COPiL CAMBRIDGE OCCASIONAL PAPERS *in* LINGUISTICS
Volume 17, Article 10: pp. 173–209, 2025 | ISSN: 2050-5949

Linguistics in the Age of Language Models: What Can Cognitively-Inspired Language Models Offer to Linguistic Theory?*

Suchir Salhan
University of Cambridge

Abstract While theoretical linguists and cognitive scientists alike have contested the contribution of Large Language Models (LLMs) to linguistic theory, small cognitively-inspired Language Models (BabyLMs) have emerged as a complementary research programme that introduces cognitively-inspired language modelling techniques that constrain the nature and volume of training dataset sizes to 'naturalistic quantities'. In this paper, I outline the potential – and potential pitfalls – of BabyLMs in linguistic theory. The paradigm can be used to simulate predictions of acquisition theories and simulate emergent phonological properties cross-linguistically. Small cognitively-inspired language models incentivise research on simulating and testing hypotheses from language acquisition across various environments for 'grammar construction' and analysing the potential and the limits of emergentist hypotheses across morphology, phonology and syntax.

1 Introduction and Theoretical Background

Modern Linguistics pursues a 'multi-model' approach to the study of Language, and despite continual theoretical advances of the Chomskyan paradigm for over half a century, 'virtually every aspect of (I-)language remains a problem' (Chomsky, Gallego & Ott 2019: 253). Some cognitive scientists, notably Piantadosi (2023), have made provocative claims that Large Language Models (LLMs) serve as 'good' theories of human cognition. In response, many linguists and cognitive scientists have refuted in this claim: LLMs do not obviously change the epistemic status of evidence in Linguistic Theory (e.g. Cuskley, Woods & Flaherty 2024, Baker 2024, Fox & Katzir 2024, Katzir 2023). One common criticism brought up by linguists and cognitive scientists about drawing inferences from LLMs is that human learners can robustly acquire their first language (L1) upon exposure to linguistic input of far fewer orders of magnitude than is currently required to train Large Language Models (LLMs). To address this problem, there has been a shift in recent

* Thanks to Zébulon Goriely and Dr Andrew Caines for their comments on this manuscript. Thanks to Prof Paula Buttery, Dr Andrew Caines, Zébulon Goriely and Richard Diehl Martinez for supervising and facilitating the work and experiments reported in this paper. Thanks also Theresa Biberauer, Núria Bosch Masip and Mila Marcheva for linguistic and modelling insights. And thanks to Steph Cooper and the rest of the COPiL Team for their help throughout the editing process.

Natural Language Processing (NLP) research to develop a **cognitively-inspired**, **compute-efficient** small-scale pre-training framework for Language Models (Huebner, Sulem, Cynthia & Roth 2021). Small Language Models trained on naturalistic quantities of textual corpora (BabyLMs) have been touted as valuable 'experimental playgrounds' that improve the sample efficiency of language models, which are increasingly data-intensive. However, their position in linguistic theory is currently undefined beyond their potential applications in NLP to support the development of equitable NLP systems cross-linguistically. If BabyLMs are feasible models of cognition or certain linguistic capabilities, then we must consider what their position is in linguistic theory.

I offer a detailed case study that utilises BabyLMs to compare competing Chomskyan theories about the first language acquisition of syntactic categories in Section 4, and applications of BabyLMs for multilingual NLP and potential convergence with emergentist theories. In Section 5, I set out 'top-down' goals for BabyLMs in linguistic theory that delineate the potential of BabyLMs for evaluating and comparing linguistic theories, in a manner that is methodologically consistent with the working assumptions of *evidentially diverse* theoretical approaches well-founded in the biolinguistic and neo-emergentist approaches in theoretical linguistics that jointly admit 'external' evidence within 'rationalist' theories of cognition.

2 BabyLMs and Cognitively-Inspired Language Modelling

The BabyLM workshop series restricts training Language Models on data that is limited by both scale, 10–100 million words, and by domain, with the pre-training corpus including data from CHILDES, among other child-centered corpora (Warstadt, Mueller, Choshen, Wilcox, Zhuang, Ciro, Mosquera, Paranjabe, Williams, Linzen & Cotterell 2023, Hu, Mueller, Ross, Williams, Linzen, Zhuang, Cotterell, Choshen, Warstadt & Wilcox 2024). Computational linguists have trained language models on child-directed speech (CDS) and simplified corpora in order to study acquisition (Yedetore, Linzen, Frank & McCoy 2023), design cognitively-inspired pretraining strategies (Diehl Martinez, McGovern, Goriely, Davis, Caines, Buttery & Beinborn 2023, Salhan, Diehl Martinez, Goriely & Buttery 2024) and develop chat-based LLMs for children (Nayeem & Rafiei 2024).

While the BabyLM Challenge does not provide any explicit restriction on the neural architecture used, most submissions rely on the Transformer architecture Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin (2017). Informally, Transformers consist of stacked blocks of self-attention and feedforward multilayer perceptions (MLPs). When predicting the next token in a sequence in the standard training objective in 'autoregressive' language models, 'vanilla' transformers attend to all previous tokens, which results in quadratic scaling of compute with sequence length. There are several subtle architectural differences between LLM systems which underdetermine their performance. Performance improvements between traditional Masked Language Models (otherwise known as *encoder-only* models), like BERT (Devlin, Chang, Lee & Toutanova 2019), and newer architectures can be achieved through using combinations of techniques like layer

normalisation, activation functions and new positional embeddings which are used during training to track token position (Warner, Chaffin, Clavié, Weller, Hallström, Taghadouini, Gallagher, Biswas, Ladhak, Aarsen et al. 2024). We will discuss this further in Subsection 5.4.

However, the BabyLM Challenge has also led to the development of new architectures (Georges Gabriel Charpentier & Samuel 2023, Charpentier & Samuel 2024) and has motivated cognitively-inspired pre-training strategies (Huebner et al. 2021, Diehl Martinez et al. 2023, Salhan et al. 2024). These have been found to perform competitively against LLMs in English (Huebner et al. 2021), and are evaluated with a standardised evaluation pipeline that targeted zero-shot morphosyntactic performance, alongside performance when finetuned on language understanding tasks. The BabyLM Evaluation Pipeline consists of benchmarks that each attempt to test for a specific dimension of linguistic competence for English-based SSLMs, such as:

i. **Benchmark of Linguistic Minimal Pairs for English (BLiMP) (Warstadt, Parrish, Liu, Mohananey, Peng, Wang & Bowman 2020)**: This is a metric for formal linguistic competence, comparing the predictions at a **critical word** in a grammatically acceptable and unacceptable minimal pair. The sentences only differ with respect to a single feature, and success is determined if $P(w_{\text{c, acceptable}}) > P(w_{\text{c, unacceptable}})$ for a critical word w_c.

ii. **SuperGLUE (Wang, Pruksachatkun, Nangia, Singh, Michael, Hill, Levy & Bowman 2019)**: A proxy for the 'functional competence' of a language model (Steuer, Mosbach & Klakow 2023), SuperGLUE evaluates for a wide range of natural language understanding (NLU) problems, including question answering, natural language inference and linguistic acceptability judgements.

The minimal pair design is common in BabyLM Evaluation datasets. Examples include semantic minimal pairs dataset that evaluate property inheritance (COMPS (Misra, Rayz & Ettinger 2023)) and SyntaxGym (Gauthier, Hu, Wilcox, Qian & Levy 2020), which focuses on minimal syntactic variations at critical regions across a range of syntactic constructions (e.g., relative clauses).

Other evaluation metrics for BabyLM architectures include the **Elements of World Knowledge (EWoK)** (Ivanova, Sathe, Lipkin, Kumar, Radkani, Clark, Kauf, Hu, Pramod, Grand, Paulun, Ryskina, Akyürek, Wilcox, Rashid, Choshen, Levy, Fedorenko, Tenenbaum & Andreas 2024), which targets conceptual knowledge from multiple knowledge domains and certain pragmatic capabilities. EWoK uses both traditional plausibility estimates via log probability and two prompt-based strategies called Likert and Choice. The metric for correctness of a given item is the recovery of the designed item structure such that

$$\text{score}(T_1 \mid C_1) > \text{score}(T_1 \mid C_2)$$

and

$$\text{score}(T_2 \mid C_1) < \text{score}(T_2 \mid C_2),$$

where score reflects P_θ for log probabilities, an integer rating for LIKERT, and the correct context index selection for CHOICE, and T is the target sentence and C is the context of the minimal pair. The second BabyLM Challenge additionally introduced a multimodal evaluation track, using an evaluation dataset called Winground (Thrush, Jiang, Bartolo, Singh, Williams, Kiela & Ross 2022) (inspired by the Winograd Schema Challenge in coreference resolution) to measure the preference of a Vision-Language Models for minimally different captions associated with a target image that are permuted to highlight different object/action relationships (i.e., given an image that shows a LIGHTBULB SURROUNDING PLANTS, the model is presented with two sentences: *Some plants surround a lightbulb* v. *Some lightbulbs surround a plant*). Table 1 shows an example of benchmarking in the BabyLM Shared Task, which uses BLiMP alongside GLUE and EWoK.

Model	BLiMP	BLiMP Suppl.	EWoK	GLUE	Av.
BabyLlama (Timiryasov & Tastet 2023)	69.8	59.5	50.7	63.3	60.8
LTG-BERT (Samuel, Kutuzov, Øvrelid & Velldal 2023)	60.6	60.8	48.9	60.3	57.7

Table 1 Example of Language Model Evaluation from the BabyLM Shared Task 2024 (Hu et al. 2024).

Engineering cognitively-inspired architectures involves modelling choices that link language models to human language processing and developing cognitively-motivated methods to improve model interpretability (Beinborn & Hollenstein 2023). Among this direction includes work on phoneme-based training of BabyLMs, where tokens consist of individual phonemes, with word boundaries removed but still only train on English text, as established in Bunzeck, Duran, Schade & Zarrieß (2024) and Goriely, Diehl Martinez, Caines, Buttery & Beinborn (2024). One motivation for this comes from computational psycholinguistics research that has utilised statistical learning models to compute string surprisal, hypothesised to correlate with the difficulty incurred by a comprehender during lexical processing (Hale 2001). Statistical learning models are used to test theories that relate a model's 'surprisal' of a substring of characters to model the cognitive cost (approximated, for example, by gaze duration) experienced by readers in tasks (Futrell, Wilcox, Morita, Qian, Ballesteros & Levy 2019, Schrimpf, Blank, Tuckute, Kauf, Hosseini, Kanwisher, Tenenbaum & Fedorenko 2021).

