Less is More: Pre-Training Cross-Lingual Small-Scale Language Models with Cognitively-Plausible Curriculum Learning Strategies

Suchir Salhan 🍋🍊 **Richard Diehl Martinez** 🍋 **Zébulon Goriely** 🍋 **Paula Buttery** 🍋🍊
🍋 Department of Computer Science & Technology, University of Cambridge, U.K.
🍊 ALTA Institute, University of Cambridge, U.K.
{sas245,rd654,zg258,pjb48}@cam.ac.uk

Abstract

Curriculum Learning has been a popular strategy to improve the cognitive plausibility of Small-Scale Language Models (SSLMs) in the BabyLM Challenge. However, it has not led to considerable improvements over non-curriculum models. We assess whether theoretical linguistic acquisition theories can be used to specify more fine-grained curriculum learning strategies, creating age-ordered corpora of Child-Directed Speech for four typologically distant language families to implement SSLMs and acquisition-inspired curricula cross-lingually. Comparing the success of three objective curricula (GROWING, INWARDS and MMM) that precisely replicate the predictions of acquisition theories on a standard SSLM architecture, we find fine-grained acquisition-inspired curricula can outperform non-curriculum baselines and performance benefits of curricula strategies in SSLMs can be derived by specifying fine-grained language-specific curricula that precisely replicate language acquisition theories.


https://github.com/suchirsalhan/MAO-CLIMB (CC BY 4.0)


https://huggingface.co/climb-mao (CC BY 4.0)

1 Introduction

Curriculum Learning (CL) has emerged as a promising method to improve the cognitive plausibility of **Small-Scale Language Models (SSLMs)** in the first BabyLM Challenge (Warstadt et al., 2023), as a way to gradually introduce more complex linguistic phenomena into the model later in training in a manner that is similar to human language acquisition. Cognitively-inspired SSLMs are models trained on corpora that approximate the volume and nature of input that a first-language learner can expect to receive during language acquisition. These have been found to perform competitively against LLMs in English (Huebner et al., 2021). CL strategies implemented in the BabyLM Challenge either specified a static measure of linguistic complexity, such as lexical frequency (Borazjanizadeh, 2023), sorted datasets according to difficulty (Opper et al., 2023), or gradually increased vocabulary sizes (Edman and Bylinina, 2023). While the majority of these strategies did not yield consistent improvements over non-curriculum learning baselines (Warstadt et al., 2023), linguistic theory suggests that children naturally focus on input that is neither too simple nor too difficult but at the right level of challenge for learning (Biberauer, 2019; Bosch, 2023). This is known as the "Goldilocks Effect", which is a form of self-selecting curriculum learning that appears to naturally occur in first language (L1) acquisition. This raises the question of whether acquisition theories can provide insights into more effective curriculum learning strategies for SSLMs, and lead to more consistent benefits of CL strategies.

Our work assesses whether language acquisition theories can provide us with better heuristics for good curriculum learning strategies to train SSLMs. We compare contrastive acquisition theories for their success when informing objective curriculum learning strategies on a standard architecture (Diehl Martinez et al., 2023). We train SSLMs with three new objective curricula called GROWING, INWARDS and MMM, each replicating the developmental sequences of contemporary acquisition theories that first-language monolingual learners are theorised to follow in the earliest stages of acquisition cross-linguistically. In practice, these curricula modify the standard masked language modelling objective in BabyBERTa-style models by varying the order and the sequence of masking using different tagsets to simulate different language acquisition theories.

The acquisition models specify different cross-lingual and language-specific developmental sequences that learners appear to follow in first language acquisition, which has not been implemented

or evaluated in the context of Deep Learning. The multilingual focus of the acquisition models is a goal strongly aligned with the spirit of the BabyLM Shared Task. We train SSLMs with these objective curricula for four typologically distant language families: Romance (French), Germanic (German), Japonic (Japanese) and Sino-Tibetan (Chinese). We introduce new age-ordered corpora of Child-Directed Speech (CDS) for these languages and select languages for pre-training based on the quantity of CDS that can be used to train SSLMs using similar volumes of data that learners can utilise in first language acquisition. We evaluate these SSLMs on syntactic minimal pair datasets. We find benefits of the cognitively-inspired objective curricula cross-linguistically, however different strategies lead to better performance for certain languages, particularly finer-grained language-specific versions of the MMM objective. Acquisition-inspired objective curricula can obtain comparable performance on minimal pair evaluation datasets to LLMs, despite requiring approximately 25X fewer parameters and 6,000X fewer words.

2 Background

We survey Curriculum Learning (CL) strategies used in the 1st BabyLM Challenge *Section* 2.1 and contrastive models of syntactic acquisition that are utilised to replicate cross-lingual developmental sequences for implementing more cognitively plausible pre-training in SSLMs in *Section* 2.2.

2.1 Curriculum Learning Strategies for Pre-training on Developmentally Plausible Corpora

While some SSLMs that utilised CL strategies outperformed the official BabyLM baselines, no CL strategies led to consistent or uniform improvements compared to stronger non-curriculum models. Many submissions for the inaugural BabyLM Challenge utilised Curriculum Learning on a small-scale masked language model architecture trained on a 5 million (5M) word corpus called BABYBERTA (Huebner et al., 2021), based on a Transformer Language Model ROBERTA (Liu et al., 2019) with $15\times$ fewer parameters, which displayed comparable grammatical capabilities to ROBERTA. In general, CL strategies, like using a pre-defined static difficulty assessment based on linguistic criteria like syntax dependency tree depth (Oba et al., 2023) or ranking sentences according to surprisal (Chobey et al., 2023) or length (DeBenedetto, 2023) or other measures of difficulty (Opper et al., 2023), showed little improvement over non-CL baselines. Diehl Martinez et al. (2023) introduce **Curriculum Learning for Infant-Inspired Model Building (CLIMB)**, which incorporates three CL strategies into BabyBERTa pre-training that each dynamically increase the difficulty of the language modelling task throughout training. CLIMB's **vocabulary curriculum** constrains the Transformer vocabulary in the initial stages of training by dynamically mask out vocabulary units over training. CLIMB's **data curriculum** varies the order of training instances based on infant-inspired expectations and the learning behaviour of the model, enabling dynamic sampling of training data according to a difficulty function. CLIMB's **objective curriculum** combines the masked language modelling task, used in RoBERTa (Liu et al., 2019) and the BabyBERTa model (Huebner et al., 2021), with coarse-grained word class prediction to reinforce linguistic generalisation capabilities. This provides functionality to change the objective function at specified discrete training steps. The objective curricula modifies the Masked Language Modelling (MLM) objective, which is the standard "denoising" objective for Pre-trained Language Models, like ROBERTA and BABYBERTA. Both models use a random token masking strategy, applying a fixed masking ratio α to mask different contexts selected randomly with a probability P_i. Diehl Martinez et al. (2023) introduce two objective curricula defined using 'curriculum units' of Universal Part of Speech (UPOS) tags. The first objective classifies [MASK] to one of [VERB, NOUN, OTHER], while the second objective classifies [MASK] to one of the 10 UPOS tags. CLIMB's objective curricula, following the submission guidelines of the 1st BabyLM Challenge, are performed using an unsupervised part-of-speech (POS) tagger. They additionally tuned the vocabulary and model size of BabyBERTa, resulting in a model that outperformed the official baselines for the first BabyLM Challenge. CLIMB's curriculum learning strategies outperformed the official baseline but the accuracy of CL-strategies was comparable to the stronger BabyBERTa-style baseline introduced by the authors. We add new **cognitively-plausible objective curricula**, as an extension to the original CLIMB submission and CLIMB's improved BABYBERTA-style as baselines.

