FW|WP/ !<3 __)
  | <1 (NP <1 /NN|DT|PRP|CD|UH|FW|WP/
         <2 /^NN|DT|PRP|CD|FW|VBG|WP/
         <3 /^NN|DT|PRP|CD|FW|WP/ !<4 __)
  ]
  <2 (VP <: /^VB/)
  <3 /^(\.|\.\.\.|!|\?)$/
)
!> __

% Wh-question (e.g., Who is talking to
    you?):
SBARQ<(/WH/$++(/SQ|S/<1(/VB|MD/)<2VP))

% Subordinating conjunction
(e.g., My feet are dry because I have
    boots):
(NP!<<CC)
$++(VP<(/VB/
$++(SBAR<(/IN|WH/$++(S<NP<VP!<<CC)))))
```

(b) Tregex Patterns needed to match the sentence "*My feet are dry because I have boots.*"

Figure 1: Example of syntactic annotation (a) and tregexes (b) used to filter CHILDES

we designed approximately 300 regular expressions targeting the sentences that can be categorized into the 13 syntactic categories. These expressions were crafted by an experienced syntactician with a graduate degree in computational linguistics and six years of professional experience in linguistics. Matches returned by the expressions are saved and reordered to curate corpus subsets. This setup also allows for corpus-specific or cross-corpus categorization. Extracted data can also be used to create filtered training data, for example, by excluding fragments or only including relative clauses.

Figure 1a shows a constituency tree of a complex sentence and Figure 1b shows examples of Tregex patterns used to match the syntactic trees to different categories.

To the best of our knowledge, our Tregex patterns constitute the most extensive syntactic analysis of CHILDES to date; prior parsing studies used much smaller subsets (65k-236k tokens; (Sagae et al., 2007; Liu and Prud'hommeaux, 2023; Yang et al., 2025)). Even so, it categorizes only 71% of sentences in the English portion of CHILDES, primarily because of the long tail of rare that would be impossible to fully cover with Tregexes and presence of noisy disfluencies (stutters, restarts, fillers), e.g., "y you know b build this like real big thing to hold t planets from colliding together."

2.3 Curriculum

Most studies on language acquisition in English-speaking children focus on a specific syntactic phenomenon or developmental period. For instance, the seminal work by Brown (2013) describes the acquisition of a variety of phenomena such as tense, possessives, and auxiliaries, yet omits others such as interrogatives and conjunctions. Similarly, Braine and Bowerman (1976) focus exclusively on the first word combinations. Many studies approach acquisition from a universalist perspective, highlighting similarities among different language speakers (Slobin, 1987).[2]

However, to create a syntactically grounded developmental curriculum, we need a more comprehensive framework representing a wider range of phenomena. To this end, we adopted the developmental stages proposed by Friedmann and Reznick (2021), based on observations of 54 Hebrew-speaking children aged 1.5 to 6 years. These stages have also been applied to English to examine whether similar learning trajectories are also

[2]For numerous language-specific studies, see the series *The Crosslinguistic Study of Language Acquisition* (ed. D. I. Slobin).

observed in the learning behavior of LMs (Evanson et al., 2023).

Friedmann and Reznick (2021) identify three main stages in syntactic development: the first stage corresponds to simple subject–verb constructions, the second to interrogatives, and the third to relative clauses and embedded structures such as infinitives. We adopt these three stages as the basis for our main curriculum, labeling them as simple, interrogative, and complex. The 13 syntactic categories are mapped to these macro-categories as shown in Table 1.

We stress that this is only one possible hypothesis about how an effective curriculum could be constructed, and any conclusions would be made only with respect to it rather than developmentally motivated CL in general.

2.4 Evaluation

We evaluated our models using the shared BabyLM evaluation pipeline (Charpentier et al., 2025). Model evaluation was conducted on the full test set, with the exception of the Age of Acquisition (AoA) Evaluation Benchmark (Chang and Bergen, 2022). The evaluation suite includes BLiMP (Warstadt et al., 2020), EWoK (Ivanova et al., 2024), COMPS (Misra et al., 2023), (Super)GLUE (Wang et al., 2018), Entity Tracking (Kim and Schuster, 2023), WUG_ADJ (Hofmann et al., 2024), WUG_PAST (Weissweiler et al., 2023), and Reading (self-paced and eye-tracking) (de Varda et al., 2024).

BLiMP is a linguistic evaluation suite and BLiMP Supplement includes tasks specifically designed for BabyLM. COMPS and EWoK are world-knowledge datasets: COMPS focuses on immutable properties and their inheritance to subordinate concepts, whereas EWoK targets more dynamic, context-dependent properties. The Entity Tracking task assesses a model's ability to follow the states of discourse entities. WUG_ADJ evaluates adjective nominalization on nonce words, while WUG_PAST assesses past-tense formation on nonce words. The Reading task measures the alignment between LM predictions and human processing through comparison with reading times. Lastly, GLUE is used for fine-tuning evaluation.