3 In Vivo and In Silico Learners: A fundamental difference?

LLMs have been characterised as '*in silico* learners' that depart from *in vivo*, or natural, language learning in humans in various ways. The existence of a fundamental difference between *in vivo* and *in silico* learners is used to justify the position that "in principle, [language models] can tell us nothing about language, language

acquisition, human cognition, anything" (Chomsky, personal communication reported in Millière (2024)). It is possible to broadly classify two positions on *in vivo* and *in silico*: (1) a weaker **substantive difference hypotheses** and (2) a stronger **irreducible difference hypotheses**.

Representative of the former, Dentella, Guenther & Leivada (2024) hypothesise that the *in silico* learning of BabyLMs differ in at least three respects. First, the type of evidence available to a learner (in particular the availability and effective utilisation of negative evidence), the absence of effectively utilising the 'poverty of stimulus' in the input to promote effective generalisation, and the occurrence of semantic hallucinations due to 'impenetrable' linguistic reference.

The 'irreducible difference' position puts forward a much stronger difference between LLMs and human learners[1]. Fox & Katzir (2024) stipulate that the inherent non-modularity of Language Models means that LLMs cannot be considered a scientific theory. Additionally, the lack of explicit consideration of constituency and entailment means, which they argue are 'parts of all the best theories of human linguistic cognition'. It discounts any possibility that LLMs can be considered an adequate or scientific theory. By implication, this has been used by linguists to justify the following typological position: LLMs process possible and impossible languages indistinguishably (Roberts, Watumull & Chomsky 2023).

Arguments that specifically criticise the unclear *explanandum* and *explanans* of LLMs are also consistent with this position of 'irreducible difference'. Baker (2024) claims this represents a fundamental limitation in the predictive utility and explanatory adequacy of LLMs. The first requirement for a language model to be an adequate theory of cognition, according to this stronger view, is that non-trivial predictive linguistic generalisations should be testable 'out of distribution' on a new test case. Additionally, to satisfy the status of being a theory (as controversially claimed by Ambridge & Blything (2024) and Piantadosi (2023)), irreducible difference arguments demand that LLMs– and, more importantly, any mechanisms that underpin the behaviour – should facilitate deductive and inferential scientific exploration.

In practice, this implies that if Language Models are to offer any form of linguistic exploration, then this should be equivalent to how theoretical linguists might formulate generalisations that derive typological distributions, acquisitional developmental sequences, markedness preferences or any other object of linguist theorising (Mallory 2024). In order for there to be any theoretical utility derived from a Language Model, this should rely on striking an equivalence between an untrained Language Model and predictive 'algorithmic theories', whereby an uninitialised model can be viewed as a space of possible grammars that can vary based on architecture and parameter size, although this is with two important caveats: (1) model selection must be cognitively-motivated and (2) it assumes a mechanistic understanding of the 'inner workings' of the architecture.[2]

[1] For further comprehensive exposition, see Millière (2024).

[2] See Baroni (2022) for an example of this argument.

Model	$\|V\|$	Human Equivalent
CHILDES Transformer	8.6M	10 months
BabyLM 10M	10M	1 year
BabyLM 100M	100M	8 years
GPT2	8B	730 years
Llama 3.2-3B	9T	821,250 years

Table 2 Human-Equivalent Approximates of Data-Constrained Training Corpora. Data from Ziv et al. (2025).

3.1 A working hypothesis: BabyLMs decompose in vivo learning

Conceiving BabyLMs programmatically, I formulate a working hypothesis that cognitively-inspired language modelling has the *potential* to simulate a good **cognitive proxy** for a learner through the explicit modelling of systematic aspects of human linguistic cognition that might otherwise with obfuscated in LLMs.

Given this and broadly assuming the weaker position of 'substantive difference hypotheses', one of the first challenges of cognitively-inspired language modelling is characterising – and iteratively recharacterising – the **initial conditions** of simulation. As noted by Cuskley et al. (2024), precise comparisons between children and computational models are challenging. While training on CDS has enhanced ecological validity, it is certainly possible to introduce 'tricks' during pretraining, such as introducing large numbers of epochs. This will increase the amount that a model will 'see' the training data by a multiplicative factor. BabyLMs have since been constrained to only 10 epochs by Charpentier, Choshen, Cotterell, Gul, Hu, Jumelet, Linzen, Liu, Mueller, Ross et al. (2025).

Beyond calibrating the conditions for cognitively-inspired pretraining, the core 'top-down', or linguistically-inspired, goal for the BabyLM research programme is to carefully control certain aspects of a model architecture to construct carefully controlled experiments utilising data-constrained training corpora with naturalistic volume of input (equivalences illustrated in Table 2).

BabyLM experiments may either introduce techniques that improve the 'opportunism' of cognitive proxies in a sample-efficient learning setting or simulate the predictions of linguistic theories (as will be discussed in Section 4). To explicate the former, variation sets are an example of drawing on structured repetition in the input to increase the amount of data that can be justifiably included in a sample-efficient learning setting. Variation sets are partial self-repetitions in Child Directed Speech (CDS) – simplified input provided by caregivers to children– that are centred around a common frame and clustered in a short time span as shown in Example (1).

(1) **Variation Sets in English CDS (Howe Corpus)**:

a. Yes yes, he's got **toes**.

b. Four **toes**.

c. Have you got **toes**, Richard?

d. Where are your **toes**?

e. Show me your **toes**.

f. Come and show me your **toes**.

Haga, Fukatsu, Oba, Bisazza & Oseki (2024) strike an analogy with the naturalistic concept of **variation sets** to generate synthetic examples of CHILDES corpora for English in the BabyLM pretraining dataset (Warstadt et al. 2023), and experiment with two methods for inputting variation sets into a model during pretraining to compare the empirical benefits of concatenating variation sets into a single sequence and distributing each sentence of a Variation Set into adjacent batches.

Other approaches seek to draw cognitively-inspired interpretations of techniques for controlling a Transformer-based architecture. For example, Press, Smith & Lewis (2022) introduce a technique to control length extrapolation in Transformers by biasing query-key attention scores with a penalty that encodes an inductive bias towards recency; penalising attention scores between distant query-key pairs, with the penalty increasing as the distance between a key and a query grows. This method is repurposed in a BabyLM context by Mita, Yoshida & Oseki (2025). Building on Clark, Oh & Schuler (2025)'s incorporation of a recency bias into attention score computation during training, they are able to simulate an exponentially growing working memory trajectory that dynamically changes during training with a globally coherent bias.

3.2 Developmentally-plausible benchmarking

Another related challenge is characterising the 'dimensions' of comparison: the training dynamics of cognitively-inspired models should ideally be compared to human developmental trajectories in language acquisition (Lavechin, De Seyssel, Titeux, Bredin, Wisniewski, Cristia & Dupoux 2022, Evanson, Lakretz & King 2023, Yang, Wang, Plonsky, Oswald & Chen 2024, Charpentier et al. 2025). This relies on appropriate 'developmental benchmarking' of Language Models. However, interpreting and drawing inferences through comparing checkpoints to human developmental sequences is unclear (Chemla & Nefdt 2024). Multimodal Vision-Language Model benchmarks have been developed, which contain subtasks that evaluate visual and linguistic abilities that emerge at different stages of children's development. Tan, Yu, Long, Ma, Murray, Silverman, Yeatman & Frank (2024) contains subtasks where (i) the model must pick the correct image associated with a given word; (ii) the model must pick the correct image corresponding to a sentence; and (iii) the model must assign appropriately higher or lower similarity scores to more or less similar images.

4 A Language Acquisition Case Study: Comparing Chomskyan acquisition theories with Small Cognitively-Inspired Language Models

I now outline a detailed case study expanded from earlier work in Salhan et al. (2024), which introduces precise implementations of the developmental sequences of contrastive acquisition theories in BabyLMs. These are based on contemporary Chomksyan acquisition models, including Biberauer's (2019) 'Maximise Minimal Means' model and the Growing Trees Hypothesis (Friedmann, Belletti & Rizzi 2021). We compare the success of three **Curriculum Learning (CL)** strategies (Growing, Inwards & MMM) that precisely replicate the predictions of contrastive acquisition theories to specify fine-grained curriculum learning strategies on a standard SSLM architecture trained on a volume of Child-Directed Speech (CDS) that a learner would expect to receive by 6 years-old (6;0). To implement **cognitively-inspired small-scale language models (*henceforth referred to as* SSLMs)** and acquisition-inspired curricula cross-lingually, we create age-ordered corpora of CDS for four typologically distant language families (Sino-Tibetan, Romance, Germanic and Japonic).

4.1 Preliminaries: three Chomskyan models for acquisition of the syntactic category system

The **biologisation issue** (Bosch 2024) is the question of how much the acquisition task for a learner is determined by innately pre-wired structures/mechanisms. First, early Generative work has proposed a cross-lingually uniform **maturation** of the functional spine at distinct points in learning. This is a strong 'biologisation' hypothesis, assuming a Universal Grammar (UG) encodes not only universal structural primitives but also when they appear. The **'Growing Trees' Hypothesis** (Friedmann et al. 2021) is the latest instantiation of this hypothesis, proposing L1 learners do not have access to the higher layers in the functional spine of the clause. Maturational developmentally hard-wired mechanisms dictate the domains of the clause that are available to the learner. The second maturational developmentally hard-wired possibility is **inward maturation**. Based on evidence of early acquisition of 'discourse'-material and interactional language (e.g., tags-questions), Heim & Wiltschko (2021) propose, in a rather programmatic proposal focusing primarily on interactional language, that acquisition 'starts from the edges, and develops inwardly'.

Neo-emergentism (Biberauer & Roberts 2015) takes a **categorial granularity** approach to language development. It assumes an initial stage where the learner makes a **bipartite** distinction between the thematic domain of the clause (vP) and some CP-structures, such as the early emergence of verb-second (V2) word order in Dutch before inflectional and tense-based knowledge (van Kampen 2010), and sim-

ilarly the early emergence of *wh*-questions and focusing in Greek between 1;9-1;11 (Tsimpli 2005)[3].

4.2 Training corpora: multilingual age-ordered & IPA 'phonemicised' CHILDES

In Salhan et al. (2024), a training corpus of **Age-ordered Child-Directed Speech (CDS)** is collected for 18 languages (French, German, Japanese and Chinese), in addition to the English Age-Ordered-CHILDES (AO-CHILDES) corpus (Huebner & 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[4]. Goriely & Buttery (2025b) further introduce **phonemised age-ordered CHILDES (IPA CHILDES)**.