2.2 Acquisition Models in Deep Learning: Three Models

To assess whether using acquisition theories can be used to formulate better-performing CL strategies, we consider three recent language acquisition models that are amenable to Deep Learning implementation, as they specify developmental sequences that can be replicated as CL strategies in SSLMs. Based on careful linguistic analysis of universal and language-specific patterns in the utterances produced by learners cross-linguistically at different stages of acquisition, linguists have formalised strict (universal or non-language-specific) or weak (language-specific) orders of syntactic categories that are sequentially acquired. Since these acquisition models have been formulated based on linguistic analysis of multilingual acquisition data, we consider whether the CL strategies that precisely replicate these models can inform better-performing curriculum learning strategies cross-lingually. This leads us to train SSLMs with these objective curricula beyond English. As schematised in *Figure* 1, we can precisely replicate these developmental sequences as stages of SSLM pre-training, defined as proportions of training steps.

We implement three contemporary cross-lingual models of syntactic acquisition:

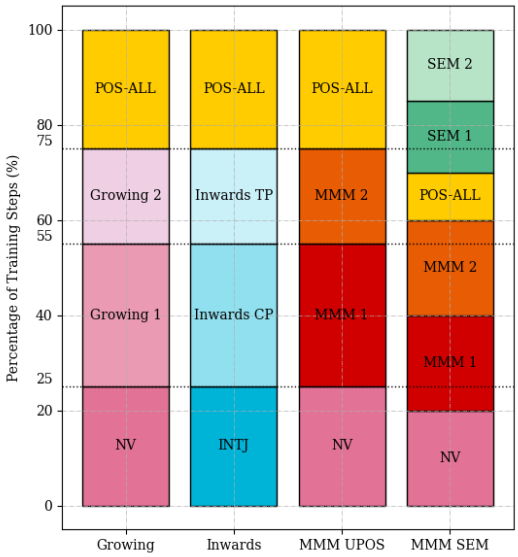

Figure 1: **Acquisition-inspired Objective Curricula:** We specify Objective Curricula GROWING, INWARDS, MMM (UPOS), MMM (SEMANTIC) for three theories of acquisition (*Section* 2.2). The Progression of Curriculum Units replicate the predicted developmental sequences by specifying curriculum units (defined in *Table* 1) defined over different pre-training stages, expressed as a percentage of training steps.

1. **GROWING:** Bottom-up maturational approaches to language acquisition (Rizzi, 1993; Radford, 1990), including the "Growing Trees Hypothesis"(Friedmann et al., 2021), predicts that first language learners begin acquiring verbs and nouns (unit NV in *Table* 1). Learners subsequently progress to acquiring predicate information to form simple sentences; and finally, acquire discourse and complementiser information, allowing them to formulate complex sentences (e.g., with relative clauses). We can assume a tripartite model of bottom-up maturational development for implementation, with units Growing 1 and Growing 2 in *Table* 1.[1]

2. **INWARDS:** Bosch (2023) introduces the predictions of a **generalised inward-growing maturational proposal (INWARDS)**, building on evidence from Heim and Wiltschko (2021) of early acquisition of "discourse"-material and interactional language (e.g. tags-questions). This predicts exactly the opposite order of acquisition of GROWING. The stages of development begin with the early acquisition of complementisers used for illocutionary/discourse-related purposes (INTJ and INWARDS- CP in *Table* 1); followed by the acquisition of tense/event-related information (INWARDS-TP); and finally, thematic information.

3. **NEO-EMERGENT (MMM):** Neo-Emergentism predicts developmental stages in language acquisition that show increasing categorial granularity, taking a language-specific, or non-maturational, approach towards syntactic acquisition (Biberauer and Roberts, 2015). The general universal prediction of one neo-emergent model called Maximise Minimal Means (MMM) is that all learners, irrespective of

[1]There are differences in the number of stages predicted in bottom-up maturational approaches. Bottom-up approaches (Rizzi, 1993; Radford, 1990) predict tripartite developmental sequence (a Verb Phrase, Tense Phrase and Complementiser Phrase), but Growing Trees involves bipartite stages (TP and VP is Stage 1, and Stage 2 involves acquiring the CP until QP to predict early acquisition of WH-questions).

the language being acquired, follow the same "coarse" stages in the acquisition of syntactic categories. They first learn to distinguish nouns and verbs (Unit NV), and then an "intermediate" set of categories (complementisers and event-related words),[2] before finally learning tense/aspectual categories (units MMM 1 and MMM 2 in *Table* 1). We implement this as a **universal "coarse" default curriculum strategy** that we implement as a default curriculum strategy (MMM (UPOS) in *Figure* 1). However, MMM also incorporates **language-specific differences in "finer-grained" curricula** where learners can acquire language-specific categories, leading to typological variation in the order of acquisition (Biberauer, 2019; Bosch, 2023, 2024), which we try to model in a CL strategy by specifying language-specific tagsets in SEM 1, SEM 2 in *Table* 1.

Unit	POS Tags
NV	[NOUN, VERB]
Growing 1	NV+ [DET, ADJ, PRON, PROPN, NUM, PRT]
Growing 2	growing_1 + [AUX, PART, ADP, ADV]
INTJ	[X, INTJ, SYM]
INWARDS CP	INTJ+ [PROPN, CCONJ, SCONJ, SYM]
INWARDS TP	CP+ [NUM, PRT, AUX PART, ADP, ADV]
MMM 1	NV+ [DET, CONJ, INTJ]
MMM 2	MMM 1 + [ADJ, ADV, PRON, PROPN, NUM, PRT]
SEM 1	UPOS $+t_{\text{sem}} \in$ [EVE, TNS, ACT, ANA]
SEM 2	SEM1 + $+t_{\text{SEM}} \in$[LOG, COM, DEM, DIS, MOD, ENT, NAM, TIM]

Table 1: Summary of Curriculum Units comprise Universal Part-of-Speech Tags and the Semantic Tags introduced by Bjerva et al. (2016) used to define GROWING, INWARDS & MMM objective curricula. The ordering of units for each acquisition-inspired curriculum is shown in *Figure* 1.

Each stage of the GROWING, INWARDS and MMM models can be defined as a 'curriculum unit' composed of POS tag sequences listed in *Table* 1.[3] To precisely replicate the developmental sequences of each acquisition model computationally, we will need to use a supervised tagger to specify curricula using strictly ordered sequences of POS tags. This is a cognitively motivated divergence from Diehl Martinez et al. (2023), who use an unsupervised tagger to define curricula. Using a supervised tagger is argued by Buttery (2006) to enable computational modelling of a more cognitively plausible starting point for first language (L1) learners – based on a view of acquisition that is not fully emergent, nor completely nativist.[4] For our purposes, it allows us to precisely replicate developmental sequences in SSLMs using curriculum learning.