2.5 A Closer Look into Datasets

This section provides an exposition of syntactic properties of corpora under study. First, we compare BabyLM sub-corpora and discuss differences in their distributions. Second, we examine age-ordered CHILDES to see whether syntactic distributions follow a developmental trajectory.

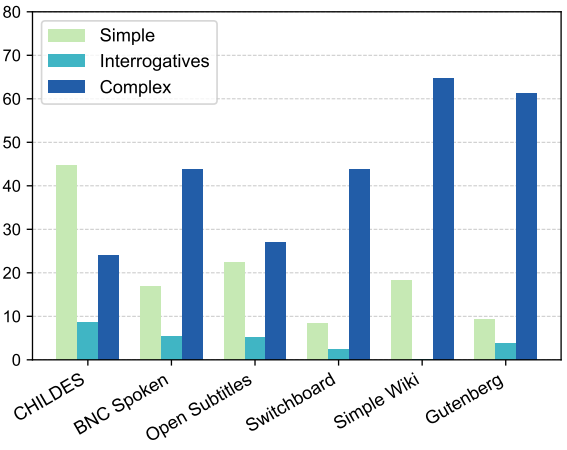


Figure 2: Distribution of macro-categories across corpora. Y-axis shows the percentage of sentences in each macro-category relative to the total number of sentences in the corpus.

2.5.1 Differences Among Corpora

In Figure 2 we present the ratio of sentences that fall under each of the macro-categories for six different corpora. Here we can see the effect of corpus genre clearly, CHILDES, being the only example of CDS differs markedly from other BabyLM corpora: Simple constructions and interrogatives account for 49% of CHILDES, compared to 10.7–27.2% in the other corpora. Among ADS corpora, BNC Spoken and Open Subtitles lean toward simpler language (16.1% simple and 5.2% interrogatives for the former; 22.0% simple and 5.2% interrogatives for the latter), whereas Switchboard has the lowest ratio of simple sentences (8.3%) and a distribution more closely aligned with text corpora.

Among written corpora, Simple English Wikipedia has the lowest proportion of interrogatives (0.04%), while Project Gutenberg is the most complex-leaning corpus, containing the highest proportion of complex sentences (59.8%).

These distributions can be useful in interpretation of model performance as identifying which constructions are rare or overrepresented in the training data provides insight into model performance across different constructions. For instance, Huebner et al. (2021) suggest that the high frequency of questions in CHILDES may explain why models trained on it perform better on interrogatives. Indeed, among the corpora analyzed here, it has the highest proportion of interrogatives (7.8%).

Padovani et al. (2025) compare models trained on CHILDES and Wikipedia. They evaluate the models on various agreement pairs and find that models trained on Wikipedia tend to perform better. This result is aligned with the distributions as relative clauses, which are one of the most challenging agreement distractors, are very scarce in CHILDES, amounting to only 0.8% of the data whereas in Simple English Wikipedia, relative clauses account for the 11.5% of the data, providing much richer training signal in terms of distractors.

2.5.2 Age-Ordered CHILDES

It is well-established that CDS is markedly different from ADS. One reason for this divergence is that adults adjust the syntactic complexity of their speech to match the child's level of comprehension (Snow, 1972; Iii and Marquis, 1977). Prior studies show that the syntactic complexity of CDS tends to increase over time, and that these changes in input correlate with children's language growth (Huttenlocher et al., 2010; Silvey et al., 2021). Given the relationship between CDS and the child's linguistic ability, we hypothesized that the age-ordered CHILDES would reflect the syntactic development of children.

Few studies have examined the differences between the age groups within CHILDES. Among them, the most relevant to our work is Bunzeck and Diessel (2025), which utilizes the morphological annotations within CHILDES with a regex-based parser, and assigns each sentence to one syntactic group among six: subject-verb constructions, interrogatives, imperatives, copular clauses, complex sentences and fragments. Their results show a subtle tendency toward interrogatives in the earlier age groups and subject-verb constructions in the older ones.