The resulting dataset contains over 45 million words of child-directed speech across 31 languages, hosted on Huggingface[5]. We encourage researchers to use this resource.

One major bottleneck is the lack of adequate evaluation resources beyond English to analyse the linguistic capabilities of Language Models. Another limitation is that the distribution of data beyond English is heavily skewed. As shown in Table 1, EnglishNA is the most represented, with close to 10 million words, however, and Farsi is the least represented, with only 43 thousand words.

4.3 Implementation

Salhan et al. (2024) compares the success of three curricula (GROWING, INWARDS & MMM) that precisely replicate the predictions of contrastive acquisition theories to specify fine-grained curriculum learning strategies on a standard SSLM architecture trained on a volume of Child-Directed Speech (CDS) that a learner would expect to receive by 6 years-old (6;0) for four typologically distant language families (Sino-Tibetan, Romance, Germanic and Japonic).

We implement three contemporary cross-lingual models of syntactic acquisition use the predicted developmental sequences of each model:

i. **GROWING**: Bottom-up maturational approaches to language acquisition (Rizzi 1993, Radford 1990), including the 'Growing Trees Hypothesis' (Friedmann

[3] Another analytic possibility is one of **continuity**: this proposes that the functional structure of children's initial grammar is not significantly different from adults' grammars. In the context of developing SSLMs, this can be considered to be a null hypothesis. However, linguistically, a **Strong Continuity Hypothesis (SCH)** (Poeppel & Wexler 1993) appears to be untenable for the evidence of selective and gradual development of a portion of functional heads during early acquisition. These accounts necessarily have to rely on the selective unavailability of certain aspects of the clause through other mechanisms like underspecification, and cannot be ascribed to phonological reduction and prosodic licensing (Mitrofanova 2018).

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

[5] `https://huggingface.co/datasets/phonemetransformers/CHILDES`

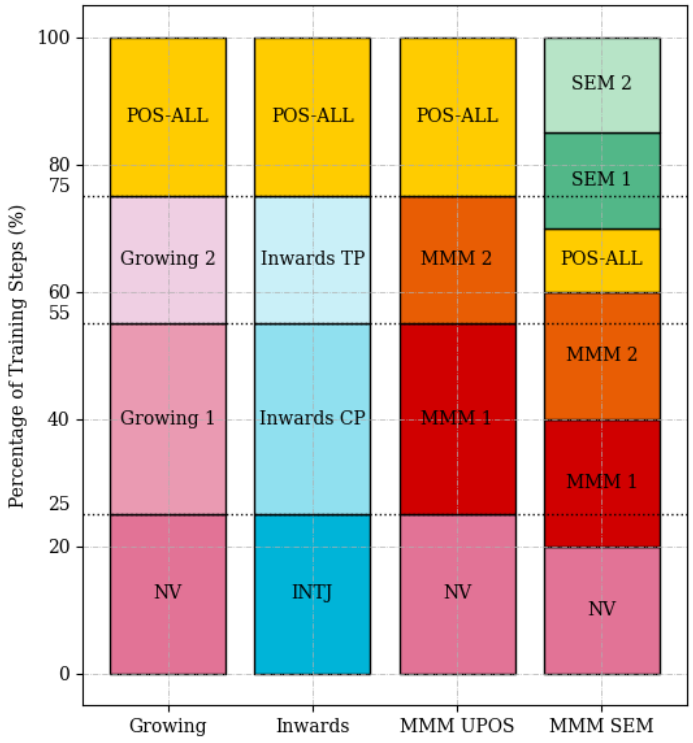

Figure 1 **Acquisition-inspired Objective Curricula (Salhan et al. 2024):** We specify Objective Curricula Growing, Inwards, MMM (UPOS), MMM (Semantic) for three theories of acquisition. The Progression of Curriculum Units replicate the predicted developmental sequences by specifying curriculum units (defined in Table 2) defined over different pre-training stages, expressed as a percentage of training steps.

et al. 2021), predicts that first language learners begin acquiring verbs and nouns (unit NV in Table 2). 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 2[6].

[6] 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).

ii. **Inwards**: Bosch (2023) tests the predictions of the predictions of a **generalised inward-growing maturational proposal (Inwards)**, building on evidence from Heim & 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 2); followed by the acquisition of tense/event-related information (Inwards-TP); and finally, thematic information.

iii. **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 & Roberts 2015, Bosch 2023). The general universal prediction of one neo-emergent model called Maximise Minimal Means (MMM) is that all learners, irrespective of 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)[7], before finally learning tense/aspectual categories (units MMM 1 and MMM 2 in Table 2). 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 2.

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 later in training. In practice, this is implemented by Salhan et al. (2024) by modifying the masked language model training objective of a BERT-style language to mask tokens belonging to increasing sets syntactic categories (in a supervised tagging setting). This relies on varying the tagset for a proportion of training.

4.4 Evaluation and results

Our BabyLMs are evaluated on syntactic minimal pair datasets that have been introduced beyond English.

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

	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 3 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 trained on the STRICT Track of the 1st BabyLM Shared Task (Warstadt et al. 2023), so this reports a results of models trained on a combination of WIKI+CDS +Other Simplified Texts. **Bolded** results indicate the highest accuracy of all the models.

i. **CLAMS (French and German)**: The Cross-Lingual Syntactic Evaluation of Word Prediction Models (CLAMS; Mueller, Nicolai, Petrou-Zeniou, Talmina & Linzen 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).

ii. **JBLiMP (Japanese)**: JBLiMP (Someya & 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*[8,9].

iii. **SLING (Chinese)**: **SLING** (Song, Krishna, Bhatt & Iyyer 2022) is a 38K minimal sentence pair dataset derived by applying syntactic and lexical transformations to Chinese Treebank 9.0, aiming to improve on the limitations of an

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

[9] Due to the small size of the JBLiMP minimal pairs dataset, Someya & Oseki (2023) recommend 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.

earlier dataset called CLiMP (Xiang, Yang, Li, Warstadt & Kann 2021), which had a lack of diversity in the vocabulary to generate minimal pair templates[10].

As shown in Table 3, fine-grained acquisition-inspired curricula can outperform non-curriculum baselines and are effective in English, Chinese and Japanese. Different strategies lead to better performance for certain languages, particularly finer-grained language-specific versions of the MMM objective. Further results in Salhan et al. (2024) find 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 with acquisition-inspired CL strategies in Japanese significantly outperform GPT-2.

4.5 Discussion and potential extensions

While proposed methodology leads to improved performance, it also raises many unanswered questions for cognitively-inspired modelling. There are important caveats: the evaluation resources beyond English are inadequate and, more importantly, the modelling approach adopted here has limitations. Transitions between stages were 'hard-coded' imprecisely based on a proportion of training steps, however, stages in acquisition crucially operates with MLU, rather than age (Bosch 2023). Ideally, we would be able to dynamically model transitions between stages in a developmental sequence using a proxy of generated textual output of a model. There have been initial attempts (e.g., Arnett, Chang, Michaelov & Bergen (2025) and Oba, Kuribayashi, Ouchi & Watanabe (2023)) to extend the BabyLM Shared Task and cognitively-inspired modelling to bilingual and second language learning.

Although both maturational 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. 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. The results could potentially be adduced as a form of external evidence supporting the inadequacy of universal or maturational models, which has been argued by Bosch (2023, 2024) *inter alia*. However, the implication that neo-emergent BabyLMs (i.e., models trained with the MMM curricula) are somehow 'better' than maturational models would lead us to expect that finer-grained curricula that are more precisely aligned with human acquisition would perform better. There are observed benefits of the more fine-grained the MMM (Sem) curriculum which incorporates two additional stages to the non-language specific strategy to define a language-specific

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

curricula that utilises semantic tags (Bjerva, Plank & Bos 2016), or *sem*-tags, to model **language-specific acquisition strategies** on certain BLiMP test sets associated with ellipsis, but in Chinese the MMM (SEM) curriculum marginally underperforms compared to original (less granular) MMM (UPOS) when handling anaphora and aspectual phenomena. This is an unclear conclusion and is a point of departure from what is expected from linguistic theory. This case study highlights the challenge of drawing inferences from BabyLMs to language acquisition in a consistent and predictive manner, although the framework is empirically 'progressive' – generating new corpora, implementing and testing competing hypotheses postulated in language acquisition.

5 The Potential, and Potential Pitfalls, of BabyLMs in Linguistic Theory

5.1 Potential pitfall I: avoiding anthropomorphism and 'anthropofabulation'

One overriding issue for BabyLM research is anthropocentrism: it is an overriding heuristic for guiding cognitively-inspired language modelling and drawing inferences from controlled experimentation for linguistic theory. Millière & Rathkopf (2024) introduce bipartite criteria for Type-I and Type-II anthropocentrism, which I summarise as follows:

i. **Type-I anthropocentrism**: If a model m fails on an instances of a task from a broader class of 'competence tasks' C, $c_i \in C$, m has not acquired the competence associated with C. This does not account for 'auxiliary factors' associated with the task space C.

ii. **Type-II anthropocentrism**: Human competence is an 'investigative kind' (Boyle 2024), or a 'reference template' for evaluating a cognitive competence, but there may be mechanistic or metalinguistic differences in realisation.

Certain claims made in 'irreducible difference' critiques may fall into Type-I errors: recent studies have suggested that LLMs might show subtle judgements on rare constructions like the English Article+Adjective+Numeral+Noun (AANN) construction ('a beautiful five days'). This is attributed by Misra & Mahowald (2024) to the 'generalisation' capabilities of language models adduced from more frequent examples. While this may not be an absolute mechanism for syntactic or morphological generalisation, interpreting these mechanisms could potentially enable empirical assessment of hypotheses associated with Construction Grammars, whose predictions have been hypothesised to align with statistical learning in Transformer-based Language Models. As surveyed in Salhan (2023), one reason for drawing this analogy is due to the design of attention heads in Transformers, which are not informally encapsulated to semantic information. Even in attention heads that are the best candidates for syntactic encapsulation, syntactic information is penetrable to semantics. McGee & Blank (2024) find that semantic implau-

sibility can reduce attention between the words that constitute the dependency for which a head is specialised.