3 Dataset

3.1 Training Corpora: MAO-CHILDES

We collect a training corpus of Age-ordered Child-Directed Speech (CDS) for four languages (French, German, Japanese and Chinese), in addition to the English Age-Ordered-CHILDES (AO-CHILDES) corpus (Huebner and Willits, 2021) used in the BabyLM Challenge, to assess the benefits of the acquisition-inspired curricula beyond English compared to non-curriculum SSLMs. MAO-CHILDES is developed from the Child Language Data Exchange System (CHILDES) (MacWhinney, 2000), which consists of in-home recordings of casual speech from caregivers to children and in-lab activities such as play, conversation and book reading directed towards first language learners for several languages.[5] We make our training corpus available on HuggingFace.[6] The distribution of CHILDES data beyond English is a practical challenge for extending the BabyLM Challenge beyond English. *Table* 6 shows the imbalance in quantities of CDS extracted from CHILDES, which is an artefact of a Western, Educated, Industrialised, Rich, and Democratic (WEIRD) bias in language acquisition research (Henrich et al., 2010). A sample of CDS in the age-ordered corpora is shown in *Figure* 2, from different stages of language acquisition. Following Huebner and Willits (2021), utterances

[2] In Chomskyan terminology, a vP-shell and a Complementiser Phrase (CP).

[3] The Chomskyan acquisition models used in this paper technically refer to syntactic projections, rather than part-of-speech tags.

[4] Note that Buttery (2006) uses a model within a Combinatorial Categorial Grammar (CCG)-based formalism, which is also a "middle ground" between fully emergent acquisition models and a traditional biologically hardwired Universal Grammar assumed in traditional Chomskyan models like Principles and Parameters.

[5] Original data can be accessed here: https://childes.talkbank.org/

[6] https://huggingface.co/climb-mao

from children and child-directed speech (CDS) produced by caregivers, and other interlocuters, to children over the age of 6;0 are disregarded, leaving CDS produced by caregivers to children less than 6;0 which is sorted using the meta-data of the age of the learner in the CHILDES database.[7]

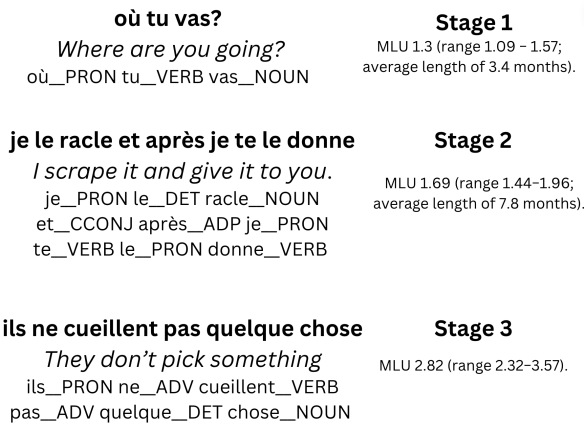

Figure 2: A sample of Child-Directed Speech (CDS) from French **MAO-CHILDES** that learners receive from caregivers at different stages of acquisition. Stages of acquisition are standardly defined in terms of mean lengths of utterances produced by learners.

3.2 Evaluation Datasets

To assess the success of three objective curricula (GROWING, INWARDS and MMM) that precisely replicate the predictions of the acquisition theories in *Section* 2.2 on a standard SSLM architecture in a multilingual setting, we extend the evaluation pipeline of the BabyLM Challenge. This consists of syntactic evaluation datasets like BLiMP (Warstadt et al., 2020) composed of minimal pairs of grammatical and ungrammatical sentences for language-specific syntactic phenomena. We use the following minimal pairs datasets to evaluate the objective curricula for the four languages in MAO-CHILDES:

1. **CLAMS (French and German):** The Cross-Lingual Syntactic Evaluation of Word Prediction Models (CLAMS) (Mueller et al., 2020) generates minimal pair datasets which we use for French and German using Attribute-Varying Grammars. The dataset assesses grammaticality in Simple Agreement, VP coordination, and across "interveners" in S-V agreement (subject/object relative clause or across a Prepositional Phrase).

2. **JBLIMP (Japanese):** JBLIMP (Someya and Oseki, 2023) is a minimal pairs dataset for targeted syntactic evaluation of Japanese. It consists of 331 minimal pairs of syntactic acceptability judgements curated from Japanese syntax articles in the *Journal of East Asian Linguistics*.[8]

3. **SLING (Chinese): SLING** (Song et al., 2022) is a 38K minimal sentence pair dataset derived by applying syntactic and lexical transformations to Chinese Treebank 9.0,[9] aiming to improve on the limitations of an earlier dataset called CLiMP (Xiang et al., 2021), which had a lack of diversity in the vocabulary to generate minimal pair templates.

Due to the small size of the JBLIMP minimal pairs dataset, we follow Someya and Oseki (2023)'s recommendation to compute accuracy using a SLOR score to mitigate the confounding effects of lexical frequencies and sentence lengths, which is defined as follows:

$$SLOR(X) = \frac{\log p_m(X) - \log p_u(X)}{|X|}$$

where $p_m(X)$ is the probability of a sentence for a Language Model and is the unigram probability of the sentence, estimated for each subword in the training corpus. Accuracy calculations for other languages follows dataset guidance to use unnormalised log-probabilities.

3.3 Universal POS Tagging

To define fine-grained objective curricula that perform masked language modelling with different subsets of syntactic and semantic tags for a specified proportion of training steps, we have to annotate child-directed speech corpora with Universal POS tags using an off-the-shelf SpaCy multilingual POS tagger. The distribution of POS tags in MAO-CHILDES (Figure 4) contains a high proportion of Nouns, whereas Verbs contribute a relatively low count. There are orthographic issues in the CHILDES dataset for East Asian Languages,

[7]The Script for Generating AO-CHILDES can be found here:https://github.com/UIUCLearningLanguageLab/AOCHILDES

[8]The JBLiMP Minimal Pair dataset can be found here: https://github.com/osekilab/JBLiMP/tree/main

[9]The SLING Dataset can be found here: https://huggingface.co/datasets/suchirsalhan/SLING

which are transcribed using Romanised characters (romaji) and a large proportion of English loan words in the Japanese portion of MAO-CHILDES, used in certain lexical domains, are incorrectly tagged automatically. These pre-processing inconsistencies were manually corrected. We also train a semantic tagger to specify language-specific curriculum strategies (see *Appendix* A for more detail).

4 Methodology

4.1 Model Architecture

Following Diehl Martinez et al. (2023), we develop non-curriculum learning models. These models are scaled-down language models based on RoBERTa (Liu et al., 2019), with 8M parameters and trained on no more than 30M words (Huebner et al., 2021). We use 8192 vocabulary items, which Diehl Martinez et al. (2023) find yields better overall performance compared to a larger vocabulary. Token unmasking is also removed, like BabyBERTa. We use a small model architecture composed of eight layers. This follows Diehl Martinez et al. (2023), who compare the role of model size ($8, 10, 12$ Transformer layers) and vocabulary size (comparing $|V| \in \{8192, 16384\}$). An AdamW optimiser with linear scheduling is used (Loshchilov et al., 2017). Each model is trained for 400,000 steps with 4 A100 GPUs. The hyperparameters used for the "vanilla" SSLMs are shown in *Table* 4. The models concatenate input sequences to capitalise on the available input length.

