# Do Construction Distributions Shape Formal Language Learning In German BabyLMs?

**Bastian Bunzeck** and **Daniel Duran** and **Sina Zarrieß**
CRC 1646 – Linguistic Creativity in Communication
Department of Linguistics
Bielefeld University, Germany
{firstname.lastname}@uni-bielefeld.de

## Abstract

We analyze the influence of utterance-level construction distributions in German child-directed/child-available speech on the resulting word-level, syntactic and semantic competence (and their underlying learning trajectories) in small LMs, which we train on a novel collection of developmentally plausible language data for German. We find that trajectories are surprisingly robust for markedly different distributions of constructions in the training data, which have little effect on final accuracies and almost no effect on global learning trajectories. While syntax learning benefits from more complex utterances, word-level learning culminates in better scores with more fragmentary utterances. We argue that LMs trained on developmentally plausible data can contribute to debates on how conducive different kinds of linguistic stimuli are to language learning.

**Project Gutenberg:** Complex sentences

Aber sie war in Angst, dass wir die Larven beschädigen würden, die zu Arbeiterinnen heranwachsen sollten. (*But she was afraid that we would damage the larvae which were supposed to grow into workers.*)

**MiniKlexikon:** Transitive SP sentences

Der Grafiker entwirft das Bild vorne auf dem Buch. (*The graphic designer designs the picture on the cover of the book.*)
Der Friseur schneidet die Haare. (*The hairdresser cuts the hair.*)

**Child-directed speech:** Fragmentary utterances

noch mehr! (*even more!*)
ja. (*yes.*)
mit dem Flugzeug. (*with the airplane.*)

Figure 1: Examples for most frequent construction types from different portions of our German BabyLM corpus

## 1 Introduction

One of the most contentious issues in language acquisition is the relationship between the input that learners receive and the resulting linguistic system (Pullum and Scholz, 2002; Clark and Lappin, 2011). Child-directed speech (or *CDS*) is structurally simple: Especially in the first three years of life, it abounds with questions, imperatives, and fragmentary utterances, but features fewer SV(X) and very few complex sentences, which instantiate "canonical" word order (Cameron-Faulkner et al., 2003). This distribution of utterance-level constructions is conducive to the *functional* side of language acquisition: caregivers talk in this way to elicit responses, steer behavior, or establish joint attention. But how do children acquire full-fledged, *formal* grammatical knowledge from such supposedly skewed input? While its advantages for aspects like speech segmentation or word learning are somewhat accepted (Yurovsky et al., 2012; Cristia et al., 2019), its influence on syntax remains debated: whereas some generativist approaches see any kind of input as too impoverished to learn a full-fledged syntactic system (cf. Chomsky, 1965; Crain and Pietroski, 2001; Guasti, 2002; Thomas, 2002; Berwick et al., 2011), constructionist and usage-based scholars argue that this supposedly skewed input actually aids syntax learning (MacWhinney, 2004; Tomasello, 2005; Bunzeck and Diessel, 2024).

The connectionist "renaissance", fueled by deep learning and Transformer language models, has opened up new avenues of investigating the relationship between an artificial learner's acquired linguistic system and the nature of its training data, more recently also from a constructionist/non-generativist viewpoint (Weissweiler et al., 2023; Piantadosi, 2024). LLMs, pretrained on raw language data only, and instruction-finetuned chatbots based on them, generate text without grammatical errors, and perform well in controlled syntactic test suites. Unfortunately, though, their massive parameter size does not preclude the possibility

that their linguistic capabilities result from memorization rather than generalization (Millière, 2024). Furthermore, the sheer amount of their pretraining data exceeds human learner's input by many orders of magnitude, putting their relevance for linguistic modeling into question. Work within the BabyLM community (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025) has demonstrated that Transformer LMs, trained on cognitively plausible amounts of data, can often acquire fairly complex syntactic structures, even without instruction-finetuning. They can also learn accurate word-level representations when trained with character-level tokenization (Bunzeck and Zarrieß, 2025; Goriely and Buttery, 2025a). This makes them ideal testbeds for the aforementioned issue: does the construction distribution found in CDS, which features a high proportion of questions and syntactic fragments, affect the acquisition of formal linguistic capabilities? In other words, does robust linguistic knowledge at the word and syntax level emerge when the training data is closer to the fragmented, "messy" input of human learners?

The goals of this paper, then, are twofold: (1) we compile a novel German BabyLM training set, for which we conduct the first utterance-level construction analysis for German. We find that distributions align with findings for English and other languages. We then (2) create three 5M-token subsets with distinct constructional profiles, varying, e.g., the proportion of fragmentary and complex utterances, and train small, character-based and subword Llama models on them. We evaluate them with lexical, syntactic, and semantic minimal pairs (Bunzeck et al., 2025; Mueller et al., 2020; He et al., 2025) to gauge the influence of different construction distributions on these levels of linguistic knowledge, and find that differences between grammatically complex training data and a developmentally plausible constructional distribution are fairly small. While certain syntactic phenomena are learned somewhat better from more complex sentences, lexical learning improves with more fragments and questions in the input. Most interestingly, input complexity only modulates the steepness of the resulting learning trajectories, but has no principal effect on the amount of input needed to kickstart learning.

## 2 Constructions in children's input

Child-directed speech can be seen as a separate linguistic register and is the primary input that children encounter in their first years. On the phonetic level, it features slower speech and exaggerated intonation patterns, which infants prefer listening to (Zangl and Mills, 2007), while its vocabulary is mostly restricted to everyday topics and children's immediate surroundings (Snow and Ferguson, 1977). Structurally, child-directed utterances are usually shorter and simpler than adult-directed ones (Genovese et al., 2020) and feature high amounts of structural and lexical repetition (Tal et al., 2024). Statistical properties of the input directly influence the children's order of acquisition for syntactic patterns (Huttenlocher et al., 2002; Ambridge et al., 2015), e.g., for relative clauses (Diessel and Tomasello, 2000; Brandt et al., 2008; Chen and Shirai, 2015).

Early studies were mostly concerned with mapping out how much CDS is ungrammatical or otherwise "wrong" (in the sense of hesitations, false starts, etc., cf. Pine, 1994), but the quantitative turn in linguistics (Janda, 2013) has enabled more holistic analyses. In a seminal study, Cameron-Faulkner et al. (2003) analyze utterance-level constructions in child-directed English via corpora of toyplay sessions featuring children and caregivers. They show that CDS features only few "canonical" SV(X)-utterances but abounds with questions, lexical fragments, or copula constructions. The reported construction distributions also hold for typologically different languages, e.g., Irish (Cameron-Faulkner and Hickey, 2011). These constructions and their real-world functions help children to quickly understand the *functional* side of language. However, the most common and repetitive utterances that English-speaking children hear represent a rather skewed sample of the presumed, underlying *formal* language system. Generativist approaches would argue that certain formal processes, like question formation from relative clauses, are not attainable from this kind of language, as the input never contains specific examples (Chomsky, 1980) (although Pullum and Scholz, 2002 find that the input frequently contains exactly such specific examples). They also partly emphasize the importance of statistical learning, e.g. for providing hypotheses about competing possible mental grammars constrained by innate, language-specific mechanisms (cf. Yang, 2004, also Ambridge and Lieven, 2011, 121f.). Constructivist approaches do not view language learning as such a re-construction of the target language's abstract grammar, but rather as the re-construction of the target language's inventory

of form-meaning pairings (Behrens, 2021). They argue that this kind of input is actually conducive to formal aspects of acquisition, by providing anchor points for first words and their semantic links to real-world reference, which then serve as building blocks for a gradual development into larger schemas (like questions with relative clauses).