We plot the macro-categories over age groups in Figure 3, the full results on the fine-grained categories are reported in Appendix A, Figure 5. Our results do not reveal a clear developmental pattern across age groups. In line with Bunzeck and Diessel (2025)'s results, there is a subtle tendency toward interrogatives in the earlier age groups, highest being 17.5% with 3 to 4 age group. Subject-verb constructions, on the other hand, follow a non-linear trajectory, they peak at 48.9% in between the ages of 1 to 2, then decrease and increase again between the ages of 5 and 6. Excluding the preverbal group, complex constructions start from 14.6% in 1 to 2 ages and increase to 23.8% at 5 to 6. In

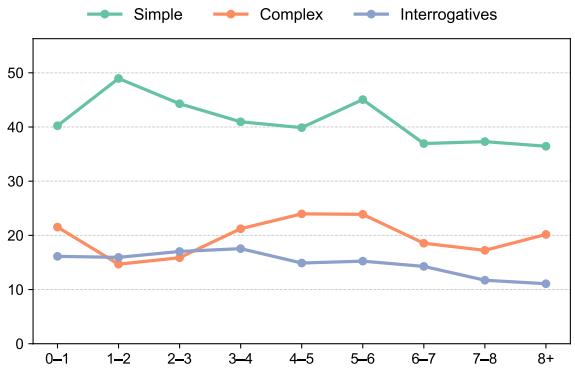


Figure 3: Distribution of macro-categories across age-ordered CHILDES. X-axis: age groups; Y-axis: percentage of sentences per macro-category.

agreement with Soderstrom (2007)'s findings, the preverbal segment of the corpus is syntactically distinct with a surprisingly high proportion of complex constructions (21%).

Our results suggest that CHILDES as a whole may not exhibit strong syntactic differentiation by age. Several factors likely contribute to this counter-intuitive outcome. The age groups aggregate data from 58 subcorpora, each containing transcripts from multiple children. Since children reach developmental milestones at individual rates (Bates et al., 2019), it may be more informative to track syntactic development longitudinally for each child, as in Brown (2013). Socioeconomic status and dialect are also known to affect language complexity (Huttenlocher et al., 2002). Lastly, CHILDES transcripts come from different sessions, such as free play and book reading, which are known to differ in their syntactic characteristics (Bunzeck and Diessel, 2025).

3 Experiments

For both CL and generalization studies, we trained a model with the GPT-2 small architecture (124M parameters) (Radford et al., 2019) from scratch using the Hugging Face Transformers library (Wolf et al., 2020). Hyperparameters are detailed in Appendix B.

3.1 Experiment 1: Curriculum

3.1.1 Methodology

This section describes experiments in which training sets are organized according to different curriculum approaches. The research question we address is "Does training on a developmentally motivated syntactic curriculum improve LM per-

Condition	BLIMP	SUPPLEMENT	EWOK	COMPS	GLUE
B1	70.24 ± 0.17	**57.66 ± 0.10**	**50.53 ± 0.28**	**52.94 ± 0.49**	57.12 ± 0.53
B2	**71.13 ± 0.62**	52.98 ± 0.70	50.27 ± 0.18	51.74 ± 0.39	57.80 ± 0.74
C1	69.88 ± 0.86	54.43 ± 2.04	50.08 ± 0.20	51.54 ± 0.58	**57.83 ± 0.51**
C2	70.45 ± 0.72	55.85 ± 0.63	50.20 ± 0.18	51.19 ± 0.41	57.45 ± 0.41
C3	70.98 ± 0.52	54.28 ± 0.29	50.06 ± 0.22	51.75 ± 0.80	57.62 ± 0.58
C4	70.03 ± 0.60	53.09 ± 0.97	49.94 ± 0.26	51.31 ± 0.09	57.80 ± 0.35
C5	70.44 ± 0.48	54.40 ± 0.89	50.19 ± 0.16	51.36 ± 0.43	57.61 ± 0.70

Table 3: Mean ± SD (over seeds) for BLiMP, Supplement, EWOK, COMPS, and GLUE. Best per column in **bold**.

Condition	ENTITY	WUG_ADJ	WUG_PAST	READING_SPR	READING_ET
B1	20.70 ± 6.09	51.10 ± 7.76	**2.28 ± 7.98**	0.04 ± 0.05	0.42 ± 0.08
B2	**41.24 ± 1.21**	**68.87 ± 1.63**	-15.81 ± 6.08	0.14 ± 0.05	0.48 ± 0.17
C1	32.34 ± 6.65	65.12 ± 1.67	-19.89 ± 10.64	**0.17 ± 0.05**	0.64 ± 0.16
C2	31.68 ± 8.75	62.06 ± 3.70	-22.81 ± 5.26	0.15 ± 0.07	**0.65 ± 0.12**
C3	38.76 ± 2.53	67.51 ± 1.10	-15.71 ± 6.15	0.08 ± 0.04	0.42 ± 0.06
C4	37.76 ± 3.71	66.84 ± 3.28	-24.32 ± 3.86	0.05 ± 0.03	0.35 ± 0.08
C5	37.83 ± 4.43	65.52 ± 4.60	-22.73 ± 2.13	0.12 ± 0.07	0.39 ± 0.04

Table 4: Mean ± SD (over seeds) for entity tracking, WUG, and reading metrics. WUG_PAST column shows correlation results multiplied by 100. Best per column in **bold**.

