

Exploring Linguistic Parameters in LMs

Cristiano Chesi, Achille Fusco, Albina Mukusheva, Guido Formichi,
Sarah Rossi, Hanna B. Cygan, Tommaso Sgrizzi, Asya Zanollo
NeTS Lab @ IUSS Pavia, Piazza della Vittoria 15, Pavia, Italy

Correspondence: cristiano.chesi@iusspavia.it

Abstract

This work places the BabyLM Multilingual perspective within the Principles and Parameters framework (Chomsky, 1981; Yang et al., 2017). First, we compared tokenization and architectural strategies that make the incremental nature of language acquisition and processing more or less transparent (recurrent versus attention-based architectures). We then assessed our trained models with respect to specific linguistic generalizations, including morphological decomposition, learning trajectories, and binary parametric options. We found that (i) morphologically informed tokenizers yield consistent benefits in both learning trajectories and performance on downstream minimal-pair tasks; (ii) recurrent architectures are substantially more data-efficient, converging earlier and achieving slightly better performance on downstream tasks; and (iii) the parameter-setting objective yields modest improvements, though not across all expected linguistic domains.

1 Introduction

In the context of Multilingual Language Modeling, a single Language Model (LM) must internally represent the properties of different languages. Adopting a perspective of cognitive plausibility, this means that the same architecture must be able not only to represent *in principle* any grammar, but also to accommodate multiple grammars simultaneously and thus successfully handle points of variation among them. In the generative linguistics tradition, a crucial notion in this respect is that of *parameter*, which differs from that assumed in language modeling. In the Principles and Parameters (P&P) framework (Chomsky, 1981; Yang et al., 2017), a *linguistic parameter* (LiP) is conceived as a categorical switch between two structurally possible options compatible with both *universal principles* and the available *linguistic evidence*. For the three languages under investigation (ENG - En-

glish, NLD - Dutch, and ZHO - Chinese), several relevant LiPs can be defined. Following Roberts (2019), a LiP can be located at the macro, meso, or micro level, depending on the affected heads and the number of lower-order options it entails. Simplifying (see § 2 for details), *Head-directionality* can be considered a macro-parameter with clear effects on word order. This can be observed by contrasting the Dutch sentence *Je kunt er toch [o ééntje] [v maken]* with its English counterpart *You can just [v make] [o one]*. The VO order is reversed as a consequence of a minimal structural parameter distinguishing English from Dutch. In this work, we explore this parametric intuition: beyond the structural machinery that, in a pre-theoretical sense, merges one token with another, the model should also internalize the higher-level generalizations captured by linguistic parameters (Chesi, 2023). The underlying assumption is that the computational machinery processing ENG, NLD, and ZHO is identical across languages, corresponding to a shared *Universal Grammar* (Chomsky, 1981) backbone, and that only a small number of parameters need to be set on top of this machinery to systematically alter the model’s expectations. In the example above, the head-directionality LiP setting determines whether the model expects a direct object to precede or follow its predicate. Building on this intuition, we approach language modeling by introducing an additional training objective: parametric discrimination across languages (LiP discovery). The architectural variable we focus on is *processing incrementality*. Incremental processing is naturally aligned with recurrent approaches (Gu and Dao, 2024; De et al., 2024), but less so with attention-based models, particularly those relying on key-value caching, such as the BabyLM baselines. A further unavoidable challenge concerns the choice of linguistic units for modeling linguistically plausible parametric distinctions. The multilingual track of the BabyLM 2026 Challenge is

particularly demanding, as the inclusion of Chinese forces us to abandon morphologically grounded, split-based segmentation procedures (Fusco et al., 2025) and instead adopt either a morphologically agnostic BPE/Unigram (bottom-up) approach or a boundary-discovery approach that treats spacing only as a soft segmentation cue.

In the following sections, we briefly explore the notion of LiPs (§ 2) and issues related to tokenization (§ 3). We then discuss specific architectural choices (§ 4) that allow us to implement supervised or unsupervised control of LiPs. In § 5, we present experiments comparing the performance of different tokenizers using the attention-based GPT-2 model, and compare this BabyLM baseline with Recurrent Neural Networks (RNNs) from the Hawk family (De et al., 2024). We also discuss various attempts to integrate parametric switches into a theory-driven computational architecture. The main results (§ 6) indicate that RNNs are more budget-efficient and slightly outperform attention-based baseline models on certain downstream evaluation tasks selected for the BabyLM Multilingual Challenge. Moreover, we find that tokenization strategy matters, with morphologically consistent approaches yielding better results on agreement phenomena. Overall, LiP supervision has only a modest effect on overall model performance, but its impact is more pronounced for specific phenomena, such as islands (§ 7).

2 Linguistic Parameters

A common criticism in linguistics is that Large Language Models (LLMs) flatten sentence structure into string-based representations, whereas humans are biased toward representing sentences as hierarchical phrase structures (Kodner et al., 2023). A related but poorly understood issue is how multilingual LMs represent cross-linguistic variation. The fact that the so-called *curse of multilinguality*—namely, the observation that, beyond a certain point, adding multilingual data appears to degrade model performance—can be mitigated by increasing model size or capacity (Conneau et al., 2020) casts further doubt on whether LLMs learn the relevant linguistic generalizations. Although the P&P framework developed in the 1980s has been substantially reformulated within the Minimalist perspective (Roberts, 2019), one of its enduring ideas is that linguistic variation is constrained: alongside a set of universal principles, each adult grammar

consistently selects a particular configuration of LiPs that may favor one word order over another. Simplifying a long-standing debate, we can identify two aspects common to linguistic theories of parameters: (i) the binary nature of parameters and (ii) the implicational nature of LiPs. Parameters are conceived as a finite set of binary options that describe the structural properties of a language, each minimally associated with the presence or absence of a functional feature that licenses particular configurations. For the child, language acquisition consists in setting the relevant parameters on the basis of the input (primary linguistic data) during a critical period (Chomsky, 1986; Guasti, 2017). More recent proposals argue that parameters emerge from the interaction among minimal universal constraints, economy principles (the so-called *third factor*), and input data (Roberts, 2019). We now consider three representative examples of classical LiPs examined in this work: *Head Directionality* (HD), *Verb Second* (V2), and *Filler–Gap Dependency* (WH).

HD parameter Assuming that a *head* is the lexical item that encodes the selectional properties of a constituent (e.g., the lexical verb, V), heads merge with their obligatory complements (e.g., the direct object (O), *a sandwich*, in $[_{VP} \text{ eat}_V [_O \text{ a sandwich}]]$). The HD parameter determines the canonical linear order of a head and its complement within a constituent. In its positive setting, the parameter yields a head-initial order (Head–Complement); in its negative setting, it yields a head-final order (Complement–Head).¹ Taking the verb phrase of a transitive verb as an illustrative case, both ENG and ZHO consistently exhibit VO order, as discussed above. NLD, by contrast, displays a split pattern, also found in other Germanic languages: word order is sensitive to clause type, with matrix clauses exhibiting VO order and embedded clauses exhibiting OV order.

V2 parameter The V2 parameter determines the presence of the so-called verb-second phenomenon, which is widespread among Germanic languages.

¹Here, we greatly simplify the linguistic analysis for the sake of clarity (see [Limitations](#) section). According to Biberauer and Roberts (2015), HD is determined by a hierarchy of binary parametric choices: (i) Is head-final present? If not, L is head-initial; if yes: (ii) is head-final for all heads? If yes, L is head-final; if not, (iii) is it present in all +V heads? If not, (iv) a subset of +V heads it affected? and so on. In several influential works, HD is recast in terms of (roll-up) movement (Kayne, 2018), which, however, must be justified by specific features.

Under V2, the finite verb in a declarative sentence surfaces in the second position from the left edge of the clause, immediately after the first constituent (Adger, 2003). NLD exhibits this phenomenon: in *Gisteren heeft Jan haar gekust* (lit. ‘Yesterday has John her kissed’), the inflected auxiliary HAVE must appear in second position (Zwart, 1997).

WH parameter As a final example, the WH parameter concerns long-distance dependencies. In languages such as ENG, a wh-item must be interpreted in a thematic position that may be arbitrarily far from its surface position, as in *What did you eat _{what}?* This follows from the assumption that thematic relations are assigned in a fixed configuration (Hale and Keyser, 1993, 2002); therefore, V and O must be merged together at some point in the derivation. Languages such as ZHO exhibit the opposite pattern: the wh-item remains in its thematic position, yielding structures such as *You eat what?*²

Table 1 summarizes a relevant set of classical clausal parameters (LiPs) that distinguish the three languages under analysis and indicates their observable effects on word order.

Parameter	Word-order effects	ENG	NLD	ZHO
Head Directionality (HD)	VO vs. OV	+	+/-	+
Verb-Second (V2)	V2 vs non-V2	-	+	-
Filler-Gap (WH)	null vs overt wh-mov.	+	+	-
Pro-drop (P0)	null vs overt subject	-	-	+
Tense morphology (T0)	null vs overt tense marking	+	+	-
Subject Verb Agreement (SV)	null vs overt agreement on V	+	+	-

Table 1: Clausal parameters (LiPs) across ENG, NLD, ZHO. + = positive setting; - = negative setting. The parameters selected for this study are in light-gray rows.