Although CDS features such a skewed construction distribution, written language aimed at children, e.g., in children's books, is characterized by a much higher rate of canonical SV(X)-constructions than CDS (Cameron-Faulkner and Noble, 2013). Questions rarely occur in books. CDS produced in shared book reading presents a middle-ground — it contains more complex and SV(X)-constructions than regular CDS, but less than book text alone (Noble et al., 2018). They argue that shared reading therefore, plays an important role in moving children from early, isolated traces of linguistic knowledge to a rich mental language system. This also aligns with the findings by Bunzeck and Diessel (2024), who show that the distribution of constructions in CDS varies with situation type (toy-play features most questions, meal sessions beget more imperatives, shared book reading features more complex constructions) and child age (questions and imperatives become less frequent with age). They suggest that CDS is therefore adapted to support children's cognitive and linguistic development. Yet, as corpus studies are necessarily descriptive and cannot establish causal/mechanistic connections on their own (e.g. what would happen if a child never hears CDS), it remains questionable if this is actually true. Here, the potential of LMs trained on little data becomes apparent for constructionist approaches: they allow controlled experiments with different kinds of input data, which can serve as additional evidence for effects hypothesized from corpus data.

## 3 Input in developmentally plausible LMs

**Authentic data** Early approaches to modeling language acquisition with neural networks used hand-picked, manually ordered data points (Rumelhart and McClelland, 1986) or synthetic data generated with hand-crafted grammars (Elman, 1993; Christiansen and Chater, 1999; Chang et al., 2006). Both lack developmental plausibility. Since then, data availability has improved with the establishment of developmental corpora. Frequently, CDS from CHILDES (MacWhinney, 2000) is used to train developmentally plausible LMs (cf. Pannitto and Herbelot, 2020; Huebner et al., 2021). While CHILDES-based models have the advantage of learning from authentic data *only*, they have the disadvantage of not accessing the *full breadth* of the linguistic input children receive. Children are exposed to many more different registers of language throughout their linguistic development, like shared (or solitary) book reading, or television shows (Montag, 2019; Gowenlock et al., 2024). In response to this, the BabyLM corpora propose a data mix of varied spoken and written sources, from CDS over adult-adult conversations to OpenSubtitles (Lison and Tiedemann, 2016), but also children's (Hill et al., 2015) and adults' books (Gerlach and Font-Clos, 2020). All data included in them could be plausibly encountered by children, which provides opportunities to ablate the influence of architecture/training on the learned linguistic knowledge.

For languages other than English, data availability is the greatest problem for the construction of developmentally plausible datasets. Salhan et al. (2024) and Padovani et al. (2025) use only data available from CHILDES for models in different languages, whereas Prévot et al. (2024) compare models trained on spoken data (child-directed + adult-adult conversations) with models trained on the French Wikipedia. As such, these first forays into more polyglot BabyLMs are still constrained to the child-directed input found in CHILDES and do not extend to the aforementioned variety of inputs (Soderstrom, 2007; Gowenlock et al., 2024). Notably, Suozzi et al. (2025) introduce an Italian BabyLM but do not elaborate on their data sources beyond CHILDES.

**Linguistic properties** The linguistic make-up of pre-training data and its influence on linguistic performance have only recently begun to receive increased scrutiny. Focusing on the lexical level, Yam and Paek (2024) measure sentence-level textual complexity with readability metrics based on text-wide word/syllable–sentence ratios for different corpora (CHILDES, BabyLM corpus, synthetic data, Project Gutenberg). They find that models trained on more complex text perform better at syntactic benchmarks, but simpler data (CHILDES) is learned better in terms of perplexity and loss convergence. Muckatira et al. (2024) filter English pre-training corpora for text spans that only contain vocabulary also found in English CHILDES

| Dataset | Description | # Words |
|---|---|---|
| CHILDES (MacWhinney, 2000) | Child-directed speech | 3,626,301 |
| | Child speech | 1,511,144 |
| OpenSubtitles (Lison and Tiedemann, 2016) | Movie subtitles | 1,543,094 |
| CallHome (Karins et al., 1997) | Phone conversations | 176,313 |
| Klexikon | Children's online encyclopedia | 1,384,891 |
| MiniKlexikon | Simplified online encyclopedia | 272,886 |
| Wikibooks Wikijunior | Educational books | 226,773 |
| Fluter | German youth magazine | 2,862,278 |
| Project Gutenberg | Literature (children's and young adult) | 2,476,133 |
| Dreambank (Domhoff and Schneider, 2008) | Dream reports | 939,197 |
| Leipzig corpus news texts (Goldhahn et al., 2012) | Short news texts | 1,541,803 |
| *Total* | | 16,560,813 |

Table 1: Lexical token counts for all subcorpora of our corpus

data and find that simplified models generate more coherent text than models trained on more complex data, and also succeed in syntactic tests if the test data is filtered accordingly. In contrast, Edman et al. (2024) change the semantic content of the pre-training data and use datasets that approximate the linguistic input second-language learners get, e.g., dictionary entries, grammar books, and paraphrases. While grammar books moderately improve syntactic evaluation, there is no positive effect for the addition of the other text types.

**Filtered corpora** While actual research on the syntactic properties of the input is rather rare, training on filtered corpora has been used in pilot studies. Patil et al. (2024) and Misra and Mahowald (2024) filter out specific grammatical constructions from the BabyLM corpora and then probe the resulting models for knowledge of these grammatical constructions (which might also be analogically learned from related constructions or constructed from their parts). Patil et al. (2024) show that their models succeed on the BLiMP benchmark (Warstadt et al., 2020), even if sentences containing structures targeted in BLiMP's minimal pair sets are removed. Similarly, Misra and Mahowald (2024) show that acceptability scores for the English AANN construction can be reliably estimated from models that have never seen it. In sum, then, models appear to be able to generalize from indirect evidence and learn language in a somewhat constructivist, bottom-up fashion.

The structural composition of child-directed data has (so far) not been scrutinized. Most studies focus on lexical or semantic properties, emphasizing content over structure; child-directed data is usually equated with a somewhat fitting vocabulary or with just being authentic data. However, findings from usage-based linguistics suggest that structural properties, like utterance-level construction distributions, play a crucial role in language acquisition. Whereas Patil et al. (2024) and Misra and Mahowald (2024) remove specific constructions from their data, we aim to explore whether different global distributions of constructions influence the resulting linguistic knowledge and learning trajectories.

## 4 A German BabyLM dataset

To construct a German dataset, we use a variety of developmentally plausible sources, similar to the English BabyLM data (Warstadt et al., 2023; Choshen et al., 2024). We use (1) all data from German CHILDES corpora (MacWhinney, 2000), including frog stories from TalkBank (Berman and Slobin, 1994) and math lessons from ClassBank (Stigler et al., 2000), (2) subtitles from OpenSubtitles (Lison and Tiedemann, 2016), (3) adult conversations from the CallHome corpus (Karins et al., 1997), and (4) written data from Project Gutenberg, from which we downloaded a manually curated sample of children's books, young adult literature and literature commonly read in German schools. We supply this data with two corpora, the DreamBank database of self-reported dreams (Domhoff and Schneider, 2008) and short news texts from the Leipzig corpus (Goldhahn et al., 2012); although they are not child-directed per se, these sources are child-available in everyday language.