However, 'substantive difference' critiques do not fall into Type-I errors, as they offer a ***systematic difference*** between a family of BabyLM and LLM architectures and desirable human cognitive behaviour. An example of this is that a learner (artificial or natural) should be able to utilise latent indirect evidence in some way. Oba, Oseki, Fukatsu, Haga, Ouchi, Watanabe & Sugawara (2024) introduce a data augmentation strategy to assess the indirect learning capabilities of a Language Model, by inserting `wug` either as a means of lexical indirect evidence (referring to training items with similar usage) or syntactic indirect evidence, finding models struggle to induce humanlike linguistic generalisation even with a degree of indirectness.

Type-II anthropocentrism bears some similarities to what Buckner (2013) calls 'anthropofabulation' in comparative cognition, which refers to the combined (1) overestimation of the consistency, domain-generality, and reflective nature of human cognition in everyday situations and (2) semantic anthropocentrism through referring to the 'theory of mind' or 'episodic memory' of non-humans, which leads to idealised and skewed perceptions of human performance. It may, for example, lead to overestimating the contribution of a 'interpretable' feature (attribute of interest) or circuit of interest from a larger network of circuits. This might be the case if that feature or circuit does not significantly contribute to the network's functionality where multiple circuits are competing for influence on model behaviour (Marks, Rager, Michaud, Belinkov, Bau & Mueller 2024).

While empirical findings of consistent 'grammar learning' trajectories in Transformer-based architecture that are invariant to model size (c.f. Choshen, Hacohen, Weinshall & Abend 2022) may be adduced, a desired attribute of cognitively-inspired BabyLMs– **developmental benchmarking** – is a strong case of Type-II anthropocentrism. There are well-studied differences between the learning dynamics of Transformer-based Language Models and human language acquisition. There are two attested training dynamics for Language Models across scales:

i. A critical phase change underlies 'grokking' behaviour and in-context learning abilities (Power, Burda, Edwards, Babuschkin & Misra 2022, Olsson, Elhage, Nanda, Joseph, DasSarma, Henighan, Mann, Askell, Bai, Chen, Conerly, Drain, Ganguli, Hatfield-Dodds, Hernandez, Johnston, Jones, Kernion, Lovitt, Ndousse, Amodei, Brown, Clark, Kaplan, McCandlish & Olah 2022, Liu, Kitouni, Nolte, Michaud, Tegmark & Williams 2022) with *sharp* and *unpredictable* transitions across scales (Schaeffer, Miranda & Koyejo 2023).

ii. Other tasks observe a steady evolution of abilities as training progresses, for example learning to reduce the perplexity of grammatical sequences containing hallucinations – with small models halting at a suboptimal distribution. Additionally, at a given perplexity and independent of model sizes, a similar subset of training tokens see the most significant reduction in loss, with the rest stagnating or showing double-descent behaviour, where performance first improves, then gets worse, and then improves again with increasing model size, data size, or training time (Xia, Artetxe, Zhou, Lin, Pasunuru,

Chen, Zettlemoyer & Stoyanov 2023). One on hand, it is a necessary attribute if BabyLMs can be used in a comparable way to other computational models of acquisition (as was attempted in the case study in Section 4), but strong alignment with human development may not be a useful hallmark for sample-efficient small language models.

Additionally, instances of sudden learning forgetting in Transformers reflect changes in model processing appears consistently: Chang, Tu & Bergen (2024) finds that this cannot be attributed to random chance or specific examples, suggesting that a sudden 'burstiness' of learning potential can be attributed to some systematic difference between a model (and associated optimisers) up until that training interval.

Another instance of potentially 'anthropofabulated' model design are Modified Transformer architectures, like **Transformer Grammars** (Sartran, Barrett, Kuncoro, Stanojević, Blunsom & Dyer 2022: i.a). These augment the *vanilla* Transformer architecture with syntax-optimised inductive biases (c.f. Kuncoro 2022: for syntax-inspired methods). As syntactic generalisation scores have been found to, at least partially, dissociate from information-theoretic metrics like perplexity (Sartran et al. 2022), it is possible that scaling behaviour (sharp grokking trajectories vs. continuous scaling trajectories) can be partitioned for syntactic and semantic generalisation (see Choshen et al. (2022) for a similar suggestion). Tree-structuredness metrics (Murty, Sharma, Andreas & Manning 2022) have been linked to the optimal depth for grokking, a phenomenon where models generalise long after overfitting their training set on structurally novel sentences – increasing gradually, long after the performance on sentences from the training distribution has plateaued (Liu et al. 2022, Murty, Sharma, Andreas & Manning 2023, Wang, Yue, Su & Sun 2024).

5.2 Potential pitfall II: don't conflate performance with algorithmic reasoning capabilities of transformers!

Transformers learn permutation-symmetric functions, which limits the algorithmic reasoning capabilities of the architecture. The architecture fails on simple copying tasks (e.g, single-digit copying tasks) where models cannot copy the 0 past a certain bitstring size, failing to length-generalise. Other algorithmic tasks, which are surveyed in detail in Salhan (2023: 84), are used to assess the expressive ability of Transformers (and other similar models). These are essential for grasping the capacity limitations of Transformers, establishing circuit complexity bounds for Transformer architecture. These circuits, if associated with standard computational complexity classes of linguistic problems or phenomena, could allow us to more precisely delimit complexity classes that are desirable to capture in cognitively-inspired language modelling, which leads us to propose the following

Occam's Razor of Cognitively-Inspired Language Modelling: The computational complexity classes associated with different algorithmic implementations of rules or generalisations have been well studied. Ideally, we might hope through the

BabyLM paradigm, we might be able to converge on a neural architecture that characterises predictively the minimal conditions for linguistic capabilities to emerge[11].

5.3 Potential pitfall III: what does a cognitively-plausible evaluation look like?

Fox & Katzir (2024) stipulate that the 'LLM Theory, on the other hand, seems content with a mechanism that can only output probabilities and where nothing even remotely similar to a distinct notion of correctness has ever been identified'. This is inaccurate, insofar, as there is a nascent evaluation paradigm for evaluating the syntactic capabilities of Language Models involving metalinguistic judgement (prompting a model with a task that requires a linguistic judgement) or direct probability measurements that estimate the probability of a sentence or a critical region/word treat Language Models as psycholinguistic subjects (Futrell et al. 2019). Dentella et al. (2024) object to the use of direct probability measures, however, comparing relative probabilities assigned to minimally different sentences does provide some means to analyse the sensitivity of a Language Model to a target syntactic feature and may reveal graded linguistic knowledge associated with judgements (Millière 2024).

There are more severe challenges with BLiMP, which has fundamental limitations as a benchmark of human linguistic competence. The dataset is **artificially generated** using from abstract grammars that exemplify syntactic phenomena – this easily yields a large number of sentences, which can help control for other possible sources of noise in test materials using generation scripts. This relies on templates to sample lexical items with selectional restrictions that annotate the morphological, syntactic, and semantic features of over 3,000 items.

BLiMP attempts to have broad syntactic coverage, with 'sub-phenomena' including coverage of minimal pairs of Filler-Gap Dependencies that arise from phrasal movement in – as in *wh*-questions – including across interveners. BLiMP's Binding dataset only covers anaphora (Principle A of Theory Binding; Chomsky 1981, 1986) in simple cases of reconstruction (e.g., *It's himself that* ***this cashier attacked/*attacked this cashier***.) and across domains (e.g., *Steven explains Kayla won't hurt herself.* vs. *Kayla explains Steven won't hurt herself.*) Since co-indexation cannot be annotated in BLiMP, Principles B and C, which characterise restrictions on pronouns and R-expressions, are not contained in the minimal pairs dataset. BLiMP's Control/raising constructions highlight syntactic and semantic differences between various types of predicates in non-finite clauses which embed an infinitival VP in three subconstructions: ***tough*-movement predicates** that involve verbs like *tough/difficult/easy* that allow the subject of the matrix clause to appear semantically as the object of the embedded clause (*Julia wasn't fun to talk to.* vs. **Julia wasn't unlikely to talk to*); cases of **existential *there*** (*William has declared there to be no guests getting fired.* vs. **William has obliged there to be no guests getting fired.*) and **expletive *it*** in simple cases of raising (e.g., *Carla could declare it to be not so important that these doctors observe Rhonda.* vs. **Carla could convince it to be not so important that these doctors observe Rhonda.*) BLiMP does not offer full

[11] See Ueda, Kuribayashi, Kando & Inui (2025) for preliminary discussion.

coverage of ellipsis, since it only considers sentences of equal length, only covering very restricted cases of N-bar Ellipsis that meet this practical constraint.

BLiMP is being used as 'template' for evaluating syntactic competence cross-lingually. BLiMP, CLiMP, SLING and JBLiMP all use a forced-choice paradigm to validate their minimal pairs with human native speakers. All papers explore the effect of training data size – CLiMP and JBLiMP found no influence of dataset size, while SLING found that smaller models may have performed better for some. The performance gap between the LMs and the native speakers is large on these cross-lingual minimal pairs datasets (and larger than it was for English). Also, models perform better at local dependencies compared to longer-distance dependencies. SLING highlights a few important properties of Mandarin syntax. Chinese has a rich system of **classifiers**, so there is an additional syntactic task of **classifier-noun agreement** when a noun is modified by a numeral or demonstrative. **Chinese Definiteness Effect** is a restriction of the distribution of *zhe* (this)/*na* (that) and the quantifier *mei* (every), which may not occur in the post-verbal position of an existential *you* (there is) sentence. Chinese has perfective aspect markers *le* and *guo*. SLING contains minimal pairs that contrast these markers with the tense and the progressive marker *zai*. JBLiMP generalises BLiMP's irregular forms dataset to incorporate minimal pairs on morphology in general. Japanese doesn't have explicit determiner-noun agreements, so JBLiMP drops BLiMP's determiner-noun agreement category for a more general Nominal Structure dataset.

5.4 Potential pitfall IV: can transformer-based BabyLMs ever be scientific models if they are not interpretable?

Transformer-based BabyLMs need to be *interpretable* to be 'good' scientific models: In joint work, we investigated the effect of frequency information in Language Models. Frequency information in token distributions forms a useful heuristic for learners during first language acquisition. As hypothesised in the Tolerance Principle (Yang 2018, Schuler, Yang & Newport 2016), token distributions drive learning trajectories, while type distributions drive learning outcome **(the generalisation gap)** to the maximum likelihood training objective that uses a cross-entropy loss between the label of the correct word and predicted probabilities from a forward pass of the model. In Chung, Hong, Salhan, Kim, Diehl Martinez, Thorne & Buttery (2025), we pretrain Language Models from scratch across scales (14M, 162M) on a diverse 420B token corpus with different regularisation strategies that affect different components of the model: (1) weight tying which produces a 'uniform' embedding space and (2) an auxiliary loss that penalises the softmax output of a Transformer. For the 14M series, the only above-chance dataset is BLiMP.