3 Tokenization issues

Standard tokenization approaches adopt four stances that are systematically implausible from a child-ecological perspective:

- (a) *Representational opacity*. Once a span has been assigned a token ID, the model has no direct access to its internal (phonotactic/orthographic) structure.
- (b) *Context independence*. A string is segmented identically regardless of its surrounding syntactic, semantic, or discourse context.
- (c) *Compression objective*. The segmentation objective optimizes compression or unigram like-

²This is, again, a linguistic simplification, since interrogative force must nonetheless be interpreted in the relevant left-peripheral position: *what you eat what?* (Cheng, 2009).

lihood. Goldman et al. (2024) suggest that this is actually already part of the language modeling (predictive) effort. However, combined with (a), it entails that sufficiently frequent related forms, e.g., *cat* and *cats*, may be represented as unrelated atomic symbols, leaving their systematic relation to be recovered from distributional evidence during subsequent training (see Fusco et al. 2025).

- (d) *Frozen learning*. The segmentation is computed once and never revised in light of subsequent structural evidence (no morphological re-analysis).

Tokenization is thus a logical bottleneck in language modeling and BPE’s merge table (Sennrich et al., 2016) or the Unigram algorithm (Kudo and Richardson, 2018) are prototypical examples that align perfectly with these anti-ecological perspectives. A way to (partially) dissolve the constraints defined above is to make the boundaries fall where predictability breaks down and systematic variation appear. Harris (1955) located morpheme boundaries at points of high branching factor (low transitional probability), Elman (1990) showed recurrent prediction error spiking at word edges, while Saffran et al. (1999) demonstrated infants segmenting on transitional-probability dips.

Following on this line of work, Fusco et al. (2025) observed that split-based, morphologically robust tokenizers (like *ParadigmFinder* and *MorPiece*) performs significantly better than merge-based tokenizers (like *BPE*) under morphological evaluation (Sigmorphon 2022), though this superiority does not uniformly translate into a better models’ performance, for instance on BLiMP tasks. ZHO language forced us to rephrase the split-based idea, since the original spacing hard-cue can hardly be used in Chinese.

3.1 MorPiece (MoP)

To align positions (a)-(d) to children’s experience, a *boundary discovery* option is introduced into MorPiece (MoP) training (Chesi, 2026). Word spacing is demoted to a soft cue, while strong punctuation serves as the only reliable cue for segmenting initial candidates. A curriculum learning approach is adopted, whereby shorter pieces are processed first to strengthen segmentation evidence in accordance with the *Tolerance/Sufficiency Principle (TP)* (Yang, 2004). Child-directed speech is

also prioritized over other sources (e.g., subtitles or Wikipedia) and is processed first. The basic idea behind MoP is to traverse each input character sequence and populate two *tries* (Fredkin, 1960): one consisting of the left-to-right sequence (*root-trie*) and the other of the reversed sequence (*infl-trie*). A split is postulated whenever TP is satisfied in both the left-to-right and right-to-left directions at a specific node. For example, "wugs" is split into "wug" and "s" if TP applies both ways, between "g" and "s" (i.e., in the root-trie: $\text{freq}(\text{node}_s) > (\text{freq}(\text{node}_g)/\ln(\text{freq}(\text{node}_g)))$) and between "s" and "g" (i.e., in the infl-trie: $\text{freq}(\text{node}_g) > (\text{freq}(\text{node}_s)/\ln(\text{freq}(\text{node}_s)))$). Each language in the byte-premium balanced multilingual corpus (ENG, NLD, ZHO) was trained independently, and its token contribution was capped at 16K. After vocabulary combination and productivity optimization (suffix fertility > 3 , minimum frequency = 10), the final multilingual vocabulary contained 39K tokens. Finally, the MoP-trained tokenizer is evaluated for order of acquisition (--ooa option) in English by inspecting productive vocabulary growth every 440k syllables of exposure (roughly three months of English language input). The resulting morpheme acquisition order shows a positive correlation with the developmental gold-standard ordering (Brown, 1973; De Villiers and Roeper, 2011): $\tau_b = .485$, $p = .031$ (two-tailed, permutation test), $N = 12$. Because frequency was itself negatively related to the gold ordering ($\tau_b = -.454$), we computed a frequency-controlled partial correlation, which remained positive, $\tau_{\text{partial}} = .231$ (Figure 1). Note that morpheme splitting sometimes overgeneralizes, and the first item selected in the graph may simply correspond to an incorrectly segmented word (e.g., *moth-er*; although unrelated to the comparative marker, this constitutes an interesting generalization worth reporting).

3.2 ParadigmFinder (PaF)

For this multilingual setting, we also extend the unsupervised ParadigmFinder (PaF) tokenizer first described by Fusco et al. (2025). Building on the paradigm-based approaches of Goldsmith (2001) and Xu et al. (2018), PaF identifies recurring root-suffix combinations in the training vocabulary and organizes them into paradigms, each comprising a set of roots and a corresponding set of suffixes that co-occur with systematic regularity. Paradigm induction begins by enumerating the possible binary splits of each word into candidate roots and suffixes

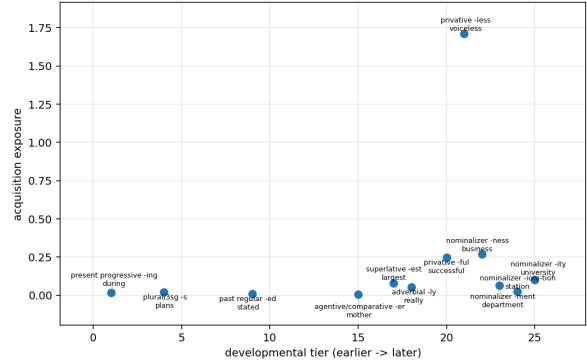


Figure 1: MoP Order of Acquisition. Training parameters: min. frequency for TP application 50; min. token freq. 10; min. suffix productivity 3; vocabulary productivity recorded every 440K syllables of exposure (~ 3 months of English exposure).

and grouping roots that occur with identical suffix sets. Candidate paradigms are subsequently normalized by factoring out prefixes shared by all their suffixes and transferring this material to the roots. This operation places segmentation boundaries at points supported by actual variation within the paradigm and prevents spurious analyses caused by root-final characters being treated as parts of suffixes. The resulting paradigms are then expanded according to TP: paradigms with smaller suffix inventories can be merged into more extensive ones when the overlap between their root sets provides sufficient evidence that the additional suffixes represent a systematic generalization rather than an accidental co-occurrence. PaF also considers nested suffixation when a word admits multiple compatible analyses. For example, if *singer* and *singers* support both the analyses *singer-Ø/singer-s* and *sing-er/sing-ers*, the relation between *-er* and *-ers* can itself be analyzed through the productive $\emptyset$, *-s* paradigm. Redundant and subsumed paradigms are subsequently pruned, and the remaining paradigms are ranked by a support score that rewards both the number of participating roots and the number of associated suffixes. Words for which no productive analysis is supported are assigned to a residual paradigm, thereby avoiding unsupported segmentations. At tokenization time, words are first matched against the induced paradigms and segmented into the corresponding root and suffix units; when no complete paradigm match is available, a longest-known-suffix strategy is used to recover partial structure.

PaF was trained on the ENG, NLD, and ZHO por-

tions of the multilingual training corpus. Paradigm induction was carried out independently for each language so that morphological regularities were inferred from language-internal evidence and patterns from one language could not influence the segmentation of another. This produced separate language-specific inventories of paradigmatic roots, suffixes, and residual units, which were subsequently combined into a single multilingual vocabulary. The vocabulary contribution of each language was capped at 16K tokens as in MoP. After the three inventories were combined, training yielded a final vocabulary of approximately 44K tokens.

4 Architectural alternatives

One of our main objectives was to evaluate whether attention-free models (no KV cache) better capture structural dependencies cross-linguistically, confirming other monolingual studies (Wilcox et al., 2023). Because memory in an RNN-based architecture is entirely encoded in the recurrent mechanism, specific wiring choices may provide varying degrees of support for the linguistic variation we aim to probe.

4.1 Attention vs. Recurrent heads

Building on De et al. (2024), we adapted the Hawk model by introducing a temporal-mixing block between two residual blocks, followed by an MLP block, and compared this architecture with the GPT-2 baseline. We opted for a Real-Gated Linear Recurrent Unit (RG-LRU) (Orvieto et al., 2023), matching the Hawk architecture, without the interleaved local attention used in the more performant Griffin variant. The recurrence is the diagonal parallelized affine scan $h_t = a_t \odot h_{t-1} + b_t$ (a $\log_2 T$ -depth Hillis–Steele pass replaces the T -step loop, exact to 6×10^{-8}). The recurrent block is a leaner block compared to the attention-based block, moreover the sequence mixing cost is linear with respect to the sequence length in Hawk, quadratic in GPT-2. Table 2 summarizes the relevant differences between the two architectures. According to De et al. (2024), Hawk model exceeds standard Transformers performance, and also more efficient RNN models based on Selective State Spaces such as Mamba (Gu and Dao, 2024), on downstream tasks when trained on half as many token.