To approximate child-available input even better, we tap into freely available child/learner-directed sources and compile four additional subcorpora for our dataset. The Wikibooks Wikijunior shelve features educational resources aimed at children, focusing on a diverse array of topics such as technology or nature. The Klexikon is a children's wiki in German, featuring more than 3,000 articles aimed

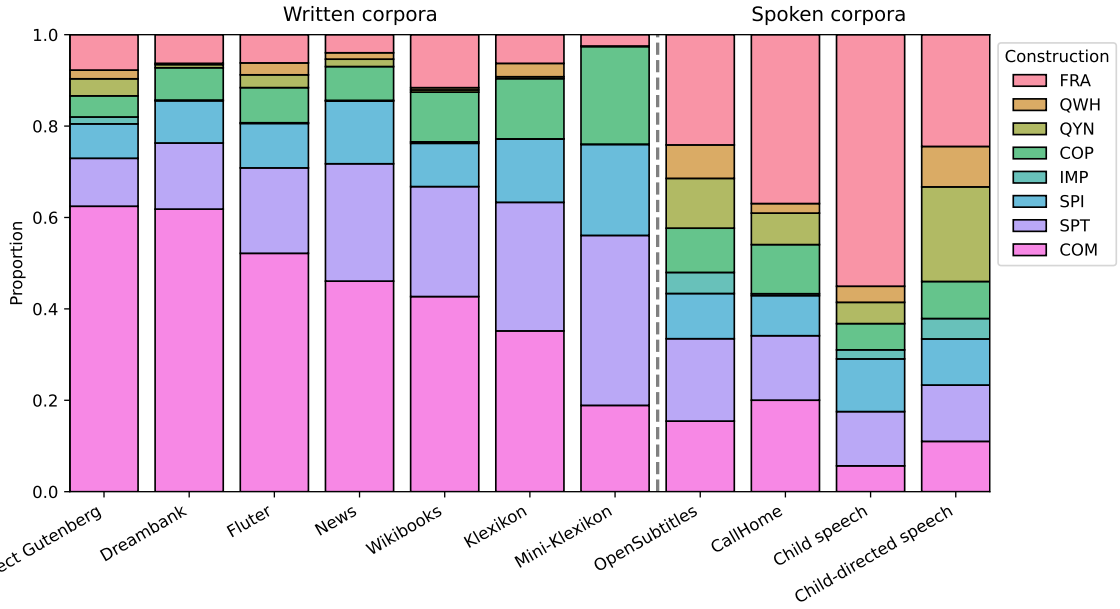



Figure 2: Proportions of utterance-level constructions for all subcorpora in our corpus

at children between 5–15. A simplified version of it is the MiniKlexikon, which features over 1,500 articles aimed at beginning readers. Finally, we also scrape the complete archives of *Fluter*, a magazine aimed at young adults published by the Federal Agency for Civic Education, which contains a large body of non-fiction. All resources are CC-licensed. Table 1 shows the raw token numbers for all corpora (16.5M overall). We extensively clean and normalize our data (details in Appendix B) and make our dataset available on Hugging Face.[1]

## 5 Construction distribution analysis

As there are no findings on the distribution of utterance-level construction in German, we conduct our own analysis using `spacy` (Honnibal et al., 2020). We first split larger paragraphs into individual sentences with the included senter and then annotate these with POS and dependency information. This information serves as the base of our construction annotation procedure. We devise standard construction categories in line with comparable efforts for English (Cameron-Faulkner et al., 2003; Cameron-Faulkner and Noble, 2013; Bunzeck and Diessel, 2024), and assign one of the following categories to each utterance:

- **FRA** – utterances that do not contain a verb
- **QWH** – wh-question (introduced by interrogative pronouns)
- **QYN** – yes/no-question (introduced by verbs/auxiliaries)
- **COP** – subject-predicate utterance where the predicate is a copula verb (a form of *sein* or *werden*)
- **IMP** – utterances introduced by verbs in imperative mood
- **SPI** – standard subject-predicate utterance (intransitive verb with no direct/accusative object)
- **SPT** – standard subject-predicate utterance (transitive verb with direct/accusative object)
- **COM** – utterances with two or more lexical verbs

This holistic taxonomy is applicable to every utterance in our corpus. For a balanced, manually annotated sample of 1,000 sentences our classifier reaches an accuracy of approx. 95%.

Figure 2 visualizes the results of this annotation process, exact proportions are reproduced in Appendix C. Generally, our results confirm earlier findings (Cameron-Faulkner et al., 2003; Cameron-Faulkner and Hickey, 2011; Cameron-Faulkner and Noble, 2013; Bunzeck and Diessel, 2024): Just like English CDS, German CDS features more questions than any other corpus, abounds with fragments, and contains comparatively few complex utterances. The Project Gutenberg data, on the other hand, is characterized by over 60% complex sentences. Interestingly, the construction distribution forms a continuum across our subcorpora. The MiniKlexikon, for example, contains considerably

[1] https://huggingface.co/datasets/bbunzeck/babylm-german

less complex sentences than the other written genres, but over half of its utterances are (in)transitive, canonical SV-sentences. This shows that even these particular sub-genres of child-directed linguistic input feature highly varied and specific constructional profiles that differ from each other.

## 6 Training data composition

We compose three different corpora of 5M words: (1) one corpus maximally resembling the construction composition of child-directed speech (`cds`), (2) one corpus containing a drastically higher amount of complex sentences, mirroring the distribution in the Project Gutenberg data (`pjg`), and (3) a corpus that is averaged between these two (`mix`). The relative distributions of construction types can be found in Table 2.

| Construction | cds | mix | pjg |
|---|---|---|---|
| FRA | 25% | 16.5% | 8% |
| QWH | 9% | 5.5% | 2% |
| QYN | 21% | 12.5% | 4% |
| COP | 8% | 6.5% | 5% |
| IMP | 5% | 3.5% | 2% |
| SPI | 10% | 9% | 8% |
| SPT | 12% | 11% | 10% |
| COM | 10% | 35.5% | 61% |

Table 2: Construction proportions of our training sets

Crucially, we sample the individual utterances for our training sets from all subcorpora in our German BabyLM dataset. By doing so, we approximate a similar (if not completely equal) mixture of sources and, therefore also a similar mixture of registers, semantic content, etc. This enables us to isolate the effect of construction distributions in our model's training data, without any interference from the possible differences between the subcorpora.

## 7 Model training and evaluation

We train small Llama models (Touvron et al., 2023) with `transformers` (Wolf et al., 2020). To account for the effect of subword tokenization, we compare character-level (3.7M parameters) and subword models (7.7M parameters) for the three datasets. We train all models for one epoch (loss curves and hyperparameters are in Appendix D) and share them on Hugging Face.[2] To test the effect of different random initializations and our sampling strategy, we reproduce pre-training for the `cds` models (see Appendix E).