4.2 Baselines: LLMs and SSLM (WIKI)

We use two families of models as baselines. First, we compare the performance of monolingual SSLMs to monolingual Large Language Models to assess the benefits of the BabyLM paradigm. For French, German and Chinese, we use RoBERTa-style monolingual LLMs.[10] The Chinese RoBERTa model is trained on around 30B words (Cui et al., 2020), which more than 10^4 times the training data we use to train our SSLMs in the Chinese portion of MAO-CHILDES.[11] We include GPT-2 Baselines for Japanese, which are reported by Someya and Oseki (2023). This is because Japanese RoBERTa monolingual language models[12] are not trained on data using Romaji orthography, which is used in the Japanese portion of MAO-CHILDES (*Section* 3). Secondly, to assess the benefits of pre-training SSLMs on Child-Directed Speech, we train SSLMs using Wikipedia text (SSLM WIKI), which is extracted to match the quantity of training data in MAO-CHILDES for each language. We keep the original hyperparameter settings used by Huebner et al. (2021).

4.3 "Vanilla" SSLMs: MAO-BabyBERTa

We train a family of SSLMs, called Monolingual Age-Ordered BabyBERTa (MAO-BABYBERTA), on language-specific training data from MAO-CHILDES using the model architecture described in *Section* 4.1 without any curriculum learning strategies. Hyperparameters are tuned for English, and we use the same settings in MAO-BabyBERTa.

4.4 Implementing Acquisition-Inspired Objective Curricula: GROWING, INWARDS & MMM

To implement the acquisition-inspired strategies, we filter our age-ordered MAO-CHILDES corpus for each language for expected utility in the acquisition process, according to the curriculum strategies of GROWING, INWARDS and MMM schematised in *Figure* 1. We then precisely implement the GROWING, INWARDS, MMM theories introduced in *Section* 2.2, using different curriculum units composed of POS tagsets (*Table* 1) to define three objective curricula that replicate the developmental sequences of each acquisition model through the progressive ordering of POS units. The logic for performing masked language modelling selectively for words annotated with a desired set of specified part of speech tags is implemented in Diehl Martinez et al. (2023), which we extend. The objective curricula modify the masked language modelling (MLM) objective in a multi-task learning setup, so the acquisition-inspired objective is activated and optimised in parallel with MLM. We fix the model architecture to be identical to the "vanilla" SSLM architecture in *Section* 4.3 to evaluate the benefits of each curriculum strategy. We modify CLIMB's objective curricula to implement the GROWING, INWARDS and MMM objective curricula by splitting 400K training steps across

[10] The French RoBERTa model is available here: https://huggingface.co/abhilash1910/french-roberta. The German RoBERTa model is available here: https://huggingface.co/uklfr/gottbert-base

[11] The Chinese RoBERTa model is available here: https://huggingface.co/hfl/chinese-roberta-wwm-ext-large.

[12] Japanese RoBERTa models is available here: https://huggingface.co/rinna/japanese-roberta-base

	Model		English	Japanese	Chinese	French	German
Non-CL	SSLM (WIKI)		64.60%	55.42%	48.01%	70.68%	59.63%
	MAO-BABYBERTA		75.48% *	61.21%	51.32%	**80.00%**	68.78%
CL	GROWING		71.13%	79.30%	56.22%	76.21%	71.13%
	INWARDS		71.05%	81.32%	54.26%	79.01%	69.34%
	MMM	(UPOS)	74.22%	**87.31%**	**58.79%**,	75.93%	**73.25%**
		(SEM)	**77.35%**		55.01%		

Table 2: Evaluation of MAO-BABYBERTA ("vanilla" SSLM architecture without objective curricula) and the three Objective Curricula (GROWING, INWARDS, and MMM) on the following syntactic minimal pairs datasets: BLIMP (English), JBLIMP (Japanese), SLING (Chinese), CLAMS (French and German). Performance is compared to SSLM (WIKI). This is the same architecture trained on non-CDS training data. *This reports the performance of the best-performing "vanilla" model by Diehl Martinez et al. (2023) on the same architecture used to train our model. **Bolded** results indicate the highest accuracy of all the models.

four non-uniform intervals that are defined as a proportion of the SSLM's training steps, defined in *Figure* 1. This is meant to roughly simulate four developmental stages of an idealised monolingual learner until 6;0. We then specify tagsets for each phase of the curricula that correspond to the acquisition theory. To illustrate this, the INWARDS curriculum begins with a unit INTJ, which performs MLM for interjections and other interactional language, which are annotated with tags INTJ, X, SYM. Then, we specify two further curriculum units INWARDS-CP which performs MLM on complementiser-like words (e.g., SCONJ), and INWARDS-TP which performs MLM on auxiliaries AUX and other tense/event-related words. At each stage of the curriculum, the objective curricula provide the vanilla SSLM model with a list of syntactic tags to use during training, taken from a pre-specified set of UPOS tags that lists all the tags used in the UPOS tagged MAO-CHILDES training set. If a tag is not used at the curriculum stage, its "ID" is set to zero so it is not a target for masked language modelling (MLM). During training, the number of part-of-speech tags that the model has to classify over are varied, according to the predictions of each acquisition model. The objective curricula end with a final curriculum unit, Pos-ALL, containing the entire Universal Part-of-Speech Tagset. The masking ratio is an important hyperparameter that impacts the pretraining of a Masked Language Model. A masking ratio of 0.4 is used for the tags specified at the curriculum stage. A 0.15 masking rate is used elsewhere if the tag is not specified at the curriculum stage. For RoBERTa-based Language Models, a masking ratio of 0.4 performs better than 0.15 in downstream tasks (Wettig et al., 2023).In addition to our "default" MMM strategy defined by Universal POS tags, MMM (UPOS), we additionally introduce a **refined version of the MMM objective**, MMM (SEM) for English and Chinese. This adds two additional stages to the non-language specific strategy to define a language-specific curricula that utilises semantic tags (Bjerva et al., 2016), or *sem*-tags, to model **language-specific acquisition strategies** (Section 2.2). Detailed methods and results are discussed in *Appendix* A. Training times for each objective are summarised in *Table* 5.

5 Results

The performance of objective curricula and cross-lingual SSLMs on minimal pairs datasets is summarised in *Table* 2. **Fine-grained objective curricula demonstrate variable effectiveness compared to non-curriculum baselines**. While MMM (UPOS) shows general promise, average benefits of MMM (UPOS), GROWING, and INWARDS, do not show statistically significant improvements on MAO-BABYBERTA cross-linguistically ($p < 0.05$). However, **the MMM (SEM) curriculum achieves a statistically significant performance improvement in both English and Chinese** ($p < 0.05$) when performing a paired t-test. Instead, **statistically significant improvements are observed with acquisition-inspired CL strategies in specific languages across minimal pairs test sets**. MMM (UPOS) only achieves a statistically significant improvement in Japanese and Chinese. GROWING leads to a statistically significant improvement in Japanese and Chinese, while INWARDS only has statistically significant improvements in Japanese. No curriculum strategy outperforms MAO-BABYBERTA in French, although INWARDS almost reaches the same accu-

racy. German CL strategies only marginally outperform the non-CL baseline. In *Figure* 3, we compare these results with a broader range of models introduced by Diehl Martinez et al. (2023), finding that the English MMM (SEM) curriculum marginally outperforms other curriculum learning strategies. See *Appendix* C for details on how t-test statistics are computed.

