

A Limitations and ethical considerations

A.1 Limitations

Our study is limited by data availability. Creating a full-fledged 100M-token BabyLM dataset is currently out of question, as neither CHILDES nor other sources contain even remotely enough data for languages other than English. Principally, synthetic corpora like the TinyStories dataset (Eldan and Li, 2023), which contains children’s stories generated by GPT-3 or TinyDialogues by Feng et al. (2024) would provide an unlimited source of training data. However, our inspection of their generated dialogues yielded that they drastically underestimate the high numbers of grammatical fragments, questions and short SV(X)-utterances in real-world data. Similarly, there are little to no evaluation sets for German beyond those that we included/creates ourselves, especially on the syntactic level.

Moreover, actual developmental plausibility also hinges on the inclusion of other modalities. For audio data, there are few CHILDES subcorpora and other corpora that contain phonetic information (Lavechin et al., 2023), but larger models need to be trained on more data, e.g. audiobooks (Lavechin et al., 2025). A middle ground is training on textual phonetic transcriptions generated from raw text, e.g. for the BabyLM data (Goriely et al., 2024; Bunzeck et al., 2025). More recently, also video recordings from infant-mounted cameras have been used to train on combined visual and auditory input modalities (Wang et al., 2023; Vong et al., 2024; Long et al., 2024). The inclusion of such data could help to disentangle learning processes further.

A.2 Ethical considerations

Given the nature of this work, there are no specific ethical concerns to address. However, we want to emphasize that BabyLMs are *not* actual babies, but rather abstractions, or *models* in the original scientific sense, of the distributional, frequency-driven aspects of their learning capacity. All claims regarding their implications for language development in the real world should be understood in this context, which we also attempted to explicate by distinguishing functional and formal aspects of learning.

B Excluded corpora

Several corpora that are — in principal — available for German were excluded from our analysis. The

Folk corpus (Reineke et al., 2023) and the Simple German corpus (Jach and Dietz, 2024) are not available under any open licenses, while the data in other German reference corpora (Kupietz et al., 2010) are not available in their entirety but can only be queried through web interfaces. Finally, Homebank features day-long audio recordings of children and their surroundings/inputs (VanDam et al., 2016), but without any written transcriptions.

C Data cleaning

In line with best practices in language modeling, we extensively clean and normalize our data. Our cleaning script is available at [link removed for anonymization].

All subcorpora We replaced all local variants of single/double quotation marks with either ' ' or " ". We further reduced multiple superfluous whitespace and newlines to singular whitespaces.

Talkbank data For the data sourced from talkbank (i.e. the CHILDES corpora and CallHome), we remove all mark-up and additional info on false starts, hesitations, implicit completions or other explanations. Furthermore, we also remove all empty utterances and those containing xxx or yyy, placeholder symbols for personally identifiable information.

Project Gutenberg For the Project Gutenberg data, we excluded all lines with more than 6 consecutive whitespaces, as these always turned out to be title pages, index pages, etc., which contain no useful language data. Additionally, we removed all textual data in square brackets, which almost always corresponded to pointers to pictures which are not found in text-only version, or additional explanations by the volunteers who digitized the respective books.

OpenSubtitles For the OpenSubtitles data, we removed all text in parentheses, which corresponds to speaker information. Also, we removed sentence-initial dashes (–) which were sometimes added. We also amended OCR errors (like mangled uppercase I and lowercase l) as far as possible.

Flutter For the data sourced from the Flutter magazine, we removed all lines containing additional metatextual data, like author info and image credits, before pre-training.

D Exact construction proportions

Table 4 shows the exact construction proportions for all of our subcorpora. This data underlies the visualization in Figure 1.