In line with current best practices to linguistic probing, we use minimal pair datasets to evaluate our LMs' linguistic knowledge in German. The datasets always consist of a correct/grammatical and a matched incorrect/ungrammatical string. We use `minicons` (Misra, 2022) to score the sentences and evaluate 19 model checkpoints per model (10 for the first 10% of training, 9 for the remaining 90%). As an additional ablation, we also evaluate the multilingual Llama 3.2 1B[3] on all probing paradigms. Currently, no monolingual German Llama models exist. Therefore, the medium-sized 1B-parameter version of Llama 3.2, which is trained on a considerable amount of German language data, is a useful baseline for expected benchmark scores enabled through a higher model capacity and more training data.

**Word-level probing** Language acquisition first involves learning what words are, i.e. which (sound) sequences map to word-level items in the mental lexicon, before learning how they combine. To gauge this most basic learning step, we adapt the experimental setup from Bunzeck et al. (2025): We use `wuggy` (Keuleers and Brysbaert, 2010) to generate 1,000 nonce words (e.g. *promsen*) from existing words (e.g. *bremsen*) and then evaluate how surprised the models are by (1) the words with the context of a prepended white space (lexical decision, Le Godais et al., 2017), (2) the words in a plausible context sequence (surprisal, Hale, 2001), and (3) the words randomly inserted into implausible contexts (antisurprisal, Shafiabadi and Wisniewski, 2025). If the model is less surprised by the existing word, we count this as a correct choice in our paradigm. We calculate accuracies over the whole dataset.

**Syntactic probing** For syntactic probing, we use the CLAMS dataset (Mueller et al., 2020), which contains syntactic minimal pairs (grammatical/ungrammatical) for German (e.g. *Die Autoren lachen/*lacht.*). The included seven phenomena all revolve around subject-verb agreement in different contexts (across PPs, relative clauses, with coordination, etc.), resulting in different degrees of difficulty. We score the sentences for their likelihood. We calculate accuracies for correctly rated

[2] https://huggingface.co/collections/bbunzeck/german-babylm-67b868e08ff8782a9814ceaf

[3] https://huggingface.co/meta-llama/Llama-3.2-1B

| | | Character | | | Subword | | | Llama 3.2 1B |
|---|---|---|---|---|---|---|---|---|
| | | cds | mix | pjg | cds | mix | pjg | – |
| Word-level | Lexical decision | 97.4% | **97.6%** | 97.4% | 84.6% | 81.9% | 80.8% | 69.6% |
| | Surprisal | 99.8% | 99.8% | **99.9%** | 91.5% | 90.3% | 90.1% | 98% |
| | AntiSurprisal | 99.3% | 98.9% | **99.7%** | 76.5% | 75.4% | 75.4% | 87.4% |
| Syntax | Simple Agreement | 90% | 90% | **95.7%** | 80% | 84.3% | 92.1% | 95.71% |
| | Across a Prepositional Phrase | 61.5% | 65.5% | 61.8% | 74.8% | 73.5% | **75.5%** | 83% |
| | Across a Subject Relative Clause | 67.1% | 66% | 62.4% | 78.4% | 73.7% | **97.9%** | 99.7% |
| | Short Verb Phrase Coordination | 69.8% | 68.8% | 67.9% | 82.6% | 93.5% | **99.5%** | 99.9% |
| | Long Verb Phrase Coordination | 53.6% | 60.6% | 63% | 60.6% | **78.8%** | 78% | 90.5% |
| | Across Object Relative Clause | 58.6% | 54.2% | 53% | 64% | 66.7% | **81.6%** | 86.1% |
| | Within Object Relative Clause | 59.8% | 56.4% | **72.5%** | 55.8% | 55.7% | 49.9% | 61.4% |
| Semantics | XCOMPS | 51.5% | 49.1% | 49.1% | 51.4% | 52% | **52.3%** | 58.9% |

Table 3: Final evaluation results (accuracies) for all benchmarks

pairs (grammatical sentence more likely) over the whole dataset.

**Semantic probing** To evaluate our models' semantic knowledge, we use the XCOMPS dataset (He et al., 2025). It contains conceptual minimal pairs (e.g. *Garnele hat einen Kopf./*Ein Bikini hat einen Kopf.*)[4] that test whether LMs have acquired knowledge about conceptual properties of real-world entities. Again, we score the sentences for likelihood and calculate accuracy over the whole dataset.

## 8 Results

### 8.1 MP probing

Table 3 shows model-wise accuracies for all minimal pair sets after training for one epoch. For the word-level evaluations, accuracy scores are generally high. Across all tasks, the character models perform with almost perfect accuracy. No effect of the constructional composition of the training data is identifiable here. For the subword models, this is not true. Here, the model trained on more questions/fragments and less complex utterances (cds) outperforms the model that approximates written language on the construction level (pjg). The improvements range from 1% for anti-surprisal to 2-3% on lexical decision. Interestingly, the very large ablation model (Llama 3.2 1B) performs the worst on isolated lexical decision, but reaches high scores in the surprisal setting.

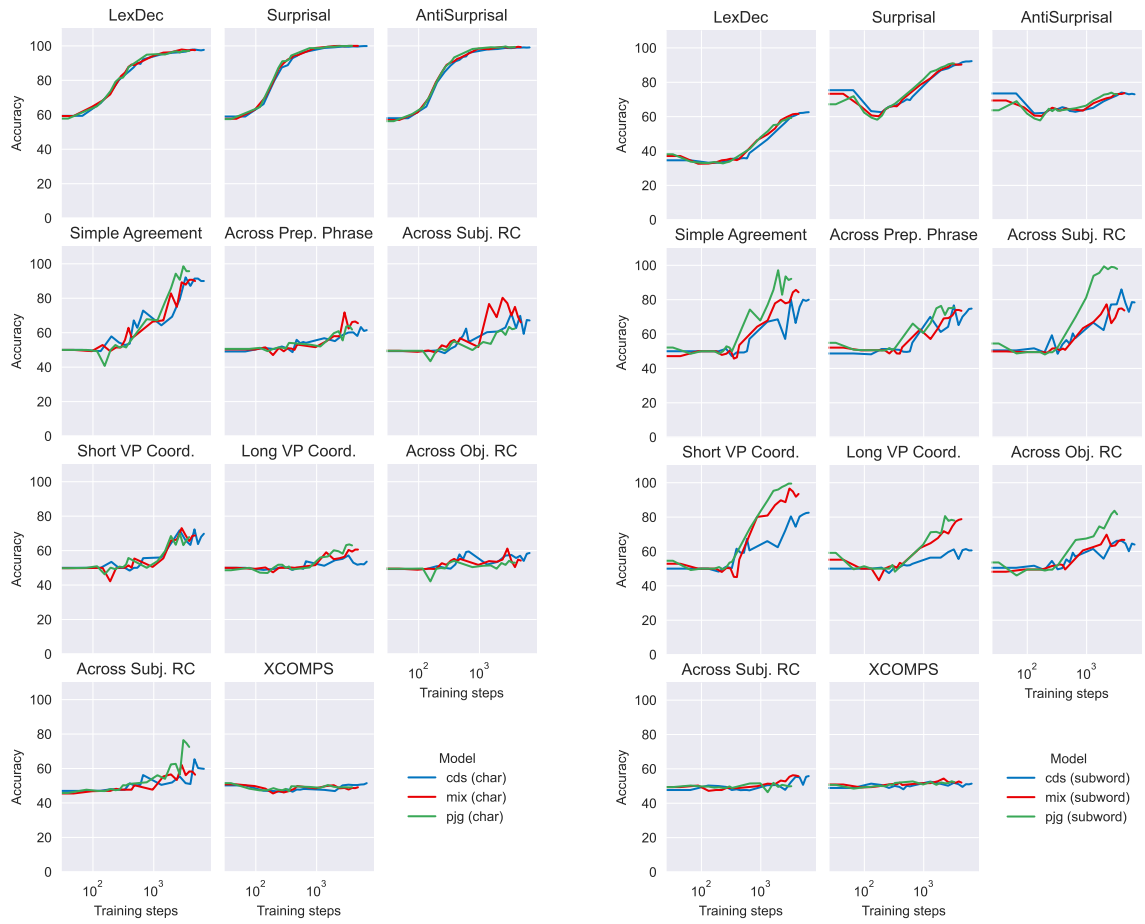
For the syntactic tests, the picture is more nuanced. Generally speaking, all our models learn to distinguish most types of grammatical and ungrammatical sentences involving agreement phenomena. The best scores are achieved on more simplistic phenomena like simple agreement or coordination with short verb phrases. Agreement phenomena that involve longer dependencies and distracting nouns, e.g. within and across relative clauses, are the hardest to learn. For the character models, the cds model outperforms the others on three out of seven tests, including both "across subj./obj. relative clause" conditions. For three other tests, the pjg model wins out, whereas the mix model achieves the highest scores on only one test (agreement across prepositional phrases). It should be noted, that for most phenomena, the character models do perform well above chance (by a margin of 10–20%), but still frequently make errors. The subword models show a somewhat different picture, with scores being generally higher and approximating perfect performance on 3/7 phenomena. Regarding construction distributions, the pjg model wins in five categories, whereas cds and mix only achieve best scores in one each. Here, the 1B-parameter Llama model outperforms our BabyLMs on 5/7 phenomena.