Language	LLM	SSLM (CL)
English	**80.10**	77.35(MMM SEM)
Japanese	77.95	**87.31 (MMM)**
Chinese	**83.41**	58.79 (MMM)
French	**83.00**	79.01(Inwards)
German	**92.16**	73.25(MMM)

Table 3: Comparison of Accuracy of LLMs and the Best Performing CL Strategy on Minimal Pairs Datasets. SEM represents Language-Specific strategies implemented for English and Chinese pre-training compared to the language-invariant MMM (UPOS) strategies.

6 Discussion

Acquisition-inspired CL strategies represent a novel large-scale application of language acquisition theory in Deep Learning, aimed at improving the performance of SSLMs. Acquisition-inspired curricula guide SSLMs, which function as large statistical learners, to generalise over frequent linguistic categories—such as nouns and verbs—early in the training process and attend to language-specific features, such as the Germanic V2 word order. This suggests that **more fine-grained, language-specific curricula may have performance benefits over non-CL strategies in SSLMs**, which is supported by results showing the limited improvements of universal/maturational theories of acquisition that inform the GROWING and INWARDS strategies. Although both acquisition models predict universal curricula that should lead to consistent benefits cross-lingually, GROWING/INWARDS only improve performance in Chinese and Japanese, while performing comparably to non-curriculum (non-CL) baselines in French/German and worse than non-CL baselines in English. An additional benefit of using fine-grained language-specific curricula is that it enables SSLMs to learn more complex grammatical phenomena that may rely on semantics like anaphora. We notice notable improvements in ellipsis performance (*Table* 7) with the MMM (SEM) curriculum. Interestingly, in Chinese, the MMM (SEM) curriculum marginally underperforms compared to MMM (UPOS) when handling anaphora and aspectual phenomena (*Table* 8), highlighting the need for further investigation into engineering optimal language-specific curriculum strategies that outperform non-CL strategies. This raises important avenues for future research. Careful analysis of developmental sequences beyond English to develop language-specific strategies similar to MMM (UPOS/SEM) will be crucial. We encourage practitioners to curate larger corpora of child-directed speech (CDS) for training SSLMs in languages beyond English and to develop more minimal pair datasets that have coverage beyond grammatical agreement in CLAMs to develop better-performing curriculum strategies for Romance and Germanic. Additionally, an important finding is that **acquisition-inspired CL strategies in Japanese significantly outperform GPT-2** (*Table* 3). The improvements observed in Japanese control/raising phenomena (*Table* 9) suggest that the properties of CDS in Japanese may lead to more robust generalisations than LLMs.

7 Conclusion

This paper assesses whether fine-grained curriculum learning strategies based on acquisition theories can provide better heuristics for CL strategies for SSLM pre-training cross-lingually, introducing the MAO-CHILDES training corpus to train SSLMs for four typologically distant language families. Mixed results of the maturational GROWING and INWARDS acquisition theories in curriculum strategies and the implementation of the coarse/universal prediction of MMM (UPOS) suggest that there is no guaranteed performance benefit just by devising universal CL strategies based on acquisition theories for SSLMs in a multilingual setting. Training SSLMs using more fine-grained language-specific curricula that precisely replicate cutting-edge linguistic theories is effective for the MMM (SEM) objective in English and Chinese and MMM (UPOS) in Japanese. Curriculum Learning can outperform non-curriculum SSLMs by specifying fine-grained language-specific curricula that precisely replicate language acquisition theories, highlighting how cognitively-inspired techniques can lead to better-performing data-efficient architectures in the spirit of the BabyLM Challenge.

Acknowledgments

Many thanks to Andrew Caines for his comments, supervision and feedback on this paper. We thank Núria Bosch-Masip for her comments on the linguistic acquisition models implemented in this paper. We thank Mila Marcheva for her thoughts on cognitively-inspired modelling, which influenced the ideas in this paper. This paper reports on work supported by Cambridge University Press & Assessment. It was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council. Additionally, we thank the NVIDIA Corporation for the donation of the Titan X Pascal GPU used in this research. Richard Diehl Martinez is supported by the Gates Cambridge Trust (grant OPP1144 from the Bill & Melinda Gates Foundation). Zébulon Goriely's work is supported by The Cambridge Trust.

References

Theresa Biberauer. 2019. Children always go beyond the input: The Maximise Minimal Means perspective. *Theoretical Linguistics*, 45(3-4):211–224.

Theresa Biberauer and Ian Roberts. 2015. Rethinking Formal Hierarchies: A Proposed Unification. *Cambridge Occasional Papers in Linguistics*, 7:1–31.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic tagging with deep residual networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541, Osaka, Japan. The COLING 2016 Organizing Committee.

Nasim Borazjanizadeh. 2023. Optimizing GPT-2 pretraining on BabyLM corpus with difficulty-based sentence reordering. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 356–365, Singapore. Association for Computational Linguistics.

Núria Bosch. 2024. On another topic, how do acquisition orders vary? The left periphery and topicalisation in bilinguals and monolinguals. 1st year PhD report.

Núria Bosch. 2023. Emergent Syntax and Maturation: A Neo-Emergentist Approach to Development. *MPhil Thesis, Department of Theoretical and Applied Linguistics, University of Cambridge*.

Paula J. Buttery. 2006. Computational models for first language acquisition. Technical Report UCAM-CL-TR-675, University of Cambridge, Computer Laboratory.

Aryaman Chobey, Oliver Smith, Anzi Wang, and Grusha Prasad. 2023. Can training neural language models on a curriculum with developmentally plausible data improve alignment with human reading behavior? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 98–111, Singapore. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Justin DeBenedetto. 2023. Byte-ranked curriculum learning for BabyLM strict-small shared task 2023. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 198–206, Singapore. Association for Computational Linguistics.

Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB – Curriculum Learning for Infant-inspired Model Building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127, Singapore. Association for Computational Linguistics.

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

Naama Friedmann, Adriana Belletti, and Luigi Rizzi. 2021. Growing trees: The acquisition of the left periphery. *Glossa: a journal of general linguistics*, 6(1):131.

Jutta Heim and Martina Wiltschko. 2021. Acquiring the form and function of interaction: a comparison of the acquisition of sentence-final particles and tag questions in the brown corpus. Talk presented at LAGB Annual Meeting 2021 (online), 8 September.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

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

Wenxi Li, Yiyang Hou, Yajie Ye, Li Liang, and Weiwei Sun. 2021. Universal semantic tagging for English and Mandarin Chinese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5554–5566, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

Brian MacWhinney. 2000. *The CHILDES Project: The Database*, volume 2. Psychology Press.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330.

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.

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 290–297, Singapore. Association for Computational Linguistics.

Mattia Opper, J. Morrison, and N. Siddharth. 2023. On the effect of curriculum learning with developmental data for grammar acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 346–355, Singapore. Association for Computational Linguistics.

Andrew Radford. 1990. The Syntax of Nominal Arguments in Early Child English. *Language Acquisition*, 1(3):195–223.

Luigi Rizzi. 1993. Some Notes on Linguistic Theory and Language Development: The case of root infinitives. *Language Acquisition*, 3(4):371–393.