Feature	GPT-2	Hawk
Temporal mixing	Multi-head causal self-attention	Real-Gated Linear Recurrent Unit (RG-LRU)
Sequence mixing cost	$O(n^2)$ in sequence length (attention matrix)	$O(n)$ — linear recurrence, fixed-size state
State at inference	Grows with context (KV-cache)	Fixed-size hidden state, independent of context length
Positional info	Learned absolute position embeddings	None needed — recurrence is inherently ordered
Long-range dependency	Direct token-to-token attention weights	Gated accumulation/decay through the recur. state
Training parallelism	Native (attention is parallel)	Associative-scan to parallelize recurrence
Params for depth/width	Attention QKVO projections + MLP	Recurrent block (RG-LRU gates + projections) + MLP

Table 2: Comparison Between GPT-2 and Hawk Architectures.

5 Materials and methods

5.1 Tokenizers evaluation

Tokenizers performance is assessed first based on morphological independent Sigmorphon 2022-like tests. Sigmorphon 2022 benchmark (Batsuren et al., 2022) has been directly adapted for ENG, while NLD and ZHO morphological benchmarks are automatically generated from on-line resources freely available: NLD is based on UniMorph dataset (inflections only); ZHO is based on the morphological annotation in public Childes datasets. We also evaluated the overall performance of each tokenizer directly on morphologically relevant minimal pairs decision tasks. For the comparison we trained each tokenizer on the byte-premium paired clean multilingual corpus of 100M tokens.³ We focused only on local and non-local agreement phenomena in minimal pairs tests⁴ using the same GPT-2 model (BabyLM Multilingual baseline). We compared BPE, UniGram and MoP tokenizers initially with the same 16K (max) per language vocabulary size.

5.2 Linguistic Parameters (LiP) setup

Two parameterization injection strategies are considered:

³All text was lowercased and any metalinguistic information (e.g., speaker tiers) was removed.

⁴BLiMP-filtered: anaphor_gender_agreement, anaphor_number_agreement, determiner_noun_agreement_*, distractor_agreement_*, irregular_plural_subject_verb_agreement_*, regular_plural_subject_verb_agreement_*, BLiMP-NL: anaphor_agreement; BLiMPZHO: anaphor_gender_agreement, anaphor_number_agreement, classifier_noun_agreement_*

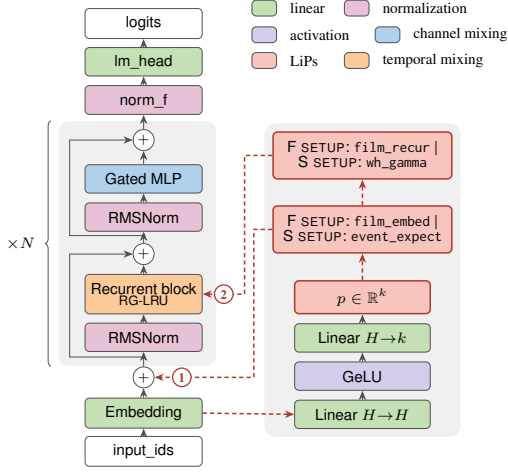


Figure 2: Hawk with linguistic-parameter (LiP) control. The compute path is shown on the left; the two alternative parameterization modes and their injection points are shown on the right. See Figure 3 in Appendix A.2 for more details.

F setup a *Feature-wise Linear Modulation* (FiLM) (Perez et al., 2017) is implemented as a router that must infer k soft bits from the input and conditions the backbone diffusely, supervised toward LiPs gold via Binary Cross-Entropy (BCE).

S setup a purely *switch*-based parameterization, where three named single-site switches on LiPs are coded and supervised.

For the current experiments we will concentrate on the three LiPs discussed in Section 2: *HD*, *V2*, and *WH*. The choice is linguistically justified because their scope is different: *HD* hinges directly on the local merge mechanism, *V2* requires the accommodation of a clause-bounded tense re-ordering, while the *WH* parameter might require a cyclic long-distance dependency spanning various clauses. The injection strategy is illustrated in Figure 2 (See Appendix A.2 for technical details).

5.3 Hawk model variants

To assess the locus of potentially linguistically relevant generalizations, we considered minimal variations of a baseline Hawk model, with and without LiPs injection. Relative to the GPT-2 baseline models, the capacity modifications were designed to test the following hypotheses:

Hawk-A Increasing the model depth (number of layers) and the recurrent state size (width of the recurrent units) at the same parameter budget as the baseline might improve both hierarchical

Setup	GPT-2	base	A	B	C
Parameters (M)	12	115	112	142	115
Layers	12	12	18	12	12
Embedding	768	704	704	704	704
Recurrent width	–	768	864	1152	768
Kernels	–	4	4	4	8
MLPs	–	3	2	3	3

Table 3: Comparison Between GPT-2 and Hawk different setups.

sensitivity (through greater depth) and information retention (through wider recurrent units).

Hawk-B A facilitation in non-local LiPs channeling (e.g., longer retention of WH dependencies) might be obtained by increasing the recurrent state size (+50%: from 768 to 1152).

Hawk-C A facilitation in local LiPs channeling (e.g., HD and local agreement phenomena such as SV agreement in the absence of local interveners) could be obtained by doubling the convolutional kernel count (from 4 to 8).

All models are trained as a single multilingual causal LM over ENG, NLD, and ZHO on the BabyLM 2026 budget (~ 100 M byte-premium-adjusted English-equivalent words), 10 epochs, sequence length 512, using a byte-fallback MoP tokenizer (vocab 39,953, see § A.1 for discussion on the impact of byte-fallback). Optimization: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0), peak LR $5e-4$ with 1% linear warmup and cosine decay to $5e-5$, grad-clip 1.0, *bf16* mixed precision, RMSNorm (instead on LayerNorm) and tied input/output embeddings. Regarding "grammatical awareness", it is worth emphasizing that, in the baseline Hawk setup, only 28.1M parameters are allocated to the embedding layer, while the remaining 87.3M are non-embedding parameters. Originally, we adopted a simple sequential recurrence and subsequently implemented Hawk’s native parallel (diagonal) scan. This dramatically reduced the training time (about 5% of our naive model’s implementation).

6 Results

6.1 Tokenizers evaluation

The SIGMORPHON-based morphological evaluation of each tokenizer is reported in Table 4, with the monolingual baseline results from Fusco et al. (2025) shown in gray. All multilingual tokenizers exhibit a substantial decline relative to their

Tokenizer	Avg. Lev. Dist.	Prec	Rec	F1
BPE	2.08	21.03	26.62	23.50
MoP	1.96	24.54	29.52	26.80
PaF	1.24	38.99	30.12	33.99
BPE-ML _{ENG}	7.25	13.52	23.15	17.07
MoP-ML _{ENG}	6.40	12.84	20.89	15.91
PaF-ML _{ENG}	12.98	3.73	1.72	2.35
BPE-ML _{NLD}	5.63	16.44	27.11	20.46
MoP-ML _{NLD}	5.97	13.04	22.78	16.59
PaF-ML _{NLD}	10.79	14.98	9.69	11.77
BPE-ML _{ZHO}	8.25	0.00	0.00	0.00
MoP-ML _{ZHO}	1.18	77.04	66.96	71.65
PaF-ML _{ZHO}	2.36	50.15	32.66	39.55

Table 4: BPE, ParadigmFinder (PaF), and MorPiece (MoP) multilingual (-ML) evaluation (ENG, NLD, and ZHO, 16K per language vocab) on SIGMORPHON-inspired tasks. Gray rows show monolingual tokenizer results (BabyLM 10M).

Tokenizer	ENG	NLD	ZHO
BPE-ML	0.776863	0.845	0.437500
UniG-ML	0.515994	0.460	0.420825
MoP-ML	0.801362	0.840	0.467525
PaF-ML	0.808980	0.835	0.475000

Table 5: BPE, UniGram (UniG), MorPiece (MoP) and ParadigmFinder (PaF) multilingual (-ML) evaluation on local and non-local agreement on BLiMP-tasks.

monolingual counterparts. This degradation is especially pronounced for PaF, whose reliance on matching words against induced, language-specific paradigms makes it particularly sensitive to reduced vocabulary coverage in the multilingual setting. BPE remains comparatively strong for ENG and NLD, whereas MoP achieves more consistent performance across all three languages. In Table 5, we report the performance of the same baseline GPT-2 model trained with different tokenizers.

In the extrinsic evaluation, MoP and PaF outperform the other tokenizer, surpassing BPE by 3 percentage points in both English and Chinese. In Dutch, they trail BPE by at most 1 percentage point. Taken together, the two evaluation tests seem to suggest that although the two multilingual morphological tokenizers might struggle to correctly identify morpheme boundaries in morphologically complex unseen words, they nonetheless give a non-negligible advantage in downstream tasks involving morphological agreement. Nevertheless, in the remainder of this work, we report experiments only with MoP, as it exhibits more consistent performance across the two tokenizer evalua-

tions. Increasing vocabulary size produces a mild improvement in BPE and MoP tokenizers, but not for Unigram.

6.2 Architectural choices impact

One clear result of our experiments is that Hawk is sample-efficient while GPT-2 is not. At the 100M-word budget, Hawk-B beats GPT-2 MoP by 6.5 points (.6110 vs .5463) and even Hawk-A beats it by 5. The Transformer’s eventual catch-up happens at a token count the challenge doesn’t reward. Recurrence wins at the developmentally-plausible data scale, and the Transformer (GPT-2 baseline model) only equalizes with $\sim 10\times$ the budget the task allows (Table 7 in the Appendix).