The scores on XCOMPS reveal that our small models do not reliably learn the conceptual knowledge underlying the included minimal pairs. Scores revolve around the chance baseline, with subword models performing slightly better than character models for 2/3 data mixtures. Nonetheless, these scores are also not considerably worse than the performance of our ablation model (58.9%).

### 8.2 Learning trajectories

Figure 3 shows the learning trajectories of our models across one training epoch. As there are no intermediate checkpoints available for the 1B-parameter ablation model, we only report trajectories for our

[4]We sample 1,000 MPs with randomized replacement, as the other conditions contain implausible/wrong minimal pairs. Furthermore, the quality of translation is not optimal, as exemplified by the missing determiner in front of *Garnele*.




(a) Trajectories for character models

(b) Trajectories for subword models

Figure 3: Learning trajectories for all minimal pair benchmarks

self-trained models. In line with best practices in ML (Viering and Loog, 2023), we log-scale the x-axis in our plots. This allows us to also trace early learning in more detail.

For our character models, word-level learning happens rapidly in an S-shaped curve. No differences are visible between the datasets, performance improvements align almost perfectly. For the subword models, the learning processes are not as nicely monotonically improving. Rather, the learning trajectories show a dip early in training, which then later on recovers to fairly good accuracy scores. Interestingly, despite differences in final scores, the improvements across models trained on quite different datasets still align with regard to turning and takeoff points.

This pattern is also confirmed by the learning trajectories for the syntactic phenomena. While the `pjg` models trained on more complex utterances frequently reach the highest final scores, it is remarkable to see how the improvements for all models seem to happen in parallel. The global shape of the trajectory is the same for all syntactic tests, regardless of the construction distribution. For example, the learning curve for simple agreement is steeper for the `pjg` models once learning has started, but take-off points are neatly aligned. These take-off points are pushed back by the individual paradigms' complexities — simple agreement and short VP coordination begin to improve earlier than MPs containing RCs. Finally, it is interesting to note that for the character models, word-level learning consistently stabilizes before syntactic learning, whereas both processes seem to happen concurrently in subword models (mirroring findings for English, cf. Bunzeck and Zarrieß, 2025). As our models do not learn to distinguish the semantic minimal pairs, the corresponding learning curves remain flat and performance differences are likely due to chance.

## 9 Discussion

This paper set out to investigate whether the constructional profile of CDS, which is shaped in a way to support the acquisition of *functional* language

competence, actually influences LMs' *formal* language learning, and whether its relative lack of complex sentences and canonical SV(X) utterances makes it less useful training data, or too "impoverished" for meaningful formal learning to happen. The results of our utterance-level corpus analysis for German align with earlier findings on CDS and book language for English (Cameron-Faulkner et al., 2003; Cameron-Faulkner and Noble, 2013; Bunzeck and Diessel, 2024) and Irish (Cameron-Faulkner and Hickey, 2011), adding to the growing evidence that this linguistic distribution is fairly universal, at least in WEIRD societies (Henrich, 2024).

From a language modeling perspective, the constructional profile of training data is *not* overly important for the resulting performance on linguistic benchmarks. Rather, starting/turning points of the resulting learning trajectories are mostly determined by the respective amount of training steps. Despite models trained with more complex input resulting in slightly better performance, they do not begin to learn earlier. Global learning trajectories are extremely similar, only the local magnitude differs between different constructional setups. This provides further evidence that LMs based on the Transformer architecture (Vaswani et al., 2017) not only memorize language from their training data, but generalize to the underlying patterns. The same holds true or word-level learning processes such as lexical decision or (anti)surprisal tests, where data with more fragments and questions even seems to be rather beneficial. Furthermore, the comparison of our results to the Llama 3.2 1B model shows that rather high scores are already attainable with small models and little data (only on long VP-coordination do our models underperform).

What does this now mean for theories of language acquisition? This study was inspired by findings of construction-based corpus analyses (Cameron-Faulkner et al., 2003; Cameron-Faulkner and Hickey, 2011; Bunzeck and Diessel, 2024), which argue that the specific constructional profile of CDS is beneficial to acquisition. Of course, LMs and minimal pair evaluations do not directly correspond to the learning processes in humans and we cannot make causal claims about them. Yet, our methodology can provide evidence as to what kinds of input data is beneficial to a purely statistical learner (that does not even tap into the functional side of language, cf. Mahowald et al., 2024), an abstraction that is highly relevant to usage-based theories (Ambridge et al., 2015). On a formal level, there seem to be comparatively little disadvantages for models trained on less "complex" or somewhat impoverished data. Despite more complex data leading to slightly better benchmark scores, the learning trajectories remain largely unaffected (although somewhat erratic, cf. Bunzeck and Zarrieß, 2024). What really shapes the learning process in our LMs is the amount of input, not its formal complexity (similar to findings for children by Huttenlocher et al., 1991; Rowe, 2012). An increase in appropriate construction types for child-rearing (like questions, imperatives, or fragments) does not hinder formal learning (if only reduce its magnitude slightly). As CLAMS only focuses on subject-verb agreement in canonical SV(X)-sentences, it is rather surprising that the much higher amount of questions in the `cds` dataset does not negatively affect performance, although the subjects' and predicates' positions are switched in German yes/no-questions. Conversely, the `cds` dataset even enables word-level learning to converge to a better end state. This also aligns with a broader trend found in language acquisition studies — the complexity and quality of input can indeed predict later language skills (Noble et al., 2020; Alroqi et al., 2023), but the ground level is always extremely high already: being a competent user of the language itself. Furthermore, quality varies with many more extralinguistic factors like the number of siblings (Laing and Bergelson, 2024) or cultural factors (Bergelson et al., 2023; Bunce et al., 2024).