We observe that Z-Loss and Tying lead to performance improvements, as shown by the detailed breakdown of accuracy by syntactic phenomena in Table 4. A combination of tying and Z-loss leads to the highest performance on binding (79.13%). Z-loss and Tying also leads to around a +10% improvement on ellipsis (75%) and irregular forms (95%). However, accuracy improvements are not observed for certain datasets, such as NPI licensing or quantification, which tend to rely on better

sample efficiency and generalisation. There is only above chance performance on Island Effects, another longer-distance syntactic generalisation task related to enhanced generalisation. However, it is unclear why we see principled increases in existing measures of syntactic competence through architectural changes.

Phenomena	Untied		Tied	
	CE Loss	Z Loss	CE Loss	Z Loss
anaphor agreement	0.922	0.9185	0.915	**0.96**
argument structure	0.6821	**0.7629**	0.7492	0.7427
binding	0.7110	0.7751	0.764	**0.7913**
control raising	0.7250	0.7656	**0.7822**	0.7326
determiner noun agreement	0.8194	0.8785	**0.8871**	0.8804
ellipsis	0.6605	0.7515	**0.7785**	0.73
filler gap dependency	0.5183	0.5417	0.5187	**0.5619**
irregular forms	0.8835	**0.9560**	0.9510	0.9385
island effects	0.4928	0.4752	0.4579	**0.5138**
npi licensing	**0.6949**	0.6693	0.6550	0.6594
quantifiers	0.5963	0.6350	0.5973	**0.6525**
subject verb agreement	0.7567	**0.8388**	0.7850	0.7875
Average	0.7052	**0.7473**	0.7367	0.7459

Table 4 Detailed BLiMP Accuracy Scores for 14M Model Series.

Theoretical linguistics provides a blueprint for interpretability: Given a minimal pairs dataset $\mathcal{D}$ of contrastive datasets, one popular approach to interpret Transformers utilises Sparse Variational Autoencoders (SAEs) to compute a decomposition of an input activation x into an approximate reconstruction $\hat{x}$ to various hidden states in models. SAEs can be trained on attention and MLP outputs and residual stream activations for feature disentanglement of each model component and then quantify the importance of an activation a on a pair of inputs $x_{\text{clean}}, x_{\text{patch}}$. A large indirect effect is a proxy for the influence of a neuron on a model's decision.

In this approach, a model is represented by a computation graph G that takes feature activations f_i and SAE errors ε at particular token positions as nodes that are part of the Language Model's computation. We use this to identify **Sparse Feature Circuits**. These are computational sub-graphs that explain model behaviours in terms of SAE features and error terms. Datasets like CausalGYM (Arora, Jurafsky & Potts 2024) take an input minimal pair that has an alternation that affects next-token prediction, then intervenes on the base forward pass using a pre-defined intervention function that operates on aligned representations from both inputs. Then, it is possible to determine how this intervention impacts next-token predic-

tion probabilities. In aggregate, such interventions assess the causal role of the intervened representation on the model's behaviour We can use **directionality** for causal effect as an intuitive test for whether they reflect features that the model uses downstream. **Distributed alignment search (DAS)** learns the **intervention direction**, potentially distributed across many neurons, that maximises the output probability of a **counterfactual label**. The counterfactual label is obtained by recasting a minimal pair, like S-V agreement, from SyntaxGYM into counterfactual pairs that elicit singular or plural verbs based on the number feature of the subject, and hold everything else (including the distractor) constant: (a) *The author near the senators* → *is* (b) The authors near the senators → are. One of the advantages of this paradigm is that it facilitates an analysis of **model learning dynamics** rather than analysing input/output relationships. Circuits that underpin simple 'linguistic tasks', such as synthetic subject-verb agreement, appear to be consistent across scale (Tigges, Hanna, Yu & Biderman 2024). In this sense, theoretical linguistics can help delimit further attributes of interest cross-linguistically to potentially identify 'circuits' for a wide range of syntactic phenomena.

Future Prospects: Cutting-edge developments in 'The Science of Language Models' (also formerly known as BERTology, or similar) highlight an implicit convergence in what computational linguists are interested in reconstructing in Language Models and (certain) goals of theoretical linguistics (Marcolli, Berwick & Chomsky 2023b). Both theoretical linguists and NLP practitioners have converged on similar mathematical formalisms like Hopf Algebras to model syntactic compositionality and the Transformer self-attention mechanism (Marcolli, Chomsky & Berwick 2023d, Marcolli, Berwick & Chomsky 2023a,c). Marcolli et al. (2023c) propose an algebraic model for the Syntax-Semantics Interface based on **Hopf Algebras**. Following Minimalist assumptions, narrow syntax is defined as a set of Syntactic Objects (lexical items and formal features), a set of *accessible* Syntactic Algebras and commutative Hopf Algebras. Commutative Hopf Algebras represent Workspaces given by Vector Space spanned by the set of Syntactic Objects. (External) Merge acts on the Workspace (Marcolli et al. 2023d). Hopf Algebras model the **symmetry** and **duality** of syntactic structure, representing syntactic structure as a series of composable elements which can be rearranged in different ways without changing the meaning of a sentence. They are tensorial bialgebras – both a tensor and a cotensor at once.

A crucial takeaway from the Hopf Algebra model is that a viable model of the syntax-semantics interface can be satisfied by *several* semantic frameworks. One consequence of using Hopf Algebras is that it suggests there are 'several approaches to the construction of possible models of semantics, which are, in our view, not entirely satisfactory and not entirely compatible' (Marcolli et al. 2023c: 9). A plurality of semantic frameworks within a syntactico-centric algebraic model is not as serious an obstacle as it may first seem. This includes truth-conditional semantics and distributional semantics techniques that underpin LLMs, which are well known to suffer from several fundamental semantic problems, ranging from grounding to quantification (Emerson 2020), alongside Pietroskian compositional

semantics (Unnsteinsson 2020)[12]. This represents an unprecedented theoretical convergence between Syntactic Theory and the mainstream approaches to Natural Language Syntax and Parsing not seen since the 1960s. Marcolli et al. (2023d) has a refreshingly clear takeaway that the 'image of Syntax' is encoded in the Transformer self-attention mechanism and that Hopf Algebras can provide a new lens of interpretability for Transformers[13]. This nascent algebraic and syntactico-centric approach, if practically translatable to data-efficient Language Models, would significantly improve the interpretability of Foundation Models in NLP; providing an explicit syntax-semantics representation that could improve performance in downstream Natural Language Understanding tasks in less data-intensive scenarios[14].

5.5 Implications for 'grammar engineering': testing linguistic theories

BabyLMs exhibit the capability to 'acquire' a language in the limit within *time* and *input* constraints of a learner and are equipotential, insofar as they can learn any possible human language in a self-supervised manner. Recent developments to train Language Models (LMs) using child-centered data in NLP research build on a wider range of computational approaches to first language (L1) acquisition, and the antecedents of the paradigm can be found in earlier computational models of acquisition, which provide a formal theory of developing grammars that link the growth of linguistic knowledge with the acquisitional mechanisms that enable it. One prominent methodology for computational models of first language acquisition learns from pairs of strings and meaning representations in a supervised manner (Siskind 1996, Villavicencio 2002, 2011, Buttery 2004, 2006, Kwiatkowski, Goldwater, Zettelmoyer & Steedman 2012), whether they learn from syntax and semantics jointly (as in Siskind 1996, Villavicencio 2002, Buttery 2006) or syntax alone (as in Gibson & Wexler 1994, Sakas & Fodor 2001, Yang 2002).

However, "vanilla" Transformer-based architecture may be inadequate for precise, controlled cognitive simulation since they utilise a several cognitively implausible or uninterpretable design choices, including arbitrary tokenisation strategies, large numbers of epochs and batched parameter updates. This is undoubtedly a challenge for precisely replicating well-motivated theoretical models of language acquisition in a BabyLM paradigm compared to other computational cognitive models that similarly utilise Child-Directed Speech (e.g., Mahon, Abend, Berger, Demuth, Johnson & Steedman (2025) and Abend, Kwiatkowski, Smith, Goldwater & Steedman (2017)'s recent categorial model of semantic bootstrapping using CDS in Hebrew and English (Szubert, Abend, Schneider et al. 2024)). But, given the programmatic status of cognitively-inspired language modelling and a *sufficiently constrained and interpretable architecture*, the same rationale that Bender,

[12] The latter, Pietroskian Semantics, is based on a single compositional principle called **Combine**(α, β), which takes every complex expression to encode a monadic concept. Marcolli et al. (2023d) reduce this to the properties of *Merge*.

[13] Cf. Nemecek (2023) for similar work on analysing Transformer self-attention using combinatorial Hopf Algebras

[14] Also of interest is the impact of interpretable Foundation Models on non-linguistic tasks, like code generation.

Flickinger & Oepen (2008) motivates for 'grammar engineering' could equally apply to BabyLMs: **validating and extending theoretical ideas through controlled models that relate strings from a fragment of natural language to interpretable grammatical representations**.

6 Takeaways and Conclusion

6.1 Implications for computational emergentism

The general approach to small-scale pre-training adopted in the BabyLM challenge is highly idealised– for example, relying solely on text input where the external environment shapes human acquisition trajectories through visual and auditory signals in concept learning and phonological acquisition (Biberauer 2011, Dresher 2009, Calabrese 1995). However, the case study outlined in Section 4 is a first-pass at implementing a holistic and crosslinguistically applicable model of syntactic development that is constrained enough to account for developmental universals, but flexible enough to capture developmental (language-specific) variation.