On English, Hawk with MoP scores highest on (72.4%), followed by Hawk with MoP+bf (70.4%), FiLM (70.3%), SWITCH (70.0%), Hawk B (69.5%), Hawk C (69.4%), BabyLM GPT-2 baseline (69.2%), and Hawk A (69.0%). Globally, none of the capacity variation on Hawk architecture produced significant changes in performance, while parametric injection (under the SWITCH supervision) only mildly improves both with respect to the un-injected Hawk baseline (+0.005) and with respect to the GPT-2 baseline (+0.017) on BLiMPs averaged (Table 6 in the Appendix).

Focusing on more fine-grained linguistic phenomena, FiLM produces the single largest category-level gain seen anywhere in this study: Island Effects⁵ rise from 56.2% (baseline) to 61.9% (+5.8 pts), consistent with the kind of hierarchical/structural sensitivity that hypotheses (Hawk A) are ultimately after. FiLM also gains slightly on Control/Raising and Binding. Against this, FiLM loses on Ellipsis (-7.0 pts), Subject-Verb Agreement (-2.6 pts) and Det-Noun Agreement (-2.2 pts), and on the fine-grained clusters it is negative on every agreement and WH-related cluster (see Tables 8-10 for fine-grained results on BLiMP tasks across the three languages).

SWITCH is closer to a flat rescaling of the baseline: small losses on most categories (NPI Licensing -3.2 pts, Quantifiers -2.4 pts, Anaphor Agreement -0.6 pts), essentially no change on Island Effects, but the largest gain of all seven models on the external MultiBLiMP-EN agreement benchmark (89.5% vs.

⁵In BLiMP: adjunct_island, complex_NP_island, coordinate_structure_constraint_*, left_branch_island_*, sentential_subject_island, wh_island; wh_island, in fact, should be considered a "weak" islands (Szabolcsi and Zwarts, 1993).

87.0% baseline), which is not fully consistent with the modest losses SWITCH shows on the internal SV/Det–Noun agreement clusters. This discrepancy is a useful reminder that "local agreement" is operationalized somewhat differently across benchmarks, and a single mechanism's ranking can depend on which agreement probe is used.

On the capacity variants, none of A, B, or C set-ups improves the overall BLiMP aggregate relative to the (also LiP-free) Hawk MoP+bf baseline: A is 2.1 points lower, B 1.6 points lower, C 1.9 points lower. Whatever local benefits the capacity changes bring, they come with a broad-based cost elsewhere in the grammar suite. Nevertheless, all three capacity variants show a small increase in Island Effects sensitivity over the baseline (A: +0.6, B: +2.7, C: +2.6 pts), mirroring the FiLM result.

7 Discussion

Consistent with previous findings, (i) the tokenization experiments show that, despite their morphological reliability (on agreement forced-choice tasks), the impact of linguistically motivated tokenizers (MoP and PaF) is not uniform across downstream tasks, including purely morphosyntactic contrasts such as those assessed by the BLiMP benchmark (Fusco et al., 2025) (ii) complex wh-dependencies remain difficult to capture (Lan et al., 2024), even with the apparently more syntactically sensitive and budget-efficient recurrent architectures (Wilcox et al., 2023). The significant improvement on island phenomena under the FiLM approach suggests that the linguistic intuition underlying the LiP may play an important role in this domain. However, the deterioration in performance across other structural domains warrants caution. Repeating the experiments across multiple runs would provide more precise uncertainty estimates and help establish the robustness of these findings. Overall, neither LiPs injection nor the capacity modifications yielded the expected performance gains on the targeted phenomena. Despite isolated, linguistically unpredictable significant improvements, the overall performance remained in line with the baseline of un-injected models. Back to the tokenization strategies, it is worth highlighting that estimating the Order of Acquisition (OoA) of morphologically productive splits is a by-product of the incremental approach adopted by the MoP tokenizer. The positive correlation with the English gold standard is noteworthy and com-

plements the Age of Acquisition (AoA) evaluations included in the BabyLM benchmark.

8 Conclusion

In this paper, we compared the attention-based GPT-2 baseline performance with a state-of-the-art recurrent model with no attention at all. The Hawk model we tested trains much more efficiently than GPT-2 model and reaches top-performance after 300/400M word budget (3 or 4 training epochs). The choice of a RNN was motivated by cognitive plausibility: only models that simulate incremental processing can be considered fully coherent with the way humans generate and parse sentences. We also believed that forcing the model to make early decision on what to retain in memory (recurrent block) and what to forget, might induce the kind of relevant bias affect linguistic processing (Chesi et al., 2026). This conclusion is only partially supported by the mild improvement Hawk models record over the GPT-2 baseline. Another central hypothesis concerned whether linguistically motivated parametric choices could help the model distinguish ENG from NLD and ZHO (Roberts, 2019). To investigate this possibility, we isolated three relevant parameters: Head Directionality (HD), which regulates local predicate–argument ordering (ENG vs. NLD); Verb Second (V2), which requires a tense-bearing element to move to the second position in the clause (NLD vs. ENG/ZHO); and Filler–Gap Dependency (WH), which requires a wh-item to move from its thematic position (ENG/NLD vs. ZHO). We incorporated these parametric switches into the RNN model and allowed it to use them to regulate both the embedding and the recurrent block. We implemented two alternative injection strategies: FiLM and a fully supervised, gold-standard SWITCH. Despite modest improvements in some targeted non-local phenomena under the FiLM approach, none of our attempts produced the expected significant overall improvement relative to the baseline. Given the current (limited and focused) linguistic parametric implementation, these results might suggest that LiPs serve as a scientific probe rather than a capacity crutch. Overall, we observe that the loss landscape is actively hostile to phrase structure and not just merely indifferent: whenever model capacity is given to a specific dimension with a specific structural hypotheses, performance oscillation often bend in unexpected linguistic directions.

9 Limitations

Due to space limits, in this study, we report results only for a specific RNN architecture (Hawk models). We conducted many additional experiments on a range of RNN architectures, all motivated by assumptions about linguistic processing, including eMG-RNNs and Selective State Space models with different channel capacities designed to bias syntactic processing according to locality constraints (local merge, raising and control, long-distance dependencies, phase-based movement, and binding). All trained models are available at [NeTSlab/babylm-2026](https://github.com/NeTSlab/babylm-2026). Source code for training tokenizers and models is available at [cristianochesi/babylm-2026](https://github.com/cristianochesi/babylm-2026).

The LiP supervision scheme is highly simplified. Considering the three parameters (HD, V2, and WH) that distinguish English (ENG), Dutch (NLD), and Chinese (ZHO), we represent them using three-dimensional vectors encoding their values across the three languages: HD = [1, 0.5, 1], V2 = [0, 1, 0], and WH = [1, 1, 0]. The injection strategy is consistent with the hypothesis that one parameter (e.g., the HD LiP) should operate directly on the embeddings, whereas another (e.g., the WH LiP) should instead be applied to the recurrent block. This is by no means the only plausible hypothesis, and further experiments are needed to explore alternative injection strategies.

Another fundamental assumption concerning linguistic parameters, which is not addressed in this paper, is that the setting of one parameter constrains the settings of others (Roberts, 2019). This implicational structure is central to the *Poverty of the Stimulus argument* and is consistent with cross-linguistic observations concerning language acquisition: children acquire language rapidly, follow similar developmental paths, and exhibit similar patterns of error (Yang, 2004; Guasti, 2017). Building on this assumption, Biberauer and Roberts (2015) propose a hierarchical model comprising four parameter levels (macro-, meso-, micro-, and nano-parameters), ordered according to the size of the grammatical class targeted by each parameter. Macro-parameters are therefore more general, applying consistently to all heads of the relevant type, and predict the simultaneous clustering of several properties, whereas micro- and nano-parameters capture increasingly idiosyncratic and lexically restricted variation. We assume that a positive setting of a higher-order parameter implies, or renders ir-

relevant, a cluster of lower-order options, whereas a negative setting blocks them (Pescarini, 2025). In this study, we focus only on three prototypical LiPs, necessarily simplifying the more nuanced linguistic analyses that recast these classical parameters in Minimalist, feature-based terms. No implicational dependencies are considered at this stage.

Regarding the OoA analysis, the gold standard considered here is available only for English. We have comparable gold standards based on Belletti and Guasti (2015) and Caprin and Guasti (2009) available at [cristianochesi/babylm-2026](https://github.com/cristianochesi/babylm-2026); similar resources would need to be developed for ZHO and NLD.

Acknowledgments

This work contains simulations carried out on the High Performance Computing DataCenter at IUSS Pavia, co-funded by Regione Lombardia through the funding programme established by Regional Decree No. 3776 of November 3, 2020.

Author Contributions

CC: conceptualization; supervision; methodology; software (training pipeline and evaluation harness); formal analysis; funding acquisition; resources (compute infrastructure); writing — original draft. **AF:** conceptualization; methodology; software (training pipeline and evaluation harness); formal analysis; writing — original draft. **AM:** data curation; software; writing — original draft. **GF, SR:** conceptualization; methodology; formal analysis; writing — original draft. **HBC, TS, AZ:** data curation; methodology; formal analysis; writing — review and editing. All authors read and approved the final manuscript and take responsibility for its content.

Use of AI Assistants

We disclose the use of generative AI assistants in the preparation of this work, in line with the ACL Policy on Publication Ethics.