Cond.	Tokens	Data order
B1	131M	Random
B2	77M	Random
C1	77M	S→I→C
C2	77M	S→C
C3	77M	S→C (gradual)
C4	77M	80% SIC, 20% Mixed
C5	77M	20% Mixed, 80% SIC, 20% Mixed

Table 5: Summary of training conditions. S=Simple, I=Interrogatives, C=Complex.

formance compared to random ordering or other curriculum variants?" To this end we train seven models: two baselines (B1, B2) and five curriculum variants (C1–C5). Table 5 summarizes all training conditions.

The baselines are B1, the full BabyLM corpus in random order, and B2, an extracted subset of BabyLM corpus containing the union of all syntactically categorized data in random order. C1 (developmental curriculum, Section 2.3) groups the syntactically categorized training data into simple, interrogative, and complex stages, shuffling within each stage before concatenating them to form the final corpus.

To contrast with the developmentally grounded approach, we also devise several alternative curricula. In the simple-to-complex curriculum (C2), we categorize each syntactic structure as either simple or complex based on the presence of nested embedding. We then concatenate these two subgroups. In C3, we use the same simple and complex division described above but interleave them such that the dataset starts from only simple examples, progresses to a balanced dataset and ends with only complex examples. To achieve this, we employ a probabilistic sampling function that decreases the probability of sampling from the simple dataset and increase the probability of sampling from the complex dataset over the course of the sampling process.

The last two CL approaches are inspired by the Learn–Focus–Review (LFR) strategy of Prakriya et al. (2025), a cognitively inspired dynamic learning paradigm. In the initial learn phase, models see a portion of randomly sampled training data. In the focus phase, more challenging portions of the data are clustered, and in the review phase, the remaining data is reintroduced to prevent forgetting. For C4, 20% of the syntactically labeled data is held out, the remaining 80% is constructed as in C1, and the held-out portion is appended as a review at the end. For C5, 40% of the data is held out, 60% is constructed as in C1, and the held-out portion is split in half, with one half appended to the beginning and the other half to the end of the corpus.

3.1.2 Results

We report averaged results over four seeds on the BabyLM test suite in Table 4 and Table 3. While

Condition	Hypernym	QA_easy	QA_tricky	SubjAuxInv	Turn_taking
B1	48.99 ± 0.35	**55.47 ± 2.71**	**39.55 ± 1.04**	84.02 ± 1.32	**60.27 ± 2.17**
B2	49.82 ± 0.50	49.61 ± 2.66	27.88 ± 3.39	87.68 ± 1.81	49.91 ± 1.94
C1	50.27 ± 0.68	52.34 ± 4.51	36.21 ± 2.29	83.77 ± 5.53	49.55 ± 0.45
C2	49.94 ± 1.08	52.73 ± 3.46	38.03 ± 3.14	**87.95 ± 0.66**	50.62 ± 1.28
C3	50.23 ± 1.52	53.90 ± 1.57	29.39 ± 1.61	87.20 ± 1.80	50.62 ± 0.68
C4	49.47 ± 1.10	50.00 ± 2.21	30.00 ± 1.60	85.88 ± 0.70	50.09 ± 1.18
C5	**50.62 ± 0.91**	52.73 ± 1.97	31.06 ± 1.94	88.54 ± 1.07	49.02 ± 1.38

Table 6: Mean ± SD over seeds for UID subtasks. Best per column in bold.

curriculum learning offers some task-specific benefits, the main finding is that models trained on parsed and categorized data perform on par with the B1 baseline despite requiring 40% fewer training steps. B1 still leads on BLiMP Supplement, EWOK, COMPS and WUG_PAST, though the EWOK and COMPS margins are small.

The difference in BLiMP Supplement scores may stem from a preprocessing decision: to make our training data parser-compatible, we removed speaker labels. As a result only the B1 model, which was trained on the whole BabyLM corpus, was shown examples with speaker labels. As shown in Table 6 (Appendix B), B1's higher Supplement score is concentrated in three subcategories, *QA_easy*, *QA_tricky*, and *turn-taking*; each containing speaker labels. Since in the main BLiMP benchmark and other Supplement categories the other models outperform B1, this suggests that presence of the speaker labels likely accounts for the observed gap.

The difference in performance on WUG_PAST is more difficult to interpret. A qualitative analysis of the predictions shows that B1 models tend to apply regular inflection to wug words more often, aligning more closely with human data. In contrast, the other models more frequently produce irregular inflections, correlating negatively with the baseline. For the WUG_ADJ task, however, B1 underperforms compared to all other models. One possible explanation is that cleaner data makes models more attentive to irregularities. This may be an advantage in tasks with a constrained prediction space, such as selecting from a limited set of adjective nominalizers, but a disadvantage in open-set tasks like WUG_PAST.

For GLUE, entity tracking, and reading tasks, models trained on categorized data outperform the B1 models. Especially for reading tasks, both self-paced reading and eye-tracking, curriculum models C1 and C2 show the highest performance, suggesting that the curriculum approaches can provide a signal that shortens the gap between human and machine processing.