Construction	Proj. Gut.	Dreamb.	Fluter	News	Wikib.	Klex.	Mini-Klex.	OpenSub.	CallHome	Child speech	CDS
FRA	7.8%	6.3%	6.2%	4.0%	11.6%	6.3%	2.5%	24.1%	37.0%	55.1%	24.5%
QWH	1.9%	0.3%	2.6%	1.4%	0.5%	2.9%	<0.1%	7.3%	2.1%	3.5%	8.8%
QYN	3.7%	0.7%	2.8%	1.6%	0.5%	0.4%	<0.1%	10.9%	6.9%	4.7%	20.7%
COP	4.6%	7.1%	7.7%	7.4%	10.9%	13.2%	21.4%	9.7%	10.7%	5.7%	8.1%
IMP	1.5%	0.1%	0.2%	0.1%	0.3%	<0.1%	<0.1%	4.6%	0.4%	2.0%	4.5%
SPI	7.5%	9.2%	9.7%	13.7%	9.5%	13.9%	19.9%	9.9%	8.8%	11.5%	10.1%
SPT	10.5%	14.5%	18.7%	25.7%	24.1%	28.1%	37.2%	18.0%	14.1%	11.9%	12.3%
COM	62.5%	61.8%	52.2%	46.1%	42.7%	35.2%	18.9%	15.4%	20.0%	5.7%	11.0%

Table 4: Exact proportions of constructions for all subcorpora

E Model hyperparameters and training details

Our models share a hidden/intermediate/embedding size of 256, 8 hidden layers and attentions heads, and a context length of 128. For the character models, the vocabulary consists of all printable ASCII characters and characters used in written German (üäöß and their uppercase variants), amounting to a vocab. size of 110 and 3,730,688 parameters. For the subword models, we train a BPE tokenizer (Gage, 1994) with a vocab. size of 8,000 and add two special tokens (BOS, EOS/PAD), resulting in 8,002 vocab. tokens and 7,771,392 parameters. Model training takes approx. 2h on a MacBook Pro with an Apple M2 Pro CPU/GPU.

We reproduce the training and test loss curves for our models in Figure 3. For the test loss, we evaluated perplexity on a held-out, randomly sampled portion of each individual training corpus. We find no principal differences in loss development, although the character models and models trained on the cds data seem to converge the fastest. As the similar curves for train and test loss indicate, all models succeed in optimizing for their next-token prediction goal.