Writing. ChatGPT (SOL, version 5.6, accessed JUNE-SEPTEMBER 2026) was used exclusively for language editing of text written by the authors: grammar and spelling correction, rephrasing for concision, and $\LaTeX$ formatting. No section of the paper was generated from a prompt describing content, and no claim, argument, related-work summary or citation was produced by the model.

All edited passages were checked by the authors against the original text.

Code. An AI coding assistant (CLAUDE OPUS, version 4 and 5, accessed March-July 2026) was used for bug fixing and significant optimization of the original code as well as auxiliary scripting tasks: boilerplate for data loading, refactoring of the evaluation scripts and visualization scripts. The core modeling code was adapted by the authors from <https://github.com/knotgrass/Griffin>. All AI-assisted code was reviewed, executed and tested by the authors, and is released in the accompanying repository, where the affected files are marked in the README.

Research content. No AI assistant was used to generate research ideas, to design experiments, to analyse or interpret results, to retrieve or select references, or to produce any data reported in the paper. The authors take full responsibility for the content of the manuscript, including any text that was language-edited with AI assistance.

References

- David Adger. 2003. *Core syntax: a minimalist approach*. Core linguistics. Oxford University Press, Oxford ; New York.
- Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. *The SIGMORPHON 2022 shared task on morpheme segmentation*. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.
- Adriana Belletti and Maria Teresa Guasti. 2015. *The Acquisition of Italian: Morphosyntax and its interfaces in different modes of acquisition*, volume 57. John Benjamins Publishing Company.
- Theresa Biberauer and Ian Roberts. 2015. *Rethinking formal hierarchies: A proposed unification*. *Cambridge Occasional Paper in Linguistics*, 7:1–31.
- Roger Brown. 1973. *A first language: The early stages*. Harvard U. Press.
- Claudia Caprin and Maria Teresa Guasti. 2009. The acquisition of morphosyntax in Italian: A cross-sectional study. *Applied psycholinguistics*, 30(1):23.
- Lisa Lai-Shen Cheng. 2009. *Wh-in-situ, from the 1980s to now*. *Language and Linguistics Compass*, 3(3):767–791.
- Cristiano Chesi. 2023. *Parameters of cross-linguistic variation in expectation-based Minimalist Grammars (e-MGs)*. *Italian Journal of Computational Linguistics*, 9(1).
- Cristiano Chesi. 2026. MorPiece: A split-based tokenization library. <https://github.com/cristianochesi/morpiece>. Version 1.4.5.
- Cristiano Chesi, Matilde Barbini, Veronica Bresnan, Achille Fusco, Sofia Neri, Maria Letizia Piccini Bianchessi, Sarah Rossi, and Tommaso Sgrizzi. 2026. *From recursion to incrementality: return to recurrent neural networks*. *Linguistics Vanguard*.
- Noam Chomsky. 1981. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.
- Noam Chomsky. 1986. *Knowledge of language: its nature, origin, and use*. Convergence. Praeger, New York.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. 2024. *Griffin: Mixing gated linear recurrences with local attention for efficient language models*. *Preprint*, arXiv:2402.19427.
- Jill De Villiers and Tom Roeper, editors. 2011. *Handbook of Generative Approaches to Language Acquisition*, volume 41 of *Studies in Theoretical Psycholinguistics*. Springer Netherlands, Dordrecht.
- Jeffrey L. Elman. 1990. *Finding Structure in Time*. *Cognitive Science*, 14(2):179–211.
- Edward Fredkin. 1960. Trie memory. *Communications of the ACM*, 3(9):490–499.
- Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo, and Cristiano Chesi. 2025. *Linguistic Units as Tokens: Intrinsic and Extrinsic Evaluation with BabyLM*. In *Proceedings of the First BabyLM Workshop*, pages 496–507, Suzhou, China. Association for Computational Linguistics.
- Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. *Unpacking tokenization: Evaluating text compression and its correlation with model performance*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2274–2286, Bangkok, Thailand. Association for Computational Linguistics.
- John Goldsmith. 2001. *Unsupervised learning of the*

- morphology of a natural language. *Computational Linguistics*, 27(2):153–198.
- Albert Gu and Tri Dao. 2024. [Mamba: Linear-Time Sequence Modeling with Selective State Spaces](#). *arXiv preprint*. ArXiv:2312.00752 [cs].
- Maria Teresa Guasti. 2017. *Language acquisition: The growth of grammar*. MIT press, Cambridge, MA.
- Kenneth L. Hale and Samuel Jay Keyser. 1993. On argument structure and the lexical expression of syntactic relations. *An annotated syntax reader*, pages 312–327.
- Kenneth L. Hale and Samuel Jay Keyser. 2002. *Prolegomenon to a theory of argument structure*. Number 39 in Linguistic inquiry Monographs. MIT Press, Cambridge, Mass.
- Zellig S Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.
- Richard S. Kayne. 2018. [The place of linear order in the language faculty](#). Ms. New York University.
- Jordan Kodner, Sarah Payne, and Jeffrey Heinz. 2023. [Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi \(2023\)](#). *arXiv preprint*. ArXiv:2308.03228 [cs].
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. [Large Language Models and the Argument from the Poverty of the Stimulus](#). *Linguistic Inquiry*, pages 1–28.
- Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. 2023. [Resurrecting recurrent neural networks for long sequences](#). *Preprint*, arXiv:2303.06349.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2017. [Film: Visual reasoning with a general conditioning layer](#). *Preprint*, arXiv:1709.07871.
- Diego Pescarini. 2025. [Quantitative modelling and syntactic microvariation](#). *Atti del Sodalizio Glottologico Milanese*, 2023(18 n.s.):91–108.
- Ian Roberts. 2019. *Parameter Hierarchies and Universal Grammar*. Oxford University Press.
- Jenny R Saffran, Elizabeth K Johnson, Richard N Aslin, and Elissa L Newport. 1999. [Statistical learning of tone sequences by human infants and adults](#). *Cognition*, 70(1):27–52.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural Machine Translation of Rare Words with Subword Units](#). *arXiv preprint*. ArXiv:1508.07909 [cs].
- Anna Szabolcsi and Frans Zwarts. 1993. [Weak islands and an algebraic semantics for scope taking](#). *Natural Language Semantics*, 1(3).
- Ethan Wilcox, Richard Futrell, and Roger Levy. 2023. [Using Computational Models to Test Syntactic Learnability](#). *Linguistic Inquiry*, pages 1–44.
- Hongzhi Xu, Mitchell Marcus, Charles Yang, and Lyle Ungar. 2018. [Unsupervised morphology learning with statistical paradigms](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 44–54, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Charles D. Yang. 2004. [Universal Grammar, statistics or both?](#) *Trends in Cognitive Sciences*, 8(10):451–456.
- Charles D. Yang, Stephen Crain, Robert C. Berwick, Noam Chomsky, and Johan J. Bolhuis. 2017. [The growth of language: Universal Grammar, experience, and principles of computation](#). *Neuroscience & Biobehavioral Reviews*, 81:103–119.
- C. Jan-Wouter Zwart. 1997. *Morphosyntax of Verb Movement*, volume 39 of *Studies in Natural Language and Linguistic Theory*. Springer Netherlands, Dordrecht.

A Appendix

A.1 Byte_fallback reduces model’s sensitivity to non-local dependencies

Two baseline Hawk models were trained one without (default), the other with (+fb) MoP byte_fallback option set. Since the performance does not match significantly on all phenomena, also considering the evaluation standard errors of each phenomenon. The Hawk+bf is not uniformly better than its Hawk predecessor: it gains on Anaphor Agreement (+1.3 pts) and Control/Raising (+0.8 pts), is essentially flat on Determiner–Noun Agreement, Ellipsis, and Subject-Verb Agreement, but loses on Irregular Forms (-3.2 pts), Quantifiers (-4.3 pts), Island Effects (-4.6 pts), NPI Licensing (-5.4 pts), and Binding (-2.2 pts). These are exactly the categories that involve tracking a licensing or structural condition across some intervening material (island boundaries, NPI licensors, quantifier scope, binding domains), so at face value the byte_fallback mechanism, on its own, does not increase, and by these metrics somewhat decreases, sensitivity to long-range structural dependencies.

A.2 Technical details of the linguistic-parameter (LiP) control

A more detailed computational graph is illustrated in Figure 3. A predictor ($\text{Linear}(H \rightarrow H) \rightarrow \text{GeLU} \rightarrow \text{Linear}(H \rightarrow k)$ over the mean-pooled embeddings) emits a parameter vector $p \in \mathbb{R}^k$, with $H = d_{\text{model}}$ and k the number of inferred parameters; it is regressed onto `default_logits`, and under SWITCH the gold $p = \text{param_targets}[\text{lang_ids}]$ is teacher-forced and the predictor is detached, i.e. it acts as an inference-time predictor only. The same p is injected at two sites. ① *Embedding site*, applied once per sequence right after `wte`: FILM sets $x \leftarrow x + W_e p$ (`film_embed`, $\text{Linear}(k \rightarrow H)$, no bias), SWITCH adds the boundary embeddings $m_{\text{HD}} e_{\text{HD}} + m_{\text{V2}} e_{\text{V2}}$ (`event_expect`). ② *Recurrent site*, broadcast to all N layers: FILM adds $W_r p$ to the recurrence-gate bias (`film_recur`, $\text{Linear}(k \rightarrow d_{\text{rnn}})$, no bias), SWITCH rescales the decay, $\log a_t \leftarrow \gamma_{\text{WH}} \odot \log a_t$ (`wh_gamma` $\in \mathbb{R}^{d_{\text{rnn}}}$).