## 10 Conclusion

Our findings add to the growing body of research on BabyLMs (Warstadt et al., 2023; Hu et al., 2024). Similarly to English models, our German BabyLMs only need little data — the `cds` dataset contains approx. 820,000 sentences, and given the estimation by Cameron-Faulkner et al. (2003) that children hear around 7,000 utterances per day, our data approximates the number of utterances heard over only 120 days — to learn a fair amount of syntax and almost impeccable lexical knowledge, with trajectories mirroring those of English models (Bunzeck and Zarrieß, 2025). We hope that our dataset enables other scholars to carry out experiments with developmentally plausible LMs beyond the dominating English LMs, and that our data provides inspiration to those compiling BabyLM corpora for other languages.

## Limitations

Our study is limited by data availability. Creating a full-fledged 100M-token BabyLM dataset with *only* child-directed speech or other explicitly child-directed materials is currently out of question, as neither CHILDES nor other sources contain even remotely enough data for languages other than English. To reach higher token counts, padding with larger data sets, e.g. more tokens from the OpenSubtitles dataset, would be necessary. 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 specifically aimed at German, beyond those that we included/creates ourselves, especially on the syntactic level. Only very recently, evaluation datasets like the massively multilingual MultiBLiMP have begun to fill this gap (Jumelet et al., 2025). Also, such minimal pair datasets are principally at odds with the usage-based, constructionist view on language development, because they are grounded in the Generativist notion of defining rules that can determine whether an utterance belongs to a language or not, whereas usage-based linguistics has adopted a network-based, associative model of linguistic knowledge (Diessel, 2019, 2023). As of late, these developments have begun to make their way into the broader LM evaluation landscape (Weissweiler et al., 2025), and novel evaluation methods like measuring affinities between lexical items and testing if different constructions manifest from them (Rozner et al., 2025a,b) provide promising future research avenues.

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; Goriely and Buttery, 2025b). 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.

## Ethical considerations

Given the nature of this work, there are no specific ethical concerns to address. However, we would like to stress that, of course, BabyLMs are not supposed to simulate real babies, but rather to instantiate 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.

## Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — CRC-1646, project number 512393437, project A02.

## References

Haifa Alroqi, Ludovica Serratrice, and Thea Cameron-Faulkner. 2023. The association between screen media quantity, content, and context and language development. *Journal of Child Language*, 50(5):1155–1183.

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Ben Ambridge and Elena Lieven. 2011. *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge University Press, Cambridge ; New York.

Heike Behrens. 2021. Constructivist Approaches to First Language Acquisition. *Journal of Child Language*, 48(5):959–983.

Elika Bergelson, Melanie Soderstrom, Iris-Corinna Schwarz, Caroline F. Rowland, Nairán Ramírez-Esparza, Lisa R. Hamrick, Ellen Marklund, Marina Kalashnikova, Ava Guez, Marisa Casillas, Lucia Benetti, Petra Van Alphen, and Alejandrina Cristia. 2023. Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, 120(52):e2300671120.

Ruth A Berman and Dan I. Slobin. 1994. *Different Ways of Relating Events in Narrative: A Crosslinguistic Study*. Erlbaum Associates, Hillsdale, NJ.

Robert C. Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the Stimulus Revisited. *Cognitive Science*, 35(7):1207–1242.

Silke Brandt, Holger Diessel, and Michael Tomasello. 2008. The acquisition of German relative clauses: A case study*. *Journal of Child Language*, 35(2):325–348.

John Bunce, Melanie Soderstrom, Elika Bergelson, Celia Rosemberg, Alejandra Stein, Florencia Alam, Maia Julieta Migdalek, and Marisa Casillas. 2024. A cross-linguistic examination of young children's everyday language experiences. *Journal of Child Language*, pages 1–29.

Bastian Bunzeck and Holger Diessel. 2024. The richness of the stimulus: Constructional variation and development in child-directed speech. *First Language*, 45(2):152–176.

Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarrieß. 2025. Small language models also work with small vocabularies: Probing the linguistic abilities of grapheme- and phoneme-based baby llamas. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6039–6048, Abu Dhabi, UAE. Association for Computational Linguistics.

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

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

Thea Cameron-Faulkner and Tina Hickey. 2011. Form and function in Irish child directed speech. *Cognitive Linguistics*, 22(3):569–594.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science*, 27(6):843–873.

Thea Cameron-Faulkner and Claire Noble. 2013. A comparison of book text and Child Directed Speech. *First Language*, 33(3):268–279.

Franklin Chang, Gary S. Dell, and Kathryn Bock. 2006. Becoming syntactic. *Psychological Review*, 113(2):234–272.

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*.

Jidong Chen and Yasuhiro Shirai. 2015. The acquisition of relative clauses in spontaneous child speech in Mandarin Chinese. *Journal of Child Language*, 42(2):394–422.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Number 11 in Massachusetts Institute of Technology. Research Laboratory of Electronics. Special Technical Report. MIT Press, Cambridge, Massachusetts.

Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences*, 3(1):1–15.

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [Call for Papers] The 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Preprint*, arXiv:2404.06214.

Morten H Christiansen and Nick Chater. 1999. Toward a Connectionist Model of Recursion in Human Linguistic Performance. *Cognitive Science*, 23(2):157–205.

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*, 1 edition. Wiley.

Stephen Crain and Paul Pietroski. 2001. Nature, Nurture And Universal Grammar. *Linguistics and Philosophy*, 24(2):139–186.

Alejandrina Cristia, Emmanuel Dupoux, Nan Bernstein Ratner, and Melanie Soderstrom. 2019. Segmentability Differences Between Child-Directed and Adult-Directed Speech: A Systematic Test With an Ecologically Valid Corpus. *Open Mind*, 3:13–22.

Holger Diessel. 2019. *The Grammar Network*. Cambridge University Press, Cambridge.

Holger Diessel. 2023. *The Constructicon: Taxonomies and Networks*, 1 edition. Cambridge University Press.

Holger Diessel and Michael Tomasello. 2000. The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, 11(1-2):131–151.

G. William Domhoff and Adam Schneider. 2008. Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition*, 17(4):1238–1247.

Lukas Edman, Lisa Bylinina, Faeze Ghorbanpour, and Alexander Fraser. 2024. Are BabyLMs second language learners? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 166–173, Miami, FL, USA. Association for Computational Linguistics.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *Preprint*, arXiv:2305.07759.

Jeffrey L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

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.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Giuliana Genovese, Maria Spinelli, Leonor J. Romero Lauro, Tiziana Aureli, Giulia Castelletti, and Mirco Fasolo. 2020. Infant-directed speech as a simplified but not simple register: A longitudinal study of lexical and syntactic features. *Journal of Child Language*, 47(1):22–44.