Suchir A. Salhan. 2023. On the potential for 'Maximising Minimal Means' in Transformer Language Models: A Dynamical Systems Theory Perspective. *Cambridge Occasional Papers in Linguistics*, page 55–110.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. Call for papers – The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus.

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.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

A MMM (SEM): Specifying Language-Specific Curricula using Semantic Tags

As a first step towards modelling language-specific curricula using curriculum learning, we use Universal Semantic Tagging (*sem-tagging*) (Bjerva et al., 2016). The set of semantic tags can differ cross-lingually. In Chinese, Li et al. (2021) specifies a language-specific semantic tagset, adding and removing tags based on Chinese's semantic and syntactic properties. The fine-grained curriculum in an SSLM set-up aims to circumvent known problems of shortcut learning in LLMs that prevent Transformer-based models from exhibiting robust structural generalisation capabilities that humans exhibit in acquisition (Salhan, 2023).

We perform *sem*-tagging to annotate the BabyLM corpus for English and the Chinese corpus in MAO-CHILDES with a set of language-neutral tags (sem-tags). For English, we only perform *sem*-tagging for the Adult Directed Speech datasets in the BabyLM Challenge dataset: the BNC, Project Gutenberg , OpenSubtitles, QCRI, Wikipedia and Switchboard corpora. This allows us to modify our UPOS curricula for English to specify a more complex curricula to simulate later stages of language acquisition. The first stage of the new MMM curriculum using semantic tags includes tags related to event, `EVE`, tense, `TNS`, and modality `MOD`. These are typically learnt later during acquisition, as part of complex tense sequences of auxiliaries and modal verbs (Biberauer and Roberts, 2015), and allow us to define a **language-specific** sem-tag objective. For Chinese, we sem-tag a corpus of Wikipedia text that contains the same amount of text as the age-ordered CHILDES corpora introduced in *Section 3*.

A.1 Semantic Tagger Accuracy

A multi-objective POS and *sem*-tagger is trained, using a Bidirectional LSTM (BiLSTM) with a Conditional Random Field (CRF) inference layer to train a multi-objective semantic and UPOS tagger for English and Chinese. This is trained on 1100 *sem*-tagged sentences from the Wall Street Journal (WSJ) section of the Penn Treebank (Marcus et al., 1993) and a 1000 *sem*-tagged sentences from Chinese TreeBank (Xue et al., 2005) annotated by Li et al. (2021). The tagger has 91.4% accuracy for Chinese and 94.6% accuracy for English.

B Training

Table 4: Hyperparameter Settings for CLIMB's "vanilla" and curriculum models and MAO-BabyBERTa (CDS)

Layers	8
Heads	8
Hidden	256
$\|V\|$	$8,192$
Layer Norm EPS	1×10^{-5}
Learning Rate	0.001
Optimizer	AdamW
Scheduler Type	Linear
Max Steps	$400,000$
Warm-up Steps	$100,000$

Type	Model	Training Time
MAO-CLIMB	GROWING	11h 51m
	INWARDS	11h 51m
	MMM (UPOS)	11h 46m
	MMM (SEM)	25h 3m
Vanilla Models	CLIMB-small-raw	12h

Table 5: Compute required to train our models. We report the model with the shortest and longest runtime for each experiment type. Each model is trained for 400,000 steps with 4 A100 GPUs.

C Statistical Significance & Detailed Results

The statistical significance of the three curriculum strategies, GROWING, INWARDS & MMM is calculated by performing t-tests on the detailed results in *Tables* 7, 8, 9, 10. For each curriculum (GROWING, INWARDS, MMM (UPOS), MMM (SEM)), we calculate the paired differences in accuracy with the Vanilla model for all the test sets in the minimal pairs evaluation dataset. We perform paired t-tests for the non-CL baseline (MAO-BABYBERTA) and the accuracy of the respective curriculum for each curriculum strategy for each language, concluding that the curriculum-based model significantly outperforms the Vanilla/MaoBabyBERTa model if the p-value is below our significance level $\alpha = 0.05$. The detailed results, below, support the findings of Huebner et al. (2021) cross-linguistically of the benefits of using less training data and paying careful attention to training artefacts and the domain of training corpora, as using CDS to train SSLMs (with/without objective curricula) outperforms SSLM (WIKI).

Figure 3: **Comparision of BLiMP Performance of English SSLMs with CLIMB curricula and GROWING, INWARDS, MMM (UPOS), MMM (SEM) (*Section* 4.4)** We report introduced by Warstadt et al. (2023) for T5-base and OPT-125m models. We include the improved BabyBERTa baseline implemented in Diehl Martinez et al. (2023), which beat the baseline used in the 1st BabyLM Shared Task. We report BLiMP performance of different CLIMB `small-raw` models (also used in the standard architecture of MAO-BABYBERTA used with the three objective curricula) for the best performing dynamic curriculum learning strategies implemented in Diehl Martinez et al. (2023). This includes CLIMB's **Data Curriculum** (Log Pacing with Source Difficulty), **Vocabulary Curriculum** (Log Pacing with Token ID Difficulty), two **Objective Curricula strategies** (MLM + ALL uses a multitask objective of masked language modelling and objective curricula specified by 10 tags throughout all training steps, MLM + NV uses three tags throughout training), and the best performing **Combination Model** (Token ID Vocabulary Curricula, Random + model ppx Data Curricula, Multitask Objective Curricula).