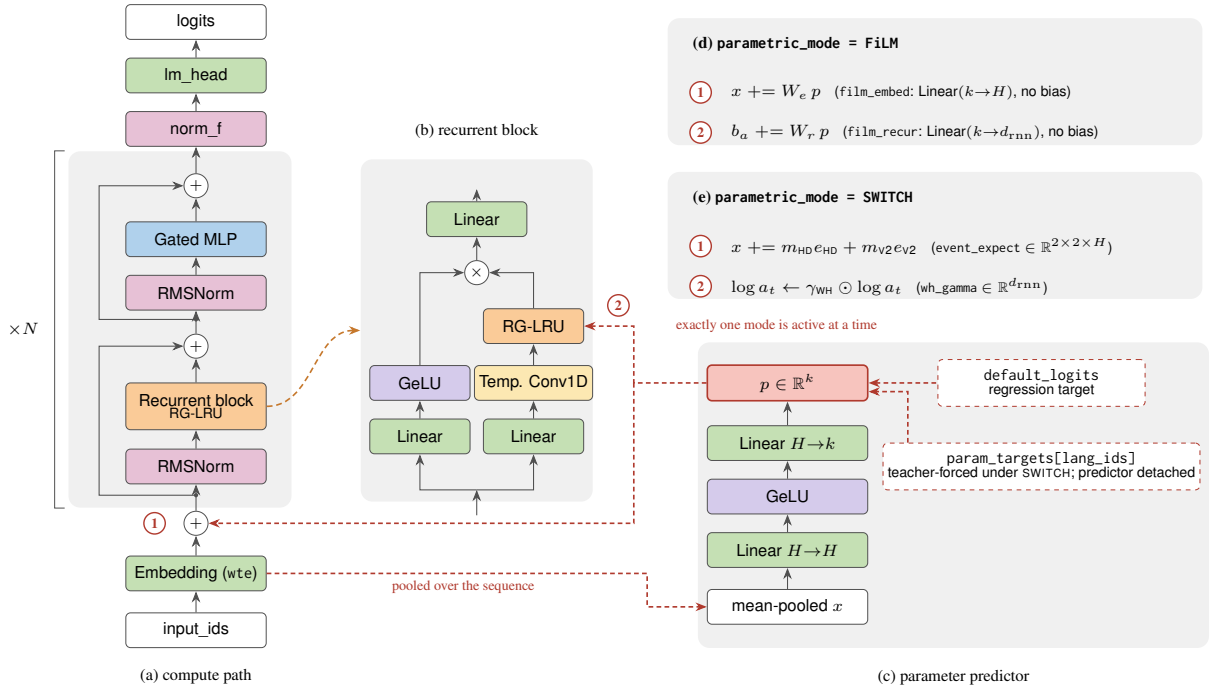


Figure 3: Expanded view of Hawk with linguistic-parameter (LiP) control. (a) The full compute path, with the embedding site ① right after wte and the recurrent site ② inside every one of the N layers. (b) The recurrent block opened up: ② acts on the RG-LRU, either on its recurrence-gate bias or on its decay. (c) The parameter predictor, together with the regression target used to train it and the gold parameters that replace its output under SWITCH, where the predictor is detached and therefore used at inference time only. (d)–(e) The update applied at each site under the two parameterization modes; $H = d_{\text{model}}$, k is the number of inferred parameters, d_{rnn} the RG-LRU state width and b_a the recurrence-gate bias. Colour coding is as in Figure 2.

Task	Lang.	GPT-2 baseline	GPT-2 PaF	GPT-2 MoP	HAWK MoP-base	HAWK MoP-A	HAWK MoP-B	HAWK MoP-C	HAWK MoP-FiLM	HAWK MoP-switch
blimp_babylm_filtered	eng	0.692	0.682	0.688	0.704	0.690	0.695	0.694	0.703	0.700
hellaswag_en_mubench	eng	0.267	0.260	0.262	0.260	0.264	0.263	0.262	0.264	0.264
multiblimp_eng	eng	0.871	0.884	0.855	0.870	0.857	0.871	0.870	0.866	0.895
winogrande_en_mubench	eng	0.488	0.509	0.508	0.510	0.501	0.482	0.513	0.519	0.506
xstorycloze_en_mubench	eng	0.499	0.496	0.490	0.484	0.511	0.501	0.494	0.491	0.496
blimp_nl	nld	0.793	0.781	0.777	0.795	0.773	0.782	0.763	0.790	0.803
hellaswag_nl_mubench	nld	0.264	0.264	0.264	0.261	0.262	0.267	0.262	0.263	0.259
multiblimp_nld	nld	0.919	0.883	0.897	0.915	0.892	0.885	0.890	0.892	0.911
winogrande_nl_mubench	nld	0.493	0.489	0.500	0.497	0.506	0.495	0.479	0.495	0.506
xcomps_nl	nld	0.527	0.543	0.548	0.561	0.556	0.557	0.562	0.565	0.563
xstorycloze_nl_mubench	nld	0.474	0.478	0.487	0.491	0.498	0.490	0.498	0.502	0.485
hellaswag_zh_mubench	zho	0.271	0.273	0.271	0.268	0.270	0.268	0.270	0.269	0.270
winogrande_zh_mubench	zho	0.497	0.505	0.505	0.495	0.480	0.487	0.485	0.494	0.479
xcomps_zh	zho	0.539	0.569	0.554	0.568	0.551	0.550	0.557	0.557	0.555
xstorycloze_zh_mubench	zho	0.506	0.502	0.503	0.509	0.502	0.514	0.519	0.517	0.503
zhoblimp	zho	0.754	0.787	0.796	0.803	0.769	0.772	0.774	0.768	0.805
global_piqa_nonparallel_en	eng	–	0.480	–	0.510	0.540	0.480	0.540	0.570	0.520
global_piqa_parallel_en	eng	–	0.233	–	0.262	0.282	0.282	0.311	0.252	0.204
global_piqa_nonparallel_nl	nld	–	0.490	–	0.520	0.570	0.540	0.500	0.550	0.530
global_piqa_parallel_nl	nld	–	0.262	–	0.233	0.272	0.301	0.252	0.233	0.301
global_piqa_nonparallel_zh	zho	–	0.460	–	0.430	0.480	0.500	0.520	0.450	0.440
global_piqa_parallel_zh	zho	–	0.194	–	0.165	0.204	0.243	0.243	0.184	0.184
arc	eng	0.308	0.258	0.237	0.240	0.240	0.265	0.252	0.262	0.252
arc	nld	0.265	0.246	0.260	0.277	0.273	0.237	0.275	0.231	0.254
arc	zho	0.279	0.269	0.290	0.262	0.265	0.244	0.258	0.265	0.240
belebele	eng	0.205	0.227	0.199	0.227	0.222	0.193	0.188	0.216	0.216
belebele	nld	0.205	0.222	0.244	0.222	0.193	0.233	0.205	0.216	0.233
belebele	zho	0.256	0.193	0.176	0.210	0.244	0.210	0.216	0.176	0.244
bmlama	eng	0.361	0.229	0.227	0.144	0.145	0.145	0.137	0.138	0.139
bmlama	nld	0.334	0.224	0.249	0.152	0.148	0.143	0.142	0.151	0.137
bmlama	zho	0.342	0.193	0.204	0.140	0.136	0.142	0.138	0.132	0.127
mnli	eng	0.666	0.560	0.552	0.430	0.385	0.387	0.392	0.414	0.436
mnli	nld	0.612	0.556	0.570	0.433	0.380	0.392	0.415	0.392	0.426
mnli	zho	0.671	0.588	0.532	0.436	0.391	0.396	0.399	0.380	0.449
sib200	eng	0.604	0.795	0.805	0.680	0.670	0.595	0.610	0.615	0.765
sib200	nld	0.662	0.790	0.785	0.700	0.615	0.530	0.600	0.625	0.730
sib200	zho	0.630	0.800	0.805	0.745	0.600	0.530	0.615	0.670	0.660
truthfulqa	eng	0.304	0.277	0.223	0.223	0.268	0.313	0.134	0.205	0.214
truthfulqa	nld	0.313	0.304	0.214	0.223	0.196	0.241	0.250	0.214	0.232
truthfulqa	zho	0.295	0.170	0.179	0.223	0.223	0.277	0.241	0.223	0.250
xnli	eng	0.645	0.544	0.545	0.444	0.447	0.438	0.447	0.437	0.458
xnli	zho	0.599	0.515	0.515	0.444	0.422	0.450	0.424	0.411	0.449
include	nld	0.348	0.223	0.312	0.339	0.384	0.232	0.304	0.277	0.277
include	zho	0.259	0.232	0.250	0.223	0.304	0.223	0.232	0.286	0.286
pos	eng	0.941	0.898	–	0.922	–	0.920	0.922	0.926	0.923
pos	nld	0.956	0.892	–	0.939	–	0.931	0.933	0.934	0.936
pos	zho	0.917	0.883	–	0.905	–	0.892	0.892	0.899	0.904

Table 6: Model scores on the BabyLM Multilingual Benchmark by task.