Martin Gerlach and Francesc Font-Clos. 2020. A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics. *Entropy*, 22(1):126.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Zébulon Goriely and Paula Buttery. 2025a. BabyLM's First Words: Word Segmentation as a Phonological Probing Task. *arXiv preprint*.

Zébulon Goriely and Paula Buttery. 2025b. IPA-CHILDES & G2P+: Feature-Rich Resources for Cross-Lingual Phonology and Phonemic Language Modeling. *arXiv preprint*.

Zébulon Goriely, Richard Diehl Martinez, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 37–53, Miami, FL, USA. Association for Computational Linguistics.

Anna Elizabeth Gowenlock, Courtenay Norbury, and Jennifer M. Rodd. 2024. Exposure to Language in Video and its Impact on Linguistic Development in Children Aged 3–11: A Scoping Review. *Journal of Cognition*, 7(1):57.

Maria Teresa Guasti. 2002. *Language Acquisition: The Growth of Grammar*. MIT Press, Cambridge, Mass.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Adrian Florea, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, Hinrich Schütze, and Nima Mesgarani. 2025. XCOMPS: A Multilingual Benchmark of Conceptual Minimal Pairs. *arXiv preprint*.

Joseph Henrich. 2024. WEIRD. In *Open Encyclopedia of Cognitive Science*, 1 edition. MIT Press.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *Preprint*, arXiv:1511.02301.

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

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. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

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, Wendy Haight, Anthony Bryk, Michael Seltzer, and Thomas Lyons. 1991. Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, 27(2):236–248.

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

Daniel Jach and Gunther Dietz. 2024. KORPUS EINFACHES DEUTSCH (KED). *Korpora Deutsch als Fremdsprache*, 4.

Laura A. Janda, editor. 2013. *Cognitive Linguistics: The Quantitative Turn: The Essential Reader*. Mouton Reader. De Gruyter Mouton, Berlin.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. *arXiv preprint*.

Krisjanis Karins, Robert MacIntyre, Monika Brandmair, Susanne Lauscher, and Cynthia McLemore. 1997. CALLHOME German Transcripts.

Emmanuel Keuleers and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German reference corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Catherine Laing and Elika Bergelson. 2024. Analyzing the effect of sibling number on input and output in the first 18 months. *Infancy*, 29(2):175–195.

Marvin Lavechin, Maureen De Seyssel, Hadrien Titeux, Guillaume Wisniewski, Hervé Bredin, Alejandrina Cristia, and Emmanuel Dupoux. 2025. Simulating Early Phonetic and Word Learning Without Linguistic Categories. *Developmental Science*, 28(2):e13606.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. BabySLM: Language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*, pages 4588–4592. ISCA.

Gaël Le Godais, Tal Linzen, and Emmanuel Dupoux. 2017. Comparing character-level neural language models using a lexical decision task. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 125–130, Valencia, Spain. Association for Computational Linguistics.

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).

Bria Long, Violet Xiang, Stefan Stojanov, Robert Z. Sparks, Zi Yin, Grace E. Keene, Alvin W. M. Tan, Steven Y. Feng, Chengxu Zhuang, Virginia A. Marchman, Daniel L. K. Yamins, and Michael C. Frank. 2024. The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. *arXiv preprint*.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Brian MacWhinney. 2004. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, 31(4):883–914.

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

Raphaël Millière. 2024. Language Models as Models of Language. *arXiv preprint*.

Kanishka Misra. 2022. Minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint*.

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

Jessica L. Montag. 2019. Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language*, 39(5):527–546.

Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. 2024. Emergent Abilities in Reduced-Scale Generative Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1242–1257, Mexico City, Mexico. 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.

Claire Noble, Thea Cameron-Faulkner, Andrew Jessop, Anna Coates, Hannah Sawyer, Rachel Taylor-Ims, and Caroline F. Rowland. 2020. The Impact of Interactive Shared Book Reading on Children's Language Skills: A Randomized Controlled Trial. *Journal of Speech, Language, and Hearing Research*, 63(6):1878–1897.

Claire H. Noble, Thea Cameron-Faulkner, and Elena Lieven. 2018. Keeping it simple: The grammatical properties of shared book reading. *Journal of Child Language*, 45(3):753–766.

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

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: Evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176, Online. Association for Computational Linguistics.

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.

Steven T. Piantadosi. 2024. Modern language models refute Chomsky's approach to language. In Edward Gibson and Moshe Poliak, editors, *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, Empirically Oriented Theoretical Morphology and Syntax. Language Science Press, Berlin.

Julian M. Pine. 1994. The language of primary caregivers. In Clare Gallaway and Brian J. Richards, editors, *Input and Interaction in Language Acquisition*, 1 edition, pages 15–37. Cambridge University Press.

Laurent Prévot, Sheng-Fu Wang, Jou-An Chi, and Shu-Kai Hsieh. 2024. Extending the BabyLM initiative : Promoting diversity in datasets and metrics through high-quality linguistic corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 147–158, Miami, FL, USA. Association for Computational Linguistics.

Geoffrey K Pullum and Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 18(1-2).

Silke Reineke, Arnulf Deppermann, and Thomas Schmidt. 2023. Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch (FOLK). In Arnulf Deppermann, Christian Fandrych, Marc Kupietz, and Thomas Schmidt, editors, *Korpora in Der Germanistischen Sprachwissenschaft*, pages 71–102. De Gruyter.

Meredith L. Rowe. 2012. A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, 83(5):1762–1774.

Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025a. Constructions are Revealed in Word Distributions. *arXiv preprint*.

Joshua Rozner, Leonie Weissweiler, and Cory Shain. 2025b. BabyLM's First Constructions: Causal interventions provide a signal of learning. *arXiv preprint*.

David E. Rumelhart and James L. McClelland. 1986. On Learning the Past Tenses of English Verbs. In *Parallel Distributed Processing*, volume 2, pages 535–551. MIT Press, Cambridge, MA.

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. *arXiv preprint*.

Nazanin Shafiabadi and Guillaume Wisniewski. 2025. Beyond surprisal: A dual metric framework for lexical skill acquisition in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6636–6641, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine E. Snow and Charles A. Ferguson, editors. 1977. *Talking to Children: Language Input and Acquisition*. Cambridge University Press, Cambridge, MA.

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

James W. Stigler, Ronald Gallimore, and James Hiebert. 2000. Using Video Surveys to Compare Classrooms and Teaching Across Cultures: Examples and Lessons From the TIMSS Video Studies. *Educational Psychologist*, 35(2):87–100.

Alice Suozzi, Luca Capone, Gianluca E. Lebani, and Alessandro Lenci. 2025. BAMBI: Developing Baby Language Models for Italian. *arXiv preprint*.

Shira Tal, Eitan Grossman, and Inbal Arnon. 2024. Infant-directed speech becomes less redundant as infants grow: Implications for language learning. *Cognition*, 249:105817.

Margaret Thomas. 2002. Development of the concept of "the poverty of the stimulus". *The Linguistic Review*, 18(1-2).

Michael Tomasello. 2005. Beyond formalities: The case of language acquisition. *The Linguistic Review*, 22(2-4).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Mark VanDam, Anne Warlaumont, Elika Bergelson, Alejandrina Cristia, Melanie Soderstrom, Paul De Palma, and Brian MacWhinney. 2016. HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings. *Seminars in Speech and Language*, 37(02):128–142.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tom Viering and Marco Loog. 2023. The Shape of Learning Curves: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819.