Modelling Phonological Emergentism. Another complementary direction of research highlighted in Section 2 aims to develop models that can be studied for their emergent phonological capabilities. As Goriely et al. (2024) note, training phoneme-based language models is sufficient to study the distributional properties of phonemes. Emergentist approaches in phonological theory propose that phonological features are proposed by learners during language acquisition. For instance, some processes, like final obstruent voicing, have shared phonetic properties (substantial impedance of airflow out of the vocal tract, and vocal fold vibration), but there are unnatural classes (e.g., the *ruki*-rule in Sanskrit: $*s \rightarrow$ ṣ $\backslash \{i, u, r, k\}\_$, e.g., $*h_1$ei $>$ èṣi). Emergentist learning algorithms, like the Successive Division Algorithm (Dresher 2009), rely on cues that learner extract from the input, and Mayer (2020) suggests that from an algorithmic perspective this necessitates consideration about the extent to which phonological classes are apparent in the distribution of sounds in a language, and to what extent do learners make use of this information. BabyLMs could be used potentially as an emergent phonological learner, which potentially may be able to draw insights about varying assumptions of feature visibility and underspecification across rule-based and Optimality Theoretic frameworks (Calabrese 1995, Nevins 2015).

6.2 BabyLMs, typology and language acquisition

The small-scale setting of the BabyLM Challenge can be a 'sandbox' for developing novel techniques that improve data efficiency, which can be extended to enhance current approaches to modelling low-resource languages– using language-specific subnetworks (Choenni, Garrette & Shutova 2022, Xu, Luo, Chang, Huang & Huang 2022) or meta-learning algorithms (Ponti 2021, Prokhorov 2021) – where data sparsity continues to be one of the main limiting factors for model performance.

Cognitively-Inspired Subword Tokenisation. The default tokenisation techniques used in Language Models segments text through a recursive process of

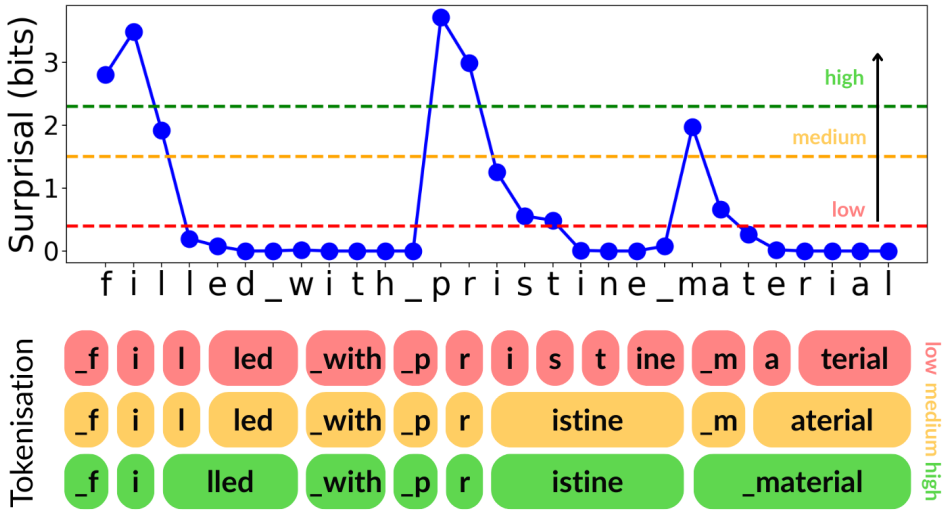

Figure 2 **Learning Subword Vocabulary using Acquisition-Inspired Tokenisation Strategy Multiple Information Thresholds**. Figure from Goriely et al. (2025b): Per-byte **surprisal** of *filled with pristine material* from a byte-level Language Model with three intermediate **information thresholds** (low, medium, high). All contiguous bytes below a threshold are grouped into a subword. The tokenisation strategy is inspired by computational models of word segmentation (Goriely et al. 2025a, Goriely & Buttery 2025a).

merging two characters in a process called **byte-pair encoding** (BPE), which does not yield a cognitively-plausible vocabulary for language modelling (Beinborn & Pinter 2023). This differs substantially from the process of word segmentation in child language acquisition. As surveyed in Goriely & Buttery (2025a), word segmentation can be modelled as a combination of chunking using distributional cues and predicting boundaries where there are spikes in surprisal or entropy that surpass a certain "information threshold". In joint work in Goriely, Salhan, Lesci, Cheng & Buttery (2025b), we use this insight to develop a novel tokenisation strategy for Language Models that is consistent with word segmentation cues. The approach, which we call ByteSpan, sets tokenisation boundaries using spikes in a language model's prediction error to group contiguous predictable byte sequences into subwords when below a certain information threshold. An example of this process is schematised in Figure 2.

This highlights a potential case study for how cognitively-inspired modelling and drawing parallels between computational models of acquisition and engineering techniques used in language modelling can lead to the generation of novel general-purpose techniques, which may potentially have greater typological extensibility than widely used NLP techniques that are potentially ill-suited for other languages.

Variation Sets and Structural Recycling. The release of MAO-CHILDES and IPA CHILDES are promising resources for fine-grained corpus analysis to develop precise implementations of hypotheses in language acquisition beyond English. Beyond the work of Haga et al. (2024), it could facilitate the development of cognitively-inspired models of acquisition in morphologically rich languages, such as Chintang, Turkish, and Inuktitut. In the acquisition of these languages, the prevalence of variation sets increases over time, as shown in the example below:

Variation Sets in Oaxaca Mixe/Ayöök (Mixe–Zoquean, Mexico) CDS (Pierson 2024):

(2) a. O'yip x'ee'k, uk va'an ëts n'ee'k?

Ø='oy-'ay-p	*x='ää'k-I*	*'uk*	*va'an*	*ëts*
3S.IND=good-VBZ-INC.INDI	2A.DEP-peel-INC.DEP	or	OPT	PRO1.SG

n='ää'k-I
1A.DEP=peel-INC.DEP

'Can you peel it, or should I peel it?'

b. Mits m'ää'kup?

mits	*m='ää'k-up*
PRO2.SG	2A.IND-peel-IRR.IND

'You'll peel it?'

c. Ää'küs!

'ää'k-ü=ts
peel-IMP=ASRT

'Peel it, then!'

This suggests that the caregivers actively employ this speech register to promote the acquisition of morphology as the children develop linguistically. Variation Sets are a specific form of **Structural Recycling** in language acquisition and cross-linguistically, there are observable systematic patterns of how variation sets, like other aspects of 'recycling', seem to be influenced by variables other than age (Moran, Lester, Gordon, Küntay, Pfeiler, Allen & Stoll 2019) and 'bootstraps' learners with memorising and chunking rote-learned forms in the input (Wirén, Björkenstam, Grigonytė & Cortes 2016, Laalo & Argus 2020). This highlights the potential to simulate different acquisition trajectories cross-linguistically, raising several potential avenues to extend the fine-grained curriculum learning strategies of Salhan et al. (2024) to simulate bilingual, heritage language and second-language acquisition scenarios beyond English.

6.3 Conclusions

Roberts et al. (2023) strongly argues against the hypothesis that LLMs contribute anything meaningfully to linguist theory: 'ML systems popular in the current AI spring are *weak AI* – brute force systems laboriously trained to 'unthinkingly' associate patterns in the input data to produce outputs that approximate those data in a process with no resemblance to human cognition (thus betraying Turing's original vision for AI)'. However, small cognitively-inspired language models (BabyLMs) should not suffer from the same issues *a priori*. Modelling Linguistic Competence, rather than performance, in a manner that is compatible with the ontology of mainstream Chomskyan approaches in theoretical linguistics is a relatively neglected focus of modern NLP (Salhan 2023). Although traditionally some theoretical linguists maintain a Platonic view of their models, incorporating more diverse data beyond the traditional internal/external evidence delineation could potentially allow theoretical linguists to delimit evidence for more unified models in Linguistics, as highlighted in Section 4. 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.

However, drawing inferences from BabyLMs that meaningfully influence linguistic theory is not straightforward and 'top-down' linguistically-motivated goals in Language Modelling are necessary. The paradigm introduces additional noisiness through the uninterpretability of backbone architectures, and there are challenges with anthropocentrism. Despite these issues, the paradigm is empirically progressive - incentivising careful linguistic research cross-linguistically that systematically address systematic points of departure between *in vivo* and *in silico* learners. *Pace* Roberts et al. (2023), the paradigm incentivises the development and conception of AI that are not just 'brute force systems' – and potentially more aligned with 'the strong-anthropic-AI [that] Turing envisioned' – while not simultaneously 'throwing the baby out with the bathwater' to capitalise on at least some of the ingredients behind the success of Transformer-based LLMs.

Small cognitively-inspired language models incentivise research on simulating and testing hypotheses from language acquisition across various environments for 'grammar construction' and analysing the potential and the limits of emergentist hypotheses across morphology, phonology and syntax.

References

Abend, O., T. Kwiatkowski, N. J. Smith, S. Goldwater & M. Steedman. 2017. Bootstrapping language acquisition. *Cognition* 164. 116–143.

Ambridge, B. & L. Blything. 2024. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics* 50(1-2). 33–48.

Arnett, C., T. A. Chang, J. A. Michaelov & B. K. Bergen. 2025. On the acquisition of shared grammatical representations in bilingual language models. *arXiv preprint* arXiv:2503.03962.

Arora, A., D. Jurafsky & C. Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In L.-W. Ku, A. Martins & V. Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14638–14663.

Baker, J. 2024. Large language models and linguistic theory: review and reflections. *SyntaxLab [5 November 2024]* .

Baroni, M. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. https://arxiv.org/abs/2106.08694.

Beinborn, L. & N. Hollenstein. 2023. *Cognitive plausibility in natural language processing*. Springer Nature.

Beinborn, L. & Y. Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In H. Bouamor, J. Pino & K. Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4478–4486.

Bender, E. M., D. Flickinger & S. Oepen. 2008. Grammar engineering for linguistic hypothesis testing. In *Proceedings of the Texas Linguistics Society X Conference: Computational linguistics for less-studied languages*, 16–36.

Biberauer, T. 2011. In defence of lexico-centric parametric variation: Two 3rd factor-constrained case studies. *Paper presented at the Workshop on Formal Grammar and Syntactic. Variation: Rethinking Parameters (Madrid)*. .

Biberauer, T. 2019. Children always go beyond the input: The maximise minimal means perspective. *Theoretical Linguistics* 45(3-4). 211–224.

Biberauer, T. & I. Roberts. 2015. Rethinking formal hierarchies: A proposed unification. *Cambridge Occasional Papers in Linguistics* 7. 1–31.

Bjerva, J., B. Plank & J. Bos. 2016. Semantic tagging with deep residual networks. In Y. Matsumoto & R. Prasad (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3531–3541.

Bosch, N. 2023. Emergent syntax and maturation: a neo-emergentist approach to development.

Bosch, N. 2024. On another topic, how do acquisition orders vary? the left periphery and topicalisation in bilinguals. *SyntaxLab [7 May 2024]* .