Budget (words)	GPT-2 MoP	Hawk-base	Hawk-A	Hawk-B	Hawk-C	Hawk-FiLM	Hawk-switch
<i>Macro average (mean of the three per-language zero-shot aggregates)</i>							
1M	0.4716	0.4851	0.4807	0.4696	0.4726	0.4743	0.4761
2M	0.4687	0.4788	0.4808	0.4826	0.4785	0.4819	0.4761
3M	0.4683	0.4840	0.4807	0.4826	0.4796	0.4834	0.4845
4M	0.4703	0.4893	0.4870	0.4851	0.4854	0.4867	0.4871
5M	0.4767	0.5026	0.4941	0.4914	0.4911	0.4952	0.4952
6M	0.4719	0.5036	0.5064	0.5007	0.5042	0.5035	0.5041
7M	0.4727	0.5178	0.5167	0.5187	0.5136	0.5224	0.5118
8M	0.4774	0.5241	0.5185	0.5247	0.5229	0.5204	0.5193
9M	0.4760	0.5271	0.5217	0.5319	0.5342	0.5266	0.5316
10M	0.4777	0.5397	0.5337	0.5362	0.5369	0.5327	0.5334
20M	0.4843	0.5614	0.5622	0.5647	0.5564	0.5627	0.5566
30M	0.4946	0.5735	0.5697	0.5750	0.5693	0.5731	0.5677
40M	0.4967	0.5869	0.5813	0.5863	0.5798	0.5808	0.5795
50M	0.5135	0.5849	0.5853	0.5864	0.5854	0.5867	0.5794
60M	0.5230	0.5927	0.5878	0.5972	0.5907	0.5881	0.5864
70M	0.5285	0.5955	0.5919	0.6023	0.5889	0.5982	0.5955
80M	0.5372	0.5941	0.5945	0.6058	0.5916	0.5972	0.5936
90M	0.5404	0.5975	0.5993	0.6058	0.5950	0.5990	0.5976
100M	0.5463	0.5990	0.5960	0.6110	0.6002	0.5946	0.6000
200M	0.5758	0.6115	0.6070	0.6163	0.6120	0.6147	0.6136
300M	0.5913	0.6157	0.6129	0.6193	0.6161	0.6164	0.6177
400M	0.6003	0.6177	0.6140	0.6212	0.6162	0.6193	0.6166
500M	0.6008	0.6160	0.6106	0.6206	0.6169	0.6223	0.6153
600M	0.6056	0.6157	0.6139	0.6194	0.6147	0.6196	0.6149
700M	0.6061	0.6116	0.6118	0.6169	0.6154	0.6189	0.6139
800M	0.6074	0.6078	0.6089	0.6142	0.6116	0.6146	0.6119
900M	0.6081	0.6078	0.6058	0.6099	0.6084	0.6135	0.6076
1000M	0.6109	0.6078	0.6029	0.6060	0.6055	0.6101	0.6119
best	0.6109	0.6177	0.6140	0.6212	0.6169	0.6223	0.6119
<i>Grammar average (mean of the five grammaticality benchmarks)</i>							
1M	0.5602	0.5914	0.5706	0.5702	0.5802	0.5626	0.5838
2M	0.5590	0.5853	0.5867	0.5915	0.5905	0.5845	0.5903
3M	0.5621	0.5955	0.5930	0.5974	0.5950	0.5878	0.5954
4M	0.5734	0.6029	0.5983	0.6084	0.5942	0.6016	0.6030
5M	0.5859	0.6310	0.6109	0.6187	0.6080	0.6092	0.6158
6M	0.5746	0.6315	0.6343	0.6335	0.6385	0.6392	0.6265
7M	0.5726	0.6543	0.6481	0.6661	0.6522	0.6654	0.6488
8M	0.5750	0.6753	0.6638	0.6781	0.6662	0.6712	0.6534
9M	0.5695	0.6819	0.6653	0.6914	0.6738	0.6837	0.6806
10M	0.5746	0.6935	0.6760	0.6934	0.6849	0.6842	0.6781
20M	0.5967	0.7393	0.7424	0.7462	0.7324	0.7439	0.7260
30M	0.6166	0.7579	0.7568	0.7611	0.7513	0.7559	0.7571
40M	0.6231	0.7776	0.7700	0.7766	0.7697	0.7731	0.7621
50M	0.6460	0.7805	0.7821	0.7800	0.7747	0.7800	0.7740
60M	0.6707	0.7847	0.7843	0.7887	0.7846	0.7810	0.7777
70M	0.6796	0.7855	0.7857	0.7949	0.7791	0.7915	0.7858
80M	0.6876	0.7921	0.7883	0.7989	0.7853	0.7930	0.7885
90M	0.6911	0.7991	0.7993	0.7985	0.7926	0.7947	0.7963
100M	0.7045	0.7948	0.7918	0.8075	0.7928	0.7935	0.7975
200M	0.7510	0.8106	0.8080	0.8183	0.8126	0.8128	0.8156
300M	0.7794	0.8166	0.8155	0.8220	0.8152	0.8191	0.8208
400M	0.7850	0.8174	0.8190	0.8220	0.8193	0.8190	0.8226
500M	0.7919	0.8141	0.8118	0.8243	0.8157	0.8233	0.8156
600M	0.7988	0.8127	0.8121	0.8192	0.8104	0.8198	0.8180
700M	0.7989	0.8061	0.8091	0.8150	0.8098	0.8186	0.8161
800M	0.8002	0.8017	0.8025	0.8123	0.8054	0.8107	0.8105
900M	0.8014	0.8017	0.7978	0.8043	0.8028	0.8083	0.8038
1000M	0.8027	0.8017	0.7928	0.7960	0.7958	0.8028	0.8105
best	0.8027	0.8174	0.8190	0.8243	0.8193	0.8198	0.8226

Table 7: Accuracy as a function of training budget. Rows are the `chck_NM` revisions, i.e. the number of training words seen (the trainer checkpoints). The shaded row corresponds to 100M-word budget for efficiency comparison. Higher is better.

Table 8: Accuracy of HAWK models on English BLiMP (*blimp_filtered*) by single phenomena.

Phenomenon	base	base bf	A	B	C	FiLM	switch
adjunct_island	0.800	0.839	0.789	0.752	0.777	0.772	0.807
anaphor_gender_agreement	0.929	0.955	0.904	0.919	0.931	0.919	0.935
anaphor_number_agreement	0.970	0.971	0.954	0.943	0.950	0.960	0.979
animate_subject_passive	0.674	0.724	0.677	0.735	0.723	0.726	0.685
animate_subject_trans	0.817	0.797	0.738	0.751	0.723	0.741	0.798
causative	0.675	0.667	0.654	0.623	0.617	0.655	0.670
complex_NP_island	0.402	0.398	0.381	0.450	0.431	0.453	0.345
coordinate_structure_constraint_complex_left_branch	0.529	0.471	0.525	0.538	0.557	0.617	0.602
coordinate_structure_constraint_object_extraction	0.813	0.753	0.681	0.700	0.679	0.670	0.739
determiner_noun_agreement_1	0.978	0.977	0.976	0.961	0.968	0.967	0.969
determiner_noun_agreement_2	0.980	0.980	0.962	0.934	0.952	0.934	0.960
determiner_noun_agreement_irregular_1	0.865	0.860	0.852	0.831	0.858	0.819	0.846
determiner_noun_agreement_irregular_2	0.891	0.902	0.909	0.878	0.905	0.906	0.895
determiner_noun_agreement_with_adj_2	0.932	0.935	0.912	0.896	0.891	0.897	0.920
determiner_noun_agreement_with_adj_irregular_1	0.811	0.816	0.811	0.801	0.823	0.791	0.831
determiner_noun_agreement_with_adj_irregular_2	0.839	0.823	0.832	0.788	0.825	0.826	0.855
determiner_noun_agreement_with_adjective_1	0.953	0.950	0.931	0.928	0.922	0.930	0.927
distractor_agreement_relational_noun	0.503	0.504	0.402	0.501	0.536	0.491	0.448
distractor_agreement_relative_clause	0.433	0.479	0.435	0.421	0.503	0.435	0.476
drop_argument	0.771	0.699	0.693	0.695	0.675	0.716	0.707
ellipsis_n_bar_1	0.620	0.613	0.611	0.559	0.517	0.526	0.539
ellipsis_n_bar_2	0.715	0.717	0.710	0.676	0.665	0.665	0.754
existential_there_object_raising	0.746	0.778	0.766	0.780	0.740	0.772	0.729
existential_there_quantifiers_1	0.988	0.956	0.941	0.937	0.926	0.941	0.947
existential_there_quantifiers_2	0.286	0.422	0.381	0.599	0.480	0.401	0.389
existential_there_subject_raising	0.831	0.853	0.834	0.842	0.807	0.813	0.827
expletive_it_object_raising	0.721	0.723	0.697	0.714	0.736	0.744	0.717
inchoative	0.542	0.556	0.557	0.563	0.530	0.585	0.513
intransitive	0.660	0.647	0.618	0.632	0.623	0.637	0.619
irregular_past_participle_adjectives	0.906	0.827	0.888	0.849	0.807	0.912	0.973
irregular_past_participle_verbs	0.774	0.790	0.736	0.745	0.806	0.736	0.709
irregular_plural_subject_verb_agreement_1	0.765	0.754	0.720	0.706	0.733	0.684	0.741
irregular_plural_subject_verb_agreement_2	0.731	0.765	0.753	0.752	0.731	0.739	0.766
left_branch_island_echo_question	0.467	0.419	0.482	0.588	0.568	0.599	0.496
left_branch_island_simple_question	0.659	0.580	0.621	0.621	0.625	0.691	0.650
matrix_question_npi_licensor_present	0.389	0.350	0.327	0.296	0.286	0.329	0.243
npi_present_1	0.622	0.593	0.589	0.579	0.629	0.600	0.558
npi_present_2	0.658	0.612	0.650	0.615	0.667	0.613	0.548
only_npi_licensor_present	0.989	0.955	0.891	0.781	0.878	0.956	0.989
only_npi_scope	0.723	0.631	0.528	0.594	0.624	0.553	0.507
passive_1	0.870	0.855	0.843	0.821	0.827	0.844	0.832
passive_2	0.807	0.822	0.821	0.814	0.790	0.816	0.808
principle_A_c_command	0.536	0.430	0.416	0.411	0.480	0.464	0.484
principle_A_case_1	1.000	1.000	0.997	0.992	0.998	0.997	1.000
principle_A_case_2	0.904	0.893	0.877	0.869	0.814	0.888	0.917
principle_A_domain_1	0.760	0.874	0.898	0.872	0.916	0.871	0.902