Category	Constructions
Subject-Verb	Subject-Verb patterns
Modifier	Adverbs, Possessives, and Prepositions
Verbal	Particle verbs, Auxiliaries, Negation, and Tense
Embedded C.	Small clauses, reported speech
Infinitives	Infinitives
Linked Clauses	Coordination, Subordination
Relative Clauses	Relative Clauses
Interrogatives	Yes/no, wh-questions

Table 7: Syntactic category groups used in the generalization study and their corresponding constructions.

3.2 Experiment 2: Generalization

3.2.1 Methodology

In this experiment, each model is trained on a single category and evaluated on eight validation sets corresponding to the distinct eight categories given in Table 7. We approach this task as a generalization study, using the perplexity values as a proxy for models' ability to learn both the category they trained on and the remaining seven unseen categories.

For each group, we sample 2M tokens for training and 200K tokens for validation from the syntactically classified portion of BabyLM corpus. Sampling is restricted to sentences matching the target group's criteria. We train GPT-2 small models from scratch on each subset for one epoch. All results are averaged over five random seeds.

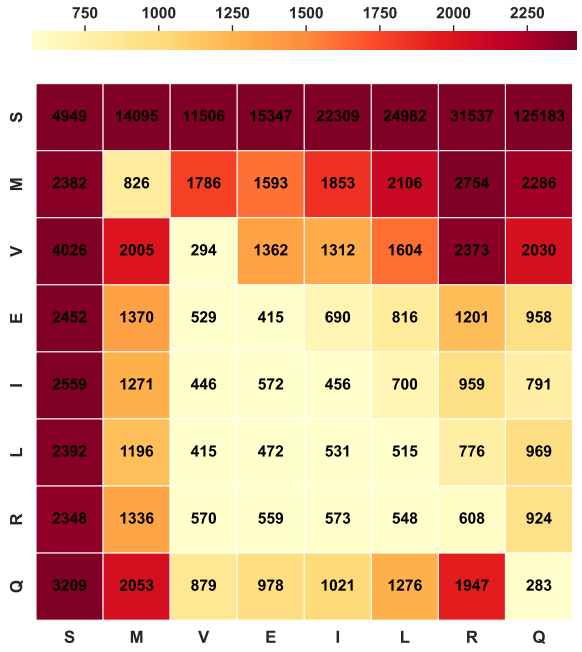


Figure 4: Cross-subset validation perplexity heatmap. Rows = training subset; columns = evaluation subset. Abbreviations: S=SVX, M=Modifiers, V=Verbal, E=Embedded, I=Infinitives, L=Coordination, R=Relative, Q=Question. Cell values are validation perplexities (lower is better).

3.2.2 Results

In Figure 4 we report mean perplexity values across seeds. As expected, each model achieves its lowest perplexity when evaluated on the same syntactic category it was trained on (diagonal entries). Off-diagonal values indicate cross-category generalization.

Performance patterns vary across categories. The *Subject–Verb* group (SVX) shows the largest drop in both in-category and cross-category performance, likely due to the high frequency of single-word (e.g., "Run!") and fragmentary utterances (e.g., "all gone"). The *Verbal* and *Modifier* groups also generalize poorly. Models trained on *questions*, despite the data exhibiting unique syntactic patterns such as subject auxiliary inversion, generalize better than those trained on *Subject–Verb*, *Verbal* and *Modifier* constructions. Models trained on complex constructions tend to generalize better to other complex categories. The *Coordination*-trained model exhibits the strongest overall generalization, with the lowest mean off-diagonal perplexity (962.20) and the lowest perplexity on the mixed test set (574.2).

Overall perplexities remain high, and there is limited evidence for genuine syntactic generalization, particularly from simpler to more complex categories. Prior work demonstrating such transfer with transformer architectures typically relies on synthetic datasets with tightly controlled syntax and vocabulary (Murty et al., 2023; Ahuja et al., 2025; Someya et al., 2024). Our subsets are selected by syntactic criteria but retain naturalistic variation in sentence form and vocabulary. These results highlight the difficulty of isolating syntactic generalization in naturalistic data and suggest that stricter control of lexical and structural properties may be necessary for clearer conclusions.

4 Conclusion

This study contributes the most detailed syntactic analysis of BabyLM data to date, implemented as an open-source toolkit for analysing, labeling and ordering training data.[1] This enabled both modeling experiments and a systematic analysis of syntactic patterns in CHILDES, where, counter-intuitively, we find no clear differences in distributions that would align with syntactic stages proposed in language acquisition research. Likewise, we find that developmentally motivated curriculum has a modest effect in language model training, compared to simply training the models on a subset of training data filtered to only syntactically categorizable sentences.