Boyle, A. 2024. Disagreement & classification in comparative cognitive science. *Noûs* 58(3). 825–847.

Buckner, C. 2013. Morgan's canon, meet hume's dictum: avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy* 28. 853–871.

Bunzeck, B., D. Duran, L. Schade & S. Zarrieß. 2024. Graphemes vs. phonemes: battling it out in character-based language models. In M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt & E. G. Wilcox (eds.), *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, 54–64.

Buttery, P. 2004. A quantitative evaluation of naturalistic models of language acquisition: the efficiency of the triggering learning algorithm compared to a categorial grammar learner. In *Proceedings of the First Workshop on Psychocomputational Models of Human Language Acquisition*, 1–6.

Buttery, P. J. 2006. Computational models for first language acquisition. Tech. Rep. UCAM-CL-TR-675 University of Cambridge, Computer Laboratory.

Calabrese, A. 1995. A constraint-based theory of phonological markedness and simplification procedures. *Linguistic Inquiry* 373–463.

Chang, T. A., Z. Tu & B. K. Bergen. 2024. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics* 12. 1346–1362.

Charpentier, L., L. Choshen, R. Cotterell, M. O. Gul, M. Hu, J. Jumelet, T. Linzen, J. Liu, A. Mueller, C. Ross et al. 2025. BabyLM Turns 3: Call for papers for the 2025 BabyLM workshop. *arXiv preprint* arXiv:2502.10645.

Charpentier, L. G. G. & D. Samuel. 2024. GPT or BERT: why not both? *arXiv preprint* arXiv:2410.24159.

Chemla, E. & R. M. Nefdt. 2024. No such thing as a general learner: Language models and their dual optimization. *arXiv preprint* arXiv:2408.09544.

Choenni, R., D. Garrette & E. Shutova. 2022. Data-efficient cross-lingual transfer with language-specific subnetworks. *arXiv preprint* arXiv:2211.00106.

Chomsky, N. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. 1986. *Knowledge of language*. New York: Praeger.

Chomsky, N., Á. J. Gallego & D. Ott. 2019. Generative grammar and the faculty of language: Insights, questions, and challenges. *Catalan Journal of Linguistics* 229–261.

Choshen, L., G. Hacohen, D. Weinshall & O. Abend. 2022. The grammar-learning trajectories of neural language models. In S. Muresan, P. Nakov & A. Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8281–8297.

Chung, W., J. Hong, S. Salhan, J. Kim, R. Diehl Martinez, J. Thorne & P. Buttery. 2025. Controlling the geometry of token embeddings in language models. *In Preparation* .

Clark, C., B.-D. Oh & W. Schuler. 2025. Linear recency bias during training improves transformers' fit to reading times. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio & S. Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, 7735–7747.

Cuskley, C., R. Woods & M. Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind* 8. 1058–1083.

Dentella, V., F. Guenther & E. Leivada. 2024. Language in vivo vs. in silico: Size matters but larger language models still do not comprehend language on a par with humans. *arXiv preprint* arXiv:2404.14883.

Devlin, J., M.-W. Chang, K. Lee & K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Diehl Martinez, R., H. McGovern, Z. Goriely, C. Davis, A. Caines, P. Buttery & L. Beinborn. 2023. CLIMB – curriculum learning for infant-inspired model building. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen & R. Cotterell (eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural*

Language Learning, 112–127.

Dresher, B. E. 2009. *The contrastive hierarchy in phonology* 121. Cambridge University Press.

Emerson, G. 2020. What are the goals of distributional semantics? In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7436–7453.

Evanson, L., Y. Lakretz & J. R. King. 2023. Language acquisition: do children and language models follow similar learning stages? In A. Rogers, J. Boyd-Graber & N. Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, 12205–12218.

Fox, D. & R. Katzir. 2024. Large language models and theoretical linguistics. *Theoretical Linguistics* 50(1-2). 71–76.

Friedmann, N., A. Belletti & L. Rizzi. 2021. Growing trees: The acquisition of the left periphery. *Glossa: a journal of general linguistics* 6(1). 131.

Futrell, R., E. Wilcox, T. Morita, P. Qian, M. Ballesteros & R. Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42.

Gauthier, J., J. Hu, E. Wilcox, P. Qian & R. Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In A. Celikyilmaz & T.-H. Wen (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 70–76.

Georges Gabriel Charpentier, L. & D. Samuel. 2023. Not all layers are equally as important: Every layer counts BERT. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen & R. Cotterell (eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 238–252.

Gibson, E. & K. Wexler. 1994. Triggers. *Linguistic Inquiry* 25. 355–407.

Goriely, Z. & P. Buttery. 2025a. Babylm's first words: Word segmentation as a phonological probing task. *arXiv preprint arXiv:2504.03338* .

Goriely, Z. & P. Buttery. 2025b. IPA CHILDES: Feature-Rich Resources for Cross-Lingual Phonology and Phonemic Language Modeling. *In Preparation for the Conference of Natural Language Learning (CoNLL)* .

Goriely, Z., A. Caines & P. Buttery. 2025a. Word segmentation from transcriptions of child-directed speech using lexical and sub-lexical cues. *Journal of Child Language* 52(1). 1–41.

Goriely, Z., R. Diehl Martinez, A. Caines, P. Buttery & L. Beinborn. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. In M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt & E. G. Wilcox (eds.), *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, 37–53.

Goriely, Z., S. Salhan, P. Lesci, J. Cheng & P. Buttery. 2025b. Bytespan: Information-driven subword tokenisation. *in press* .

Haga, A., A. Fukatsu, M. Oba, A. Bisazza & Y. Oseki. 2024. BabyLM challenge: Exploring the effect of variation sets on language model training efficiency. In M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt & E. G. Wilcox (eds.), *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, 252–261.

Hale, J. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, .

Heim, J. & M. 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.

Hu, M. Y., A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, R. Cotterell, L. Choshen, A. Warstadt & E. G. Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt & E. G. Wilcox (eds.), *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, 1–21.

Huebner, P. A., E. Sulem, F. Cynthia & D. Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. Online: Association for Computational Linguistics.

Huebner, P. A. & J. A. Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In *Psychology of learning and motivation*, vol. 75, 279–331.

Ivanova, A. A., A. Sathe, B. Lipkin, U. Kumar, S. Radkani, T. H. Clark, C. Kauf, J. Hu, R. T. Pramod, G. Grand, V. Paulun, M. Ryskina, E. Akyürek, E. Wilcox, N. Rashid, L. Choshen, R. Levy, E. Fedorenko, J. Tenenbaum & J. Andreas. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. https://arxiv.org/abs/2405.09605.

van Kampen, J. 2010. Typological guidance in the acquisition of V2 Dutch. *Lingua* 120(2). 264–283.

Katzir, R. 2023. Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi (2023). *Manuscript. Tel Aviv University* .

Kuncoro, A. S. 2022. *Scalable syntactic inductive biases for neural language models*: University of Oxford dissertation.

Kwiatkowski, T., S. Goldwater, L. Zettelmoyer & M. Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, .

Laalo, K. & R. Argus. 2020. Linguistic recycling in language acquisition: Child-directed speech and child speech in the study of language acquisition. *AILA Review* 33(1). 86–103.

Lavechin, M., M. De Seyssel, H. Titeux, H. Bredin, G. Wisniewski, A. Cristia & E. Dupoux. 2022. Can statistical learning bootstrap early language acquisition? A modeling investigation. *OSF* .

Liu, Z., O. Kitouni, N. S. Nolte, E. Michaud, M. Tegmark & M. Williams. 2022. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems* 35. 34651–34663.

MacWhinney, B. 2000. *The CHILDES project: The database*, vol. 2. Psychology Press.

Mahon, L., O. Abend, U. Berger, K. Demuth, M. Johnson & M. Steedman. 2025. A language-agnostic model of child language acquisition. *Computer Speech Language* 90. 101714.

Mallory, F. 2024. Generative linguistics and the computational level. *Croatian Journal of Philosophy* 24(71). 195–218.

Marcolli, M., R. C. Berwick & N. Chomsky. 2023a. Old and New Minimalism: a Hopf algebra comparison. *arXiv preprint* arXiv:2306.10270.

Marcolli, M., R. C. Berwick & N. Chomsky. 2023b. Syntax-semantics interface: an algebraic model. https://arxiv.org/abs/2311.06189.

Marcolli, M., R. C. Berwick & N. Chomsky. 2023c. Syntax-semantics interface: an algebraic model. *arXiv preprint* arXiv:2311.06189.

Marcolli, M., N. Chomsky & R. Berwick. 2023d. Mathematical structure of syntactic merge. *arXiv preprint* arXiv:2305.18278.

Marks, S., C. Rager, E. J. Michaud, Y. Belinkov, D. Bau & A. Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint* arXiv:2403.19647.

Mayer, C. 2020. An algorithm for learning phonological classes from distributional similarity. *Phonology* 37(1). 91–131.

McGee, T. A. & I. A. Blank. 2024. Evidence against syntactic encapsulation in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, .

Millière, R. & C. Rathkopf. 2024. Anthropocentric bias and the possibility of artificial cognition. In *ICML 2024 Workshop on LLMs and Cognition*, .

Millière, R. 2024. Language models as models of language. https://arxiv.org/abs/2408.07144.

Misra, K. & K. 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*, 913–929.

Misra, K., J. Rayz & A. Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In A. Vlachos & I. Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2928–2949.

Mita, M., R. Yoshida & Y. Oseki. 2025. Developmentally-plausible working memory shapes a critical period for language acquisition. *arXiv preprint* arXiv:2502.04795.

Mitrofanova, N. 2018. Early underspecification of functional categories: Evidence from the acquisition of locative pps in russian. *Language Acquisition* 25(4).

341–365.

Moran, S., N. A. Lester, H. Gordon, A. Küntay, B. Pfeiler, S. Allen & S. Stoll. 2019. Variation sets in maximally diverse languages. In *Proceedings of the 43rd annual Boston University Conference on Language Development*, 427–440.

Mueller, A., G. Nicolai, P. Petrou-Zeniou, N. Talmina & T. Linzen. 2020. Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5523–5539.

Murty, S., P. Sharma, J. Andreas & C. Manning. 2023. Grokking of hierarchical structure in vanilla transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 439–448.

Murty, S., P. Sharma, J. Andreas & C. D. Manning. 2022. Characterizing intrinsic compositionality in transformers with tree projections. In *The Eleventh International Conference on Learning Representations*, .