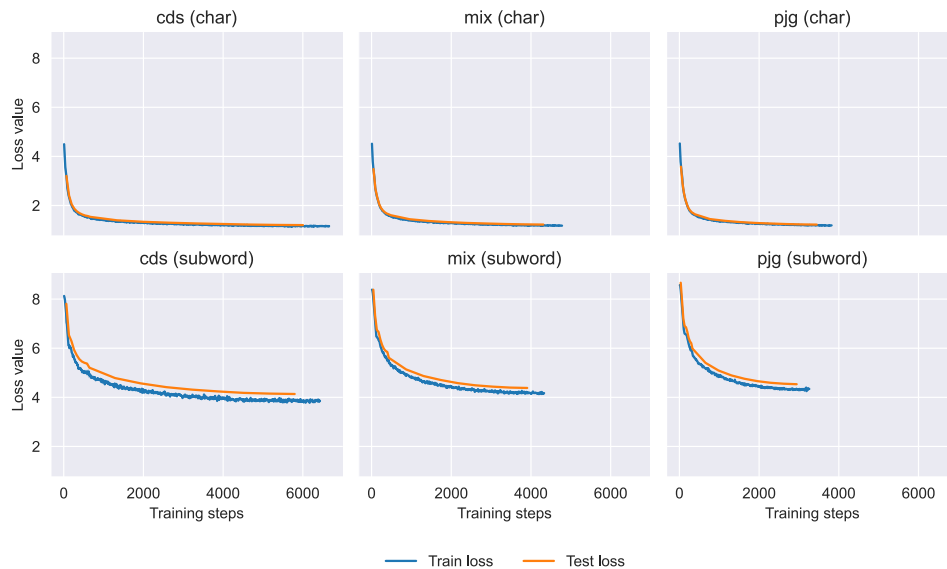
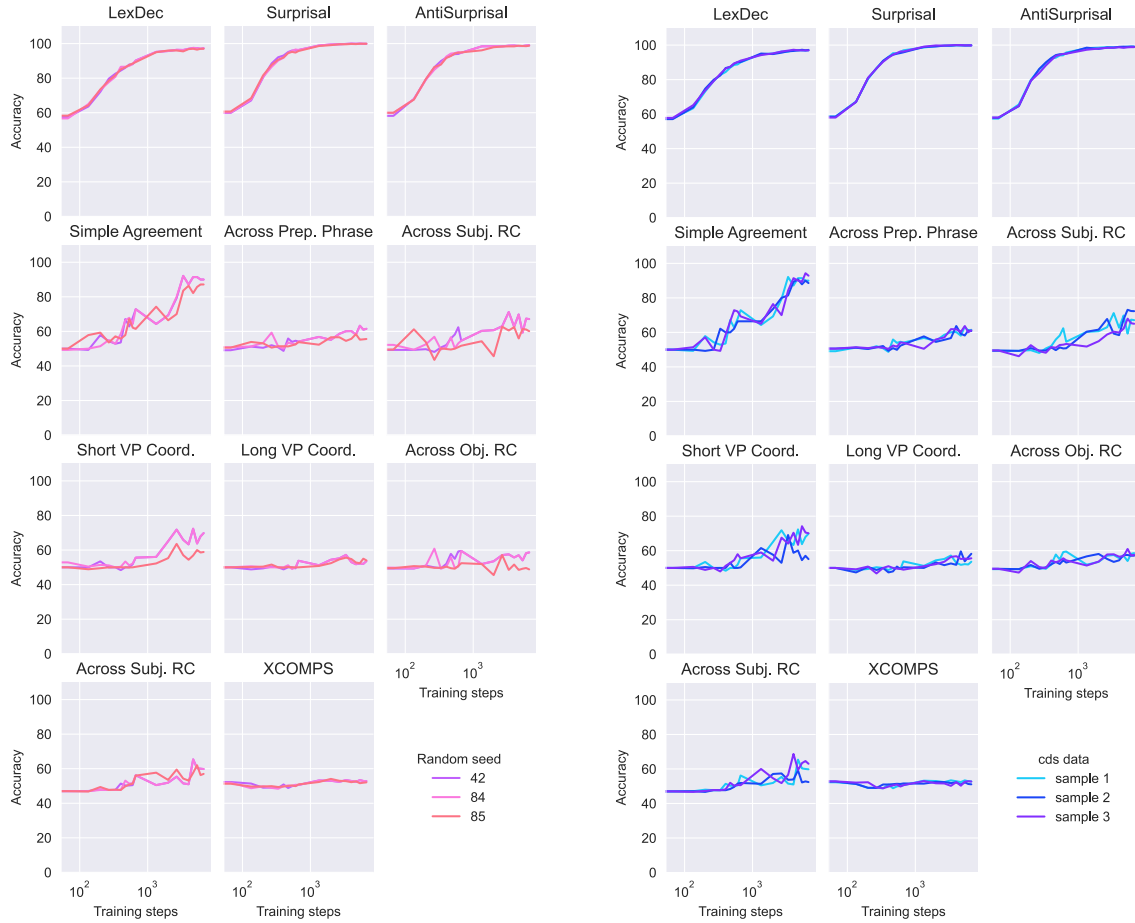


Figure 3: Loss curves for our self-trained character and subword models

F Repeated training runs

A common criticism in the BabyLM paradigm is the purported effect of training noise on model performance, which is hard to disentangle from real training data effects. While training and evaluating multiple random seeds for all our models would be too costly, we repeated two additional training runs for the character-level cds model with different random initializations (learning trajectories in Figure 4a) and two additional training runs where we re-sampled the cds dataset from our whole corpus with the exact same construction composition, but different content (learning curves in Figure 4b). In both cases, the learning trajectories do not differ tremendously. For the word-level phenomena (LexDec, Surprisal, AntiSurprisal), the curves overlap almost perfectly. For the syntax phenomena, we can see some variation and oscillation in the curves, but the trajectories still remain extremely similar (and do not differ in their steepness, the main effect that we see in Figure 2 between the datasets with different construction compositions).



(a) Trajectories for different random initializations

(b) Trajectories for different samples of cds data

Figure 4: Learning trajectories for our comparison models