Efficient curriculum learning for language models that is inspired by human learning stages remains an elusive goal. The results of this study suggest that continued focus solely on syntax may be counter-productive, and that the noise in popular resources such as CHILDES may by itself have an outsized effect in studies relying on it.

Limitations

We note the following limitations of this study:

1. We did not observe developmental patterns in the aggregated CHILDES data, but our analysis did not extend to a more fine-grained level where confounding factors could be mitigated.

2. Our syntactic categorization covered 71% of the BabyLM; some of the remaining gap is attributable to our data cleaning practices, but a portion remains unexplained.

3. The absence of clear effects from CL or generalization may stem from several factors, and this study does not establish which ones are

the most relevant. It is possible that isolating syntactic properties alone could be insufficient, or our method of isolation may not capture the most relevant distinctions. Alternatively, the targeted developmental progression and generalization may not be reproducible with the transformer architecture or training conditions used.

References

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. 2025. Learning Syntax Without Planting Trees: Understanding Hierarchical Generalization in Transformers. *Transactions of the Association for Computational Linguistics*, 13:121–141. Place: Cambridge, MA Publisher: MIT Press.

Elizabeth Bates, Philip Dale, and Donna Thal. 2019. Individual Differences and their Implications for Theories of Language Development. pages 95–151.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar Customization. *Research on Language and Computation*, 8(1):23–72.

Martin D. S. Braine and Melissa Bowerman. 1976. Children's First Word Combinations. *Monographs of the Society for Research in Child Development*, 41(1):1.

Roger Brown. 2013. *A First Language: The Early Stages*. Harvard University Press.

Bastian Bunzeck and Holger Diessel. 2025. The richness of the stimulus: Constructional variation and development in child-directed speech. *First Language*, 45(2):152–176. Publisher: SAGE Publications Ltd.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

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

BNC Consortium. 2007. British national corpus, XML edition. Literary and Linguistic Data Service.

William Croft. 2002. *Typology and Universals*, 2 edition. Cambridge University Press.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.

Steven Y. Feng, Noah Goodman, and Michael 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.

Naama Friedmann and Julia Reznick. 2021. Stages rather than ages in the acquisition of movement structures: Data from sentence repetition and 27696 spontaneous clauses. *Glossa: a journal of general linguistics*, 39(1).

Clare Gallaway, editor. 1999. *Input and interaction in language acquisition*, 1. publ. [transferred to digital reprinting] edition. Cambridge University Press, Cambridge.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Preprint*, arXiv:1812.08092.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520 vol.1, San Francisco, CA, USA. IEEE.

Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, and Janet Pierrehumbert. 2024. Derivational Morphology Reveals Analogical Generalization in Large Language Models. *arXiv preprint*. ArXiv:2411.07990 [cs].

Yaling Hsiao, Nicola J. Dawson, Nilanjana Banerji, and Kate Nation. 2023. The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar. *Journal of Child Language*, 50(3):555–580.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. *arXiv preprint*. ArXiv:2412.05149 [cs].

Philip Huebner. 2018. Order matters: Distributional properties of speech to young children bootstrap-slearning of semantic representations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 40(0).

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.

Janellen Huttenlocher, Marina Vasilyeva, Elina Cymerman, and Susan Levine. 2002. Language input and child syntax. *Cognitive Psychology*, 45(3):337–374.

Janellen Huttenlocher, Heidi Waterfall, Marina Vasilyeva, Jack Vevea, and Larry V. Hedges. 2010. Sources of Variability in Children's Language Growth. *Cognitive psychology*, 61(4):343–365.

John Neil Bohannon Iii and Angela Lynn Marquis. 1977. Children's Control of Adult Speech. *Child Development*, 48(3):1002.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Najoung Kim and Sebastian Schuster. 2023. Entity Tracking in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Jakob Lesage, Hannah J. Haynie, Hedvig Skirgård, Tobias Weber, and Alena Witzlack-Makarevich. 2022. Overlooked Data in Typological Databases: What Grambank Teaches Us About Gaps in Grammars. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2884–2890, Marseille, France. European Language Resources Association.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Zoey Liu and Emily Prud'hommeaux. 2023. Data-driven Parsing Evaluation for Child-Parent Interactions. *Transactions of the Association for Computational Linguistics*, 11:1734–1753.

Brian MacWhinney. 2000. The childes project. 1: Transcription format and programs. Erlbaum, Mahwah. Num Pages: 159.

Richard Diehl Martinez, Zebulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB: Curriculum Learning for Infant-inspired Model Building. *arXiv preprint*. ArXiv:2311.08886 [cs].

Maggie Mi. 2023. Mmi01 at The BabyLM Challenge: Linguistically Motivated Curriculum Learning for Pretraining in Low-Resource Settings. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 269–278, Singapore. Association for Computational Linguistics.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Grokking of Hierarchical Structure in Vanilla Transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.

Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Oseki. 2023. BabyLM Challenge: Curriculum learning based on sentence complexity approximating language acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 262–269, Singapore. Association for Computational Linguistics.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-Directed Language Does Not Consistently Boost Syntax Learning in Language Models. *arXiv preprint*. ArXiv:2505.23689 [cs].

Neha Prakriya, Jui-Nan Yen, Cho-Jui Hsieh, and Jason Cong. 2025. Accelerating large language model pretraining via LFR pedagogy: Learn, focus, and review. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 268–290, Vienna, Austria. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. High-accuracy Annotation and Parsing of CHILDES Transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

Catriona Silvey, Özlem Ece Demir-Lira, Susan Goldin-Meadow, and Stephen W. Raudenbush. 2021. Effects of Time-Varying Parent Input on Children's Language Outcomes Differ for Vocabulary and Syntax. *Psychological Science*, 32(4):536–548.

Dan I. Slobin, editor. 1987. *The crosslinguistic study of language acquisition, Vol. 1: The data*. Erlbaum, Hillsdale, NJ.

Catherine E. Snow. 1972. Mothers' Speech to Children Learning Language. *Child Development*, 43(2):549.

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. Targeted syntactic evaluation on the Chomsky hierarchy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15595–15605, Torino, Italia. ELRA and ICCL.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum Learning: A Survey. *International Journal of Computer Vision*, 130(6):1526–1565.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, Singapore. Association for Computational Linguistics.

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

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

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

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. UD-English-CHILDES: A Collected Resource of Gold and Silver Universal Dependencies Trees for Child Language Interactions. *arXiv preprint*. ArXiv:2504.20304 [cs].

A Category Details

Below, we list our categories ordered by an increasing number of terminals and combinatorial possibilities. We start from simple noun phrases (NP), verb phrases (VP), adjective phrases (ADJP) and Subject-Verb constructions that can be built with them. For the categories with simpler constructions without any nested structures, the Tregex patterns match entire sequences and tightly constrain the contents of each node to exclude any complex expansions within the tree. For the more complex categories, we switch to partial matching, without constraining the preterminal nodes.

- *Subject-Verb Constructions*: For the sake of readability we use the term Subject-Verb Constructions, but the structures included are intransitive sentences (SV), transitive sentences (SVO), imperatives and copular sentences (SVC). Preterminals included in this category are simple NPs, VPs and ADJPs that have limited amount of nodes and no nested structures under them. Along with the well formed structures, we include sequences that consist

of phrases such as *Beautiful girl, the doll, all toys, love you Baby* etc. For the following categories up to the interrogatives, the sentence structures are limited to the ones described here.

- *Possessives and Adverbials*: For this category, we add POS and ADV preterminals to the former group. The NPs are extended to include possessives e.g., *The girl's hat is beautiful*. Adverbial phrases are allowed both under VPs and directly under the S node.

- *Prepositions*: Phrases headed by PPs (*at the table*), NPs governing over PPs (*the girl with the blue ribbon*), ADJPs governing over PPs (*good for you*) and VPs governing over PPs (*walk to me*) are included both as standalone phrases and as participants in the SVX structures.

- *Particles*: VP categories are extended to include particle verbs (*take off, put on*). This category forms one of the smallest categories in terms of how many sentences it captures, along with auxiliaries and tense.

- *Auxiliaries*: Here we repeat all the canonical sentence types from the former categories, SVX, SVX with adverbs, SVX with PPs and so on and modify the VPs to govern over an auxiliary.

- *Negation*: The scope is again limited to all the canonical sentence types from the former categories and VPs are modified to govern over the negation particle.

- *Tense*: Although we have not differentiated between simple present or simple past tenses in the former categories, the more complex tenses such as progressive and perfective require a specific VP category. Again, we repeat all the canonical sentence types from the former categories, and modify the VPs to allow for the capture of complex tenses.

- *Interrogatives*: Here we include different types of interrogatives: Yes/no questions (*Is she coming?*), Wh-questions (*What is she doing?*), tag questions (*She doesn't know, does she?*) and question fragments (*What?, Did she?*).

Hyperparameter	Value
Model type	GPT-2 small
Parameters	124M
Vocabulary size	50,257
Context size	1024
Dropout	0.1
Learning rate	1.88×10^{-4}
Scheduler	Linear
Weight decay	0
Epochs	1
Batch size	8
Optimizer	AdamW

Table 8: Training hyperparameters for GPT-2 small

- *Embedded Clauses*: This group captures a variety of nested structures in which at least two predicates are present. This includes let-constructions such as *let me go*, causatives (*I will make him bite mommy*) and small clauses (*I think you can fix it*).