Wai Keen Vong, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.

Wentao Wang, Wai Keen Vong, Najoung Kim, and Brenden M. Lake. 2023. Finding Structure in One Child's Linguistic Experience. *Cognitive Science*, 47(6):e13305.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–6, Singapore. Association for Computational Linguistics.

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

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction Grammar Provides Unique Insight into Neural Language Models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95.

Leonie Weissweiler, Kyle Mahowald, and Adele Goldberg. 2025. Linguistic Generalizations are not Rules: Impacts on Evaluation of LMs. *arXiv preprint*.

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.

Hong Meng Yam and Nathan Paek. 2024. What should baby models read? Exploring sample-efficient data composition on model performance. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 284–291, Miami, FL, USA. Association for Computational Linguistics.

Charles D. Yang. 2004. Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10):451–456.

Daniel Yurovsky, Chen Yu, and Linda B. Smith. 2012. Statistical Speech Segmentation and Word Learning in Parallel: Scaffolding from Child-Directed Speech. *Frontiers in Psychology*, 3.

Renate Zangl and Debra L. Mills. 2007. Increased Brain Activity to Infant-Directed Speech in 6- and 13-Month-Old Infants. *Infancy*, 11(1):31–62.

## A 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.

## B Data cleaning

In line with best practices in language modeling, we extensively clean and normalize our data.

**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.

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

## C Exact construction proportions

Table 4 shows the exact construction proportions for all of our subcorpora. This data underlies the visualization in Figure 2.

| 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

## D 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 (üääöß 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 4. 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. It should be noted that due to longer/shorter sequences in the different data mixtures and our choice of padding to the maximum sequence length, some models are trained for more *steps*, although the number of *lexical tokens* remains the same.

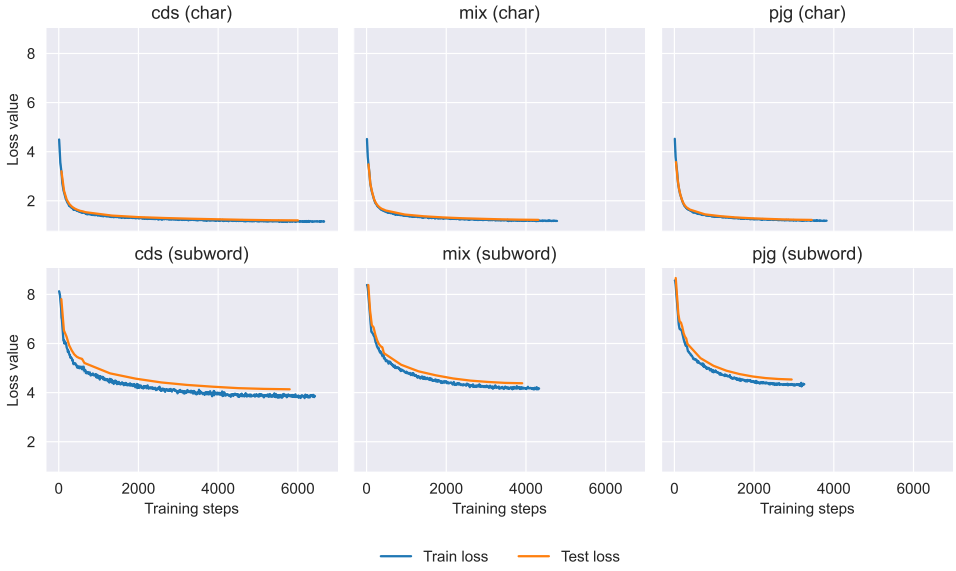
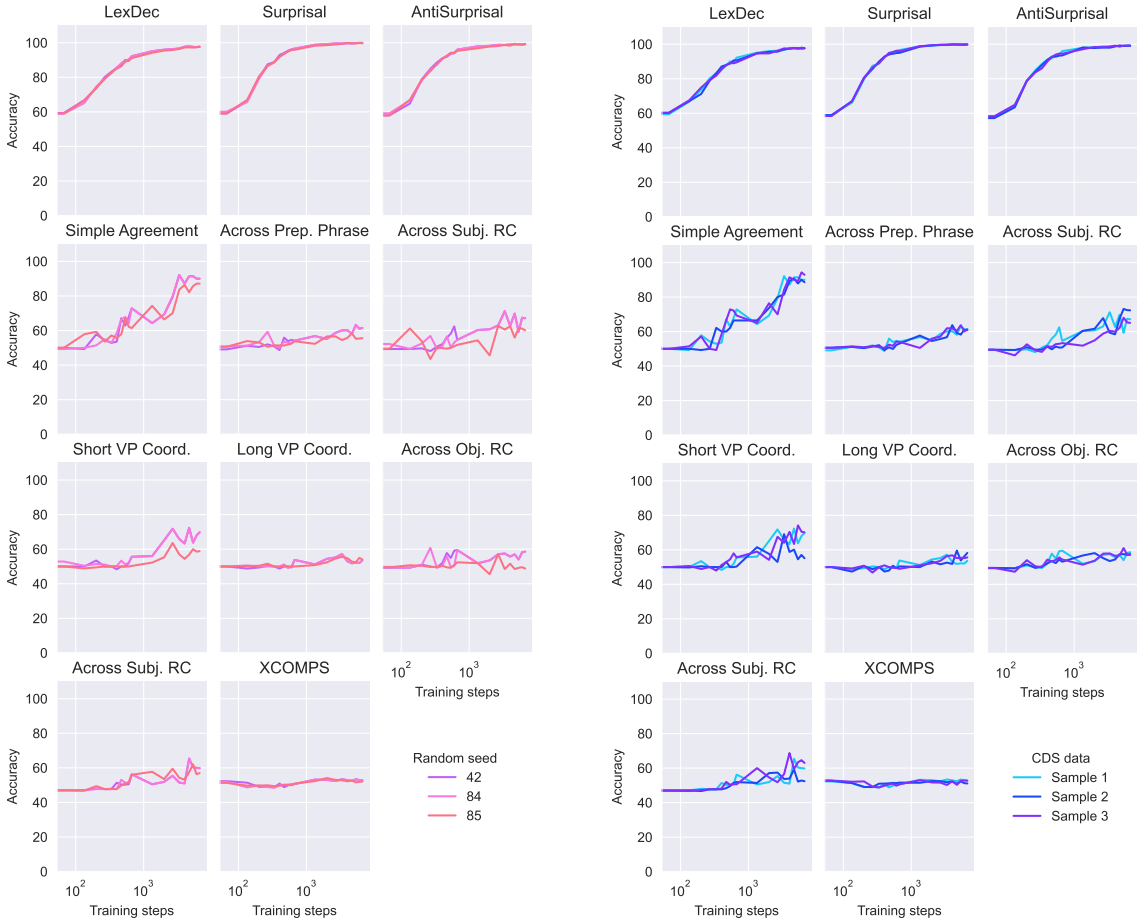



Figure 4: Loss curves for our self-trained character and subword models

## E Repeated training runs

A common criticism towards 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 5a) 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 5b). 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 3 between the datasets with different construction compositions).




(a) Trajectories for different random initializations

(b) Trajectories for different samples of cds data

Figure 5: Learning trajectories for our comparison models