Nayeem, M. T. & D. Rafiei. 2024. KidLM: Advancing language models for children – early insights and future directions. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4813–4836.

Nemecek, A. 2023. Coinductive guide to inductive transformer heads. *arXiv preprint* arXiv:2302.01834.

Nevins, A. 2015. Triumphs and limits of the contrastivity-only hypothesis. *Linguistic Variation* 15(1). 41–68.

Oba, M., T. Kuribayashi, H. Ouchi & T. Watanabe. 2023. Second language acquisition of neural language models. In A. Rogers, J. Boyd-Graber & N. Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, 13557–13572.

Oba, M., Y. Oseki, A. Fukatsu, A. Haga, H. Ouchi, T. Watanabe & S. Sugawara. 2024. Can language models induce grammatical knowledge from indirect evidence? *arXiv preprint arXiv:2410.06022* .

Olsson, C., N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish & C. Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread* https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai & S. Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, vol. 32, .

Piantadosi, S. T. 2023. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett* 353–414.

Pierson, S. G. 2024. *The acquisition of verbal morphology in Ayöök: child-directed speech, child language, and learning in the home*: UT Austin dissertation.

Poeppel, D. & K. Wexler. 1993. The Full Competence Hypothesis of Clause Structure in Early German. *Language* 69(1). 1–33.

Ponti, E. 2021. Inductive bias and modular design for sample-efficient neural language learning.

Power, A., Y. Burda, H. Edwards, I. Babuschkin & V. Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint* arXiv:2201.02177.

Press, O., N. Smith & M. Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, .

Prokhorov, V. 2021. Injecting inductive biases into distributed representations of text.

Radford, A. 1990. The syntax of nominal arguments in early child English. *Language Acquisition* 1(3). 195–223.

Rizzi, L. 1993. Some notes on linguistic theory and language development: The case of root infinitives. *Language acquisition* 3(4). 371–393.

Roberts, I., J. Watumull & N. Chomsky. 2023. Universal grammar. In D. A. Vakoch & J. Punske (eds.), *Xenolinguistics: Towards a science of extraterrestrial language*, 165–181. Taylor Francis.

Sakas, W. & J. D. Fodor. 2001. The structural triggers learner. *Language acquisition and learnability* 172–233.

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

Salhan, S., R. Diehl Martinez, Z. Goriely & P. Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt & E. G. Wilcox (eds.), *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, 174–188.

Samuel, D., A. Kutuzov, L. Øvrelid & E. Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In A. Vlachos & I. Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, 1954–1974.

Sartran, L., S. Barrett, A. Kuncoro, M. Stanojević, P. Blunsom & C. Dyer. 2022. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. *Transactions of the Association for Computational Linguistics* 10. 1423–1439.

Schaeffer, R., B. Miranda & S. Koyejo. 2023. Are emergent abilities of large language models a mirage? *arXiv preprint* arXiv:2304.15004.

Schrimpf, M., I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum & E. Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* 118(45). e2105646118.

Schuler, K. D., C. Yang & E. L. Newport. 2016. Testing the tolerance principle: Children form productive rules when it is morecomputationally efficient to do so. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 38, .

Siskind, J. M. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61. 39–91.

Someya, T. & Y. Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In A. Vlachos & I. Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, 1581–1594.

Song, Y., K. Krishna, R. Bhatt & M. Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In Y. Goldberg, Z. Kozareva & Y. Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4606–4634.

Steuer, J., M. Mosbach & D. Klakow. 2023. Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen & R. Cotterell (eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 142–157.

Szubert, I., O. Abend, N. Schneider et al. 2024. Cross-linguistically consistent semantic and syntactic annotation of child-directed speech. *Language Resources Evaluation* .

Tan, A., C. Yu, B. Long, W. Ma, T. Murray, R. Silverman, J. Yeatman & M. C. Frank. 2024. Devbench: A multimodal developmental benchmark for language learning. *Advances in Neural Information Processing Systems* 37. 77445–77467.

Thrush, T., R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela & C. Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Tigges, C., M. Hanna, Q. Yu & S. Biderman. 2024. LLM circuit analyses are consistent across training and scale. *arXiv preprint* arXiv:2407.10827.

Timiryasov, I. & J.-L. Tastet. 2023. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 279–289.

Tsimpli, I. M. 2005. Peripheral positions in early Greek. In M. Stavrou & A. R. Terzi (eds.), *Advances in Greek Generative Syntax: In honor of Dimitra Theophanopoulou-Kontou*, 179–216. Amsterdam: John Benjamins.

Ueda, R., T. Kuribayashi, S. Kando & K. Inui. 2025. Syntactic learnability of echo state neural language models at scale. https://arxiv.org/abs/2503.01724.

Unnsteinsson, E. 2020. Compositionality and expressive power: Comments on Pietroski. *Croatian Journal of Philosophy* 20(60). 295–310.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser & I. Polosukhin. 2017. Attention is All You Need. *Advances in neural information processing systems* 30.

Villavicencio, A. 2002. The acquisition of a unification-based generalised categorial grammar. Tech. Rep. UCAM-CL-TR-533 University of Cambridge, Computer Laboratory.

Villavicencio, A. 2011. Language acquisition with a unification-based grammar. In R. Borsley & K. Borjars (eds.), *Non-transformational syntax*, Blackwell.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy & S. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* 32.

Wang, B., X. Yue, Y. Su & H. Sun. 2024. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. *Advances in Neural Information Processing Systems* 37. 95238–95265.

Warner, B., A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint* arXiv:2412.13663.

Warstadt, A., A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen & R. Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen & R. Cotterell (eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 1–34.

Warstadt, A., A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang & S. R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8. 377–392.

Wirén, M., K. N. Björkenstam, G. Grigonytė & E. E. Cortes. 2016. Longitudinal studies of variation sets in child-directed speech. In *Proceedings of the 7th workshop on cognitive aspects of computational language learning*, 44–52.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest & A. Rush. 2020. Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Xia, M., M. Artetxe, C. Zhou, X. V. Lin, R. Pasunuru, D. Chen, L. Zettlemoyer & V. Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13711–13738.

Xiang, B., C. Yang, Y. Li, A. Warstadt & K. Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In P. Merlo, J. Tiedemann & R. Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2784–2790.

Xu, R., F. Luo, B. Chang, S. Huang & F. Huang. 2022. S4-tuning: A simple cross-lingual sub-network tuning method-tuning: A simple cross-lingual sub-

network tuning method. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 530–537.

Yang, C. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, C. 2018. Some consequences of the tolerance principle. *Linguistic Approaches to Bilingualism* 8(6). 797–809.

Yang, Q., P. Wang, L. D. Plonsky, F. L. Oswald & H. Chen. 2024. From babbling to fluency: Evaluating the evolution of language models in terms of human language acquisition. *arXiv preprint* arXiv:2410.13259.

Yedetore, A., T. Linzen, R. Frank & R. T. McCoy. 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In A. Rogers, J. Boyd-Graber & N. Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9370–9393.

Ziv, I., N. Lan, E. Chemla & R. Katzir. 2025. Large language models as proxies for theories of human linguistic cognition. *arXiv preprint* arXiv:2502.07687.

Appendices

Language	IPA CHILDES	Words	Phonemes	% Child
EnglishNA	EnglishNA (49)	9,993,744	30,986,218	35.83
EnglishUK	EnglishUK (16)	7,147,541	21,589,842	39.00
German	German (10)	5,825,166	21,442,576	43.61
Japanese	Japanese (11)	2,970,674	11,985,729	44.20
Indonesian	Indonesian (1)	2,347,642	9,370,983	34.32
French	French (15)	2,973,318	8,203,649	40.07
Spanish	Spanish (18)	2,183,992	7,742,550	45.93
Mandarin	Mandarin (16)	2,264,518	6,605,913	38.89
Dutch	Dutch (5)	1,475,174	4,786,803	35.08
Polish	Polish (2)	1,042,841	4,361,797	63.26
Serbian	Serbian (1)	1,052,337	3,841,600	29.14
Estonian	Estonian (9)	843,189	3,429,228	44.71
Welsh	Welsh (2)	666,350	1,939,286	69.18
Cantonese	Cantonese (2)	777,997	1,864,771	33.54
Swedish	Swedish (3)	581,451	1,782,692	44.63
PortuguesePt	PortuguesePt (4)	499,522	1,538,408	39.47
Korean	Korean (3)	263,030	1,345,276	36.76
Italian	Italian (5)	352,861	1,309,489	39.02
Croatian	Croatian (1)	305,112	1,109,696	39.24
Catalan	Catalan (6)	319,726	1,084,594	36.49
Icelandic	Icelandic (2)	279,939	1,057,235	35.21
Basque	Basque (2)	230,500	942,725	48.82
Hungarian	Hungarian (3)	237,062	918,002	47.95
Danish	Danish (1)	275,170	824,314	41.71
Norwegian	Norwegian (2)	227,856	729,649	42.58
PortugueseBr	PortugueseBr (2)	174,845	577,865	44.42
Romanian	Romanian (3)	152,465	537,669	42.62
Turkish	Turkish (2)	79,404	421,129	50.58
Irish	Irish (2)	105,867	338,425	34.37
Quechua	Quechua (2)	46,848	281,478	40.06
Farsi	Farsi (2)	43,432	178,523	40.45

Table 1 A breakdown of each language available in IPA CHILDES (Goriely & Buttery 2025b) and MAO-CHILDES (Salhan et al. 2024). The bracketed number in the **CHILDES Collection column** refers to the number of corpora downloaded from that collection. The **Words** and **Phonemes** columns refer to the number of words and tokens in each subset and **% Child** refers to the percentage of the data that is spoken by a child.

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 2 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. These units are ordered to implement acquisition-inspired strategies of Salhan et al. (2024)

We conduct our experiments using the `PyTorch` framework (Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga, Desmaison, Kopf, Yang, DeVito, Raison, Tejani, Chilamkurthy, Steiner, Fang, Bai & Chintala 2019) and the `Transformers` library (Wolf, Debut, Sanh, Chaumond, Delangue, Moi, Cistac, Rault, Louf, Funtowicz, Davison, Shleifer, von Platen, Ma, Jernite, Plu, Xu, Le Scao, Gugger, Drame, Lhoest & Rush 2020) using a server with one NVIDIA A100 80GB PCIe GPU, 32 CPUs, and 32 GB of RAM for all experiments.

Suchir Salhan
University of Cambridge
sas245@cam.ac.uk