- *Infinitives*: This category captures the to-infinitives and gerunds e.g., *She wants to drink from her cup.*

- *Clause Linking*: Here we include coordinating conjunctions (*She ate an apple but the apple was rotten*) and subordinating conjunctions (*My feet are dry because I have boots.*).

- *Relative Clauses*: This category is adapted from Hsiao et al. (2023), which includes relative clauses of subject (*The man who kicked the ball*), object (*the fun I had*) and passive (*the houses that were built*) types.

- *Fragments*: While we allow phrase level constructions when they represent a well formed phrase, malformed phrases and interjections fall into this group.

B Model Details

We tuned hyperparameters with a sweep: learning rate sampled log-uniformly in $[5\times10^{-6},\ 5\times10^{-4}]$ and per-device train batch size $\in \{8, 16, 32\}$; the best model was selected by validation-set perplexity. Remaining hyperparameters were taken from Radford et al. (2019). The full set of hyperparameters is shown in Table 8.

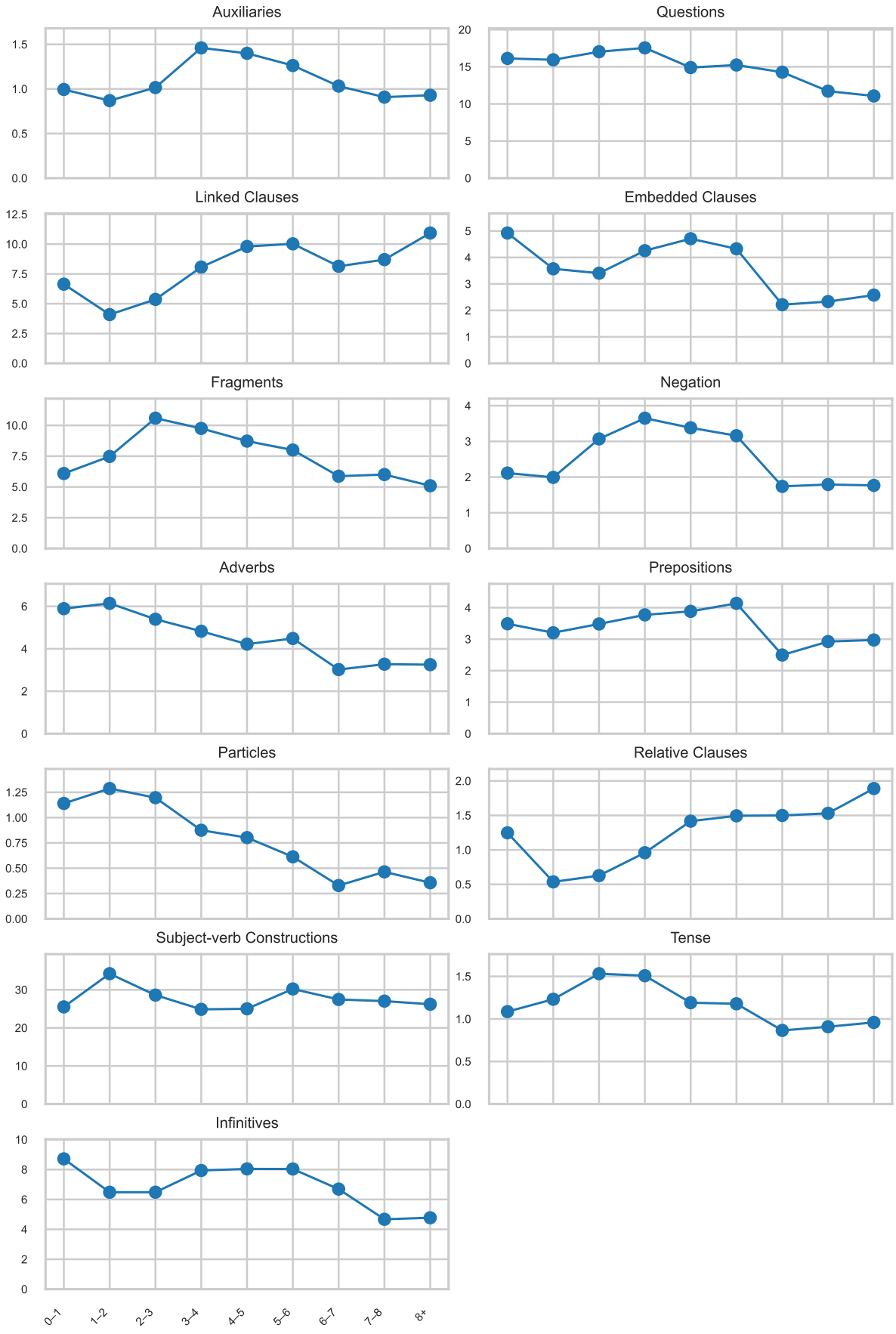


Figure 5: Percentage distribution of syntactic categories across age groups in CHILDES.