Continued on next page

Table 8 – continued

Phenomenon	base	base bf	A	B	C	FiLM	switch
principle_A_domain_2	0.643	0.696	0.562	0.616	0.593	0.645	0.632
principle_A_domain_3	0.577	0.579	0.493	0.570	0.506	0.567	0.536
principle_A_reconstruction	0.392	0.185	0.299	0.349	0.299	0.328	0.353
regular_plural_subject_verb_agreement_1	0.831	0.772	0.769	0.729	0.758	0.773	0.806
regular_plural_subject_verb_agreement_2	0.714	0.694	0.729	0.716	0.714	0.690	0.693
sentential_negation_npi_licensor_present	0.962	0.905	0.917	0.866	0.934	0.908	0.968
sentential_negation_npi_scope	0.551	0.467	0.625	0.431	0.485	0.526	0.476
sentential_subject_island	0.480	0.378	0.456	0.412	0.390	0.431	0.403
superlative_quantifiers_1	0.724	0.539	0.464	0.663	0.538	0.573	0.514
superlative_quantifiers_2	0.875	0.785	0.675	0.759	0.695	0.774	0.758
tough_vs_raising_1	0.360	0.323	0.325	0.444	0.389	0.440	0.359
tough_vs_raising_2	0.821	0.841	0.811	0.754	0.805	0.766	0.841
transitive	0.759	0.736	0.745	0.743	0.765	0.759	0.762
wh_island	0.710	0.655	0.606	0.647	0.675	0.724	0.515
wh_questions_object_gap	0.782	0.765	0.740	0.725	0.730	0.729	0.736
wh_questions_subject_gap	0.910	0.901	0.831	0.859	0.823	0.878	0.869
wh_questions_subject_gap_long_distance	0.881	0.898	0.823	0.775	0.786	0.777	0.887
wh_vs_that_no_gap	0.970	0.980	0.956	0.935	0.944	0.957	0.951
wh_vs_that_no_gap_long_distance	0.971	0.984	0.949	0.936	0.931	0.952	0.977
wh_vs_that_with_gap	0.313	0.260	0.293	0.321	0.288	0.272	0.318
wh_vs_that_with_gap_long_distance	0.063	0.035	0.067	0.095	0.112	0.105	0.064

Table 9: Accuracy of HAWK models on Dutch BLiMP (*blimp_nl*) by single phenomena.

Phenomenon	base	base bf	A	B	C	FiLM	switch
adpositional_phrases	0.980	0.955	0.965	0.955	0.945	0.975	0.990
adverbial_modification	0.925	0.930	0.940	0.885	0.950	0.900	0.940
anaphor_agreement	0.830	0.860	0.835	0.855	0.850	0.825	0.860
argument_structure	0.789	0.767	0.789	0.796	0.783	0.776	0.792
auxiliaries	0.942	0.956	0.940	0.864	0.910	0.926	0.898
binding_principle_a	0.955	0.965	0.845	0.945	0.925	0.950	0.930
complementive	0.844	0.840	0.772	0.796	0.812	0.830	0.850
crossing_dependencies	0.990	1.000	0.960	0.980	0.950	0.950	0.990
determiners	0.807	0.723	0.762	0.740	0.782	0.785	0.805
extraposition	0.922	0.892	0.818	0.882	0.802	0.842	0.902
finite_argument_clause	0.843	0.838	0.842	0.827	0.770	0.835	0.833
infinitival_argument_clause	0.833	0.890	0.841	0.824	0.816	0.840	0.869
nominalization	0.900	0.880	0.735	0.830	0.835	0.875	0.900
parasitic_gaps	0.226	0.251	0.294	0.436	0.281	0.346	0.149
passive	0.838	0.765	0.790	0.762	0.750	0.797	0.820
quantifiers	0.920	0.940	0.920	0.880	0.915	0.940	0.890
r_words	0.965	0.960	0.850	0.890	0.730	0.840	0.995
relativization	0.667	0.687	0.730	0.797	0.753	0.750	0.723
topicalization	0.695	0.613	0.623	0.605	0.563	0.647	0.705
verb_second	0.985	0.970	0.930	0.865	0.920	0.875	0.995
wh_movement	0.750	0.736	0.710	0.725	0.715	0.740	0.760
wh_movement_restrictions	0.741	0.743	0.677	0.704	0.661	0.704	0.744

Table 10: Accuracy of HAWK models on Chinese BLiMP (*zhoblmp*) by single phenomena.

Phenomenon	base	base bf	A	B	C	FiLM	switch
BA_BEI_subj_drop	0.920	0.937	0.840	0.797	0.787	0.723	0.837
BA_deletion	0.663	0.853	0.760	0.560	0.927	0.770	0.797
BA_duplicate_argument	1.000	1.000	0.993	0.953	0.957	0.997	1.000
BA_inversion	0.903	0.880	0.780	0.833	0.843	0.787	0.893
BA_meiba	1.000	1.000	1.000	1.000	1.000	1.000	1.000
BA_negation	0.770	0.877	0.760	0.757	0.577	0.680	0.923
BA_no_progressive	0.997	0.970	0.957	0.977	0.967	0.973	0.947
BA_no_stative_verb	0.180	0.153	0.247	0.333	0.247	0.073	0.483
BA_suo_adverbial_a	0.997	1.000	0.943	0.997	0.997	0.980	1.000
BA_suo_adverbial_b	1.000	1.000	1.000	0.997	1.000	1.000	1.000
BA_verb_le_a	1.000	1.000	0.993	0.993	0.990	0.987	1.000
BA_verb_le_b	0.787	0.823	0.670	0.670	0.707	0.570	0.713
BEI_construction_a	0.873	0.827	0.613	0.633	0.700	0.713	0.730
BEI_construction_b	0.907	0.857	0.707	0.803	0.700	0.740	0.850
BEI_deletion	0.300	0.417	0.653	0.573	0.647	0.780	0.770
BEI_preposition	0.967	0.977	0.857	0.807	0.833	0.763	0.960
PN_numP_a	1.000	0.993	0.997	0.990	0.997	0.997	1.000
PN_numP_b	0.993	0.983	0.827	0.893	0.853	0.870	0.990
adjective_transitive_dui	0.813	0.720	0.887	0.790	0.737	0.877	0.787
agent_animacy_adv	1.000	1.000	0.983	0.980	0.970	0.980	1.000
agent_animacy_passive	0.630	0.733	0.693	0.713	0.723	0.717	0.680
agent_animacy_subj	0.633	0.710	0.680	0.747	0.687	0.680	0.663
agent_causative	0.690	0.687	0.623	0.653	0.713	0.677	0.753
agent_deletion	0.767	0.593	0.850	0.640	0.643	0.647	0.703
anaphor_gender_agreement	0.560	0.577	0.530	0.547	0.540	0.583	0.557
anaphor_number_agreement	0.237	0.117	0.143	0.333	0.253	0.217	0.050
causative_shi_ba	0.740	0.713	0.727	0.627	0.730	0.700	0.873
classifier_noun_agreement	0.990	0.957	0.977	0.977	0.953	0.953	0.987
classifier_noun_agreement_no_gap	0.237	0.283	0.230	0.327	0.250	0.343	0.187
classifier_noun_subj	1.000	0.993	0.913	0.963	0.927	0.870	1.000
control_modal_vs_raising_modal	0.997	1.000	0.970	0.957	0.993	0.947	0.990
ellipsis_adj	0.463	0.487	0.500	0.543	0.513	0.500	0.510
ellipsis_double_object	0.990	0.947	0.947	0.983	0.957	0.980	1.000
ellipsis_n_bar_class	0.767	0.660	0.683	0.690	0.683	0.670	0.733
existential_there_subject_raising	0.807	0.810	0.630	0.753	0.687	0.633	0.747
fci_renhe_dou	1.000	1.000	0.930	0.810	0.987	0.997	0.993
fci_renhe_prepP	1.000	1.000	1.000	0.997	1.000	1.000	1.000
fci_renhe_ruguo	0.997	1.000	0.970	0.880	0.993	0.840	1.000
fci_renhe_subj	0.880