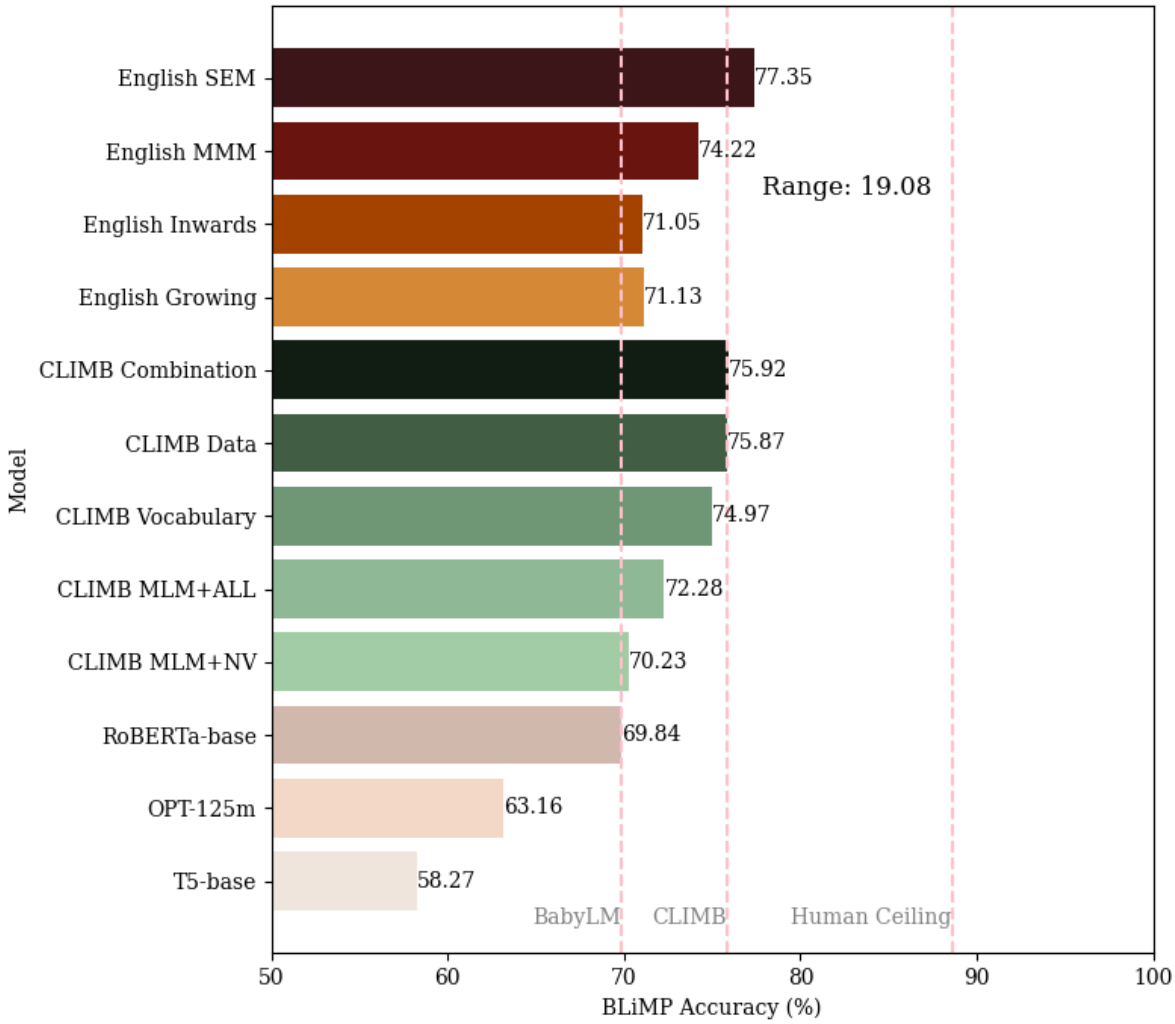

Table 6: Corpus Statistics for the **Child-Directed Speech (CDS)** files extracted from CHILDES for 24 languages, which are used to select four languages for training. The MAO-CHILDES corpus is selected based on the frequency of CDS, along additional considerations of evaluation.

lang	Samples	$\|V\|$	Tokens	Sentence Length μ	Children	Utterances
Chinese	857,792	518,172	850,510	258.28	949	3,293
German	582192	516,147	867,704	107.05	65	8105
Japanese	537,164	280,807	528,930	38.67	122	13,678
Indonesian	537,235	286,448	521,759	202.31	9	2,579
French	488,094	284,381	469,258	175.69	204	2,671
Spanish	332,903	211,559	331,009	167.85	291	1,972
Dutch	261,786	160,520	259,263	97.50	96	2,659
Portuguese	100,512	59,205	98,620	39.72	195	2,483
Polish	82,977	71,072	82,940	43.04	14	1,927
Swedish	80,936	53,719	79,739	49.34	6	1,616
Norwegian	55,262	31,310	40,215	32.62	6	1,233
Catalan	54,518	37,250	53,157	29.73	7	1,788
Romanian	33,130	20,700	32,986	16.58	6	1,990
Croatian	51,948	36,922	51,809	27.33	3	1,896
Czech	45,122	33,185	44,117	27.15	6	1,625
Danish	44,909	25,039	44,909	24.94	2	1,801
Bulgarian	31,715	21,435	31,714	32.76	1	968
Afrikaans	22,021	18,475	21,984	18.68	52	1,177
Irish	18,973	13,598	18,869	9.82	5	1,921
Russian	7,008	5,963	7,007	4.42	2	1,585
Icelandic	47,945	27,775	46,516	11.36	1	4,094
Slovenian	1,384	1,243	1,382	10.39	1	133
Thai	38,550	27,084	38,329	100.34	18	382

Figure 4: Distribution of Silver Tags across all languages in the MAO-CHILDES corpus, annotated using a SpaCy Multilingual UPOS Tagger

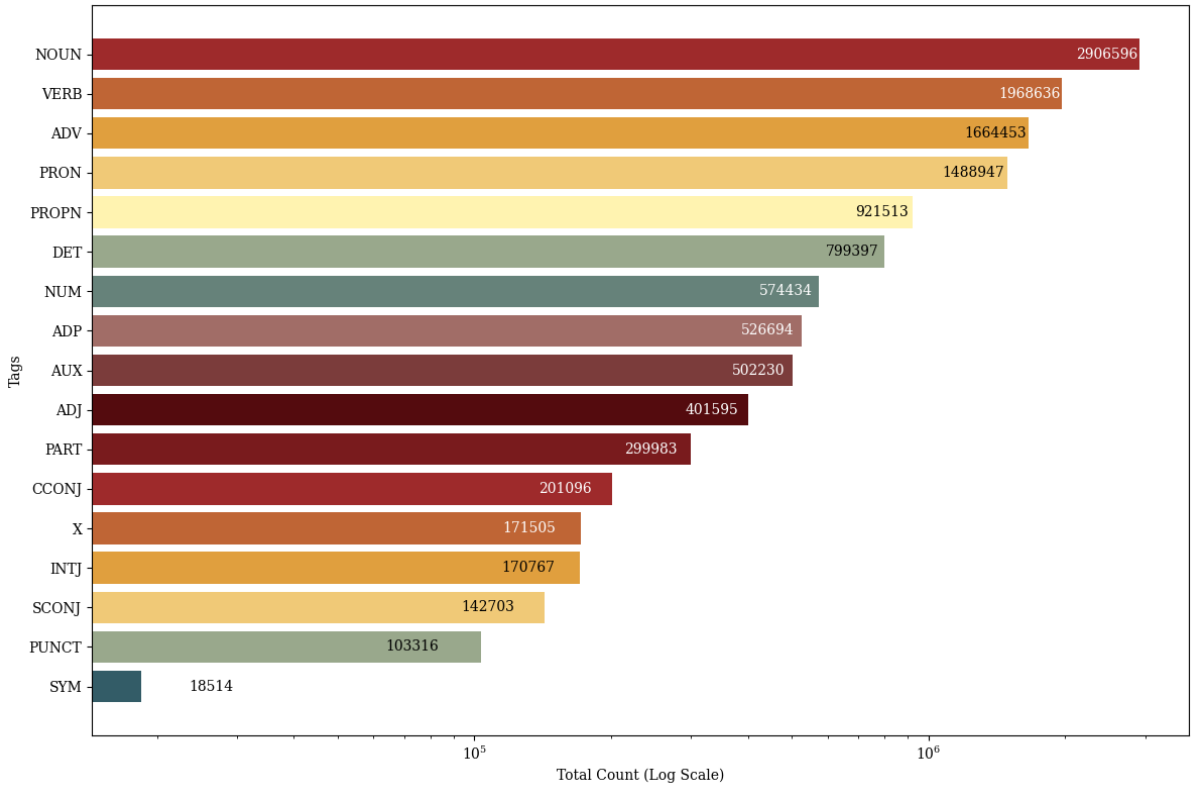

Table 7: **(English)** Evaluation of BabyBERTa model with four Cognitively-Plausible Curriculum Learning Strategies on BLIMP. English GROWING based on "Growing Trees" (Friedmann et al., 2021), INWARDS based on "Inward Maturation" (Heim and Wiltschko, 2021) and MMM (UPOS) and the language-specific *sem*-tag MMM (SEM) curricula based on Biberauer and Roberts (2015).

Grammatical Phenomenon	**Growing**	**Inwards**	**MMM** (UPOS)	**MMM** (SEM)
Anaphor	**96.22%**	84.67%	81.13%	**90.89%**
Arg Str	79.13%	79.86%	84.79%	**85.99%**
Binding	46.47%	71.75%	**83.42%**	77.76%
Control-Raising	77.03%	73.82%	**88.02%**	82.10%
Det-N Agreement	65.49%	65.19%	**84.38%**	79.31%
Ellipsis	58.24%	53.26%	42.77%	**70.94%**
Filler Gap	80.70%	**88.47%**	85.60%	73.11%
Irregular	**76.34%**	44.85%	54.42%	74.91%
Island	69.53%	62.87%	**96.62%**	68.64%
NPI	69.21%	76.02%	**83.42%**	74.13%
Quantifiers	44.54%	**84.79%**	58.43%	71.86%
Subject-Verb	65.98%	64.89%	**68.37%**	**79.03%**
Average Accuracy	71.13%	71.05%	74.22%	**77.35%**

Table 8: **(Chinese)** Comparison of accuracy of Chinese MAO-BABYBERTA ("vanilla") and GROWING, INWARDS, MMM (UPOS), MMM (SEM) objective curricula compared to a Chinese RoBERTa LLM baseline on the SLING minimal pairs dataset (Song et al., 2022)

Category	**Subcategory**	**Vanilla**	**LLM**	**Growing**	**Inwards**	**MMM (UPOS)**	**MMM (SEM)**
RelativeClause	rc_resumptive_pronoun	50.50	60.30	50.50	49.50	53.10	50.70
RelativeClause	rc_resumptive_noun	48.00	27.60	48.90	47.80	58.00	48.50
Anaphor	baseline_female	86.70	75.60	86.30	83.90	36.90	85.60
Anaphor	pp_female	70.50	71.80	70.80	67.50	41.80	69.80
Anaphor	baseline_male	12.50	38.50	45.20	45.30	81.90	45.20
Anaphor	Plural	51.98	97.95	53.10	51.20	52.33	52.10
Anaphor	self_male	14.30	92.60	47.80	46.10	81.40	46.90
Anaphor	pp_male	28.00	77.60	49.50	48.70	76.90	49.30
Anaphor	self_female	86.60	98.50	86.70	84.10	42.00	85.10
PolarityItem	any	54.20	85.60	55.30	52.70	49.20	54.60
PolarityItem	more_or_less	20.20	98.90	46.80	46.50	46.70	46.80
PolarityItem	even_wh	56.90	92.40	57.90	53.60	57.30	55.90
DefinitenessEffect	definiteness_every	85.70	94.60	85.40	83.30	88.50	84.20
DefinitenessEffect	definiteness_demonstrative	78.80	96.20	78.60	75.20	55.00	77.30
Aspect	zai_guo	49.30	97.30	49.70	47.90	43.10	49.20
Aspect	temporal_le	40.70	63.40	50.40	49.10	63.70	50.30
Aspect	zai_le	49.80	74.40	49.90	48.20	69.00	48.90
Aspect	temporal_guo	40.30	88.10	50.30	47.60	60.20	50.10
Aspect	zai_no_le	56.40	77.90	56.70	53.80	86.80	55.20
WhFronting	mod_wh	54.70	99.70	54.40	51.90	36.10	53.10
WhFronting	bare_wh	53.30	100.00	53.50	50.30	46.00	52.40
Classifier-Noun	cl_simple_noun	51.30	98.00	51.80	49.70	57.40	50.70
Classifier-Noun	cl_adj_simple_noun	52.60	96.30	52.10	50.10	61.80	51.30
Classifier-Noun	dem_cl_swap	51.10	99.60	51.20	49.20	60.70	50.60
Classifier-Noun	cl_adj_comp_noun	48.20	70.60	48.70	46.90	66.00	47.50
Classifier-Noun	cl_comp_noun_v2	49.60	88.80	49.30	47.30	61.90	48.80
Classifier-Noun	cl_comp_noun	51.00	72.00	51.60	49.80	61.30	50.90
Classifier-Noun	cl_adj_comp_noun_v2	52.20	89.50	52.50	50.70	60.90	52.10
AlternativeQuestion	haishi_ma	43.00	95.00	45.70	45.90	49.10	45.70
Average		51.32	83.41	56.23	54.27	58.79	55.48

Table 9: **(Japanese)** Accuracy of the "vanilla" SSLM for Japanese (MAO-BabyBERTa) trained on CDS and the best performing objective curricula +MMM on each phenomenon in the Japanese Benchmark of Linguistic Minimal Pairs (Someya and Oseki, 2023) compared to a Japanese monolingual GPT-2 LLM baseline trained on $\approx 30B$ words and a SSLM (WIKI) Baseline.

Phenomena	GPT2	WIKI	Vanilla	MMM
Control/Raising	16.67	50.00	25.00	70.00
Island Effects	75.76	64.00	72.06	92.19
Binding	58.97	79.05	57.86	89.62
NPI Licensing	50.00	83.33	75.00	90.00
Argument Structure	89.05	41.6	54.82	94.86
Ellipsis	85.96	49.36	56.13	97.68
Verbal Agreement	53.55	57.82	69.22	87.37
Filler-Gap	55.56	44.29	76.19	85.71
Morphology	82.86	49.77	55.08	82.05
Nominal Structure	95.65	41.51	55.87	92.12
Quantifiers	73.81	48.96	60.56	78.52
Average	**77.95**	**55.42**	**61.21**	**87.31**

Table 10: (**French and German CLAMS)** Performance of GROWING, INWARDS, MMM (UPOS) in French and MMM (UPOS) in German (the best performing objective curricula) on CLAMS (Mueller et al., 2020) compared to MAO-BABYBERTA SSLM ("vanilla") and the LLM and SSLM (WIKI) baselines. We report the LLM baselines obtained by Mueller et al. (2020) for mBERT in French and German, which does not report results for "within objective relative" (object rel within) as all focus verbs for that particular language and construction were out-of-vocabulary. Chance CLAMS accuracy is 0.5.

Language	Model	Average	S-V	Obj Rel (within)	Obj Rel (across)	VP Coord	Prep Animate	Subject Relative	Long VP Coord
FRENCH	LLM	83.00%	100.00	–	86.00	100.00	57.00	57.00	98.00
	WIKI	70.68%	67.48	73.40	73.80	71.27	66.80	70.80	71.27
	Vanilla	80.00%	82.0	64.90	84.8	78.6	84.8	83.1	82.1
	Growing	76.21%	73.70	69.57	79.51	71.12	86.53	80.01	73.70
	Inwards	79.01%	76.95	68.50	84.10	75.86	83.80	87.00	76.89
	MMM	75.93%	82.33	72.60	74.40	81.79	65.80	70.90	83.71
GERMAN	LLM	92.16%	95.00	–	93.00	97.00	95.00	73.00	100.00
	WIKI	59.63%	56.55	47.90	60.60	55.32	57.20	60.60	79.28
	MMM	73.25%	75.32	79.80	66.40	78.52	68.40	66.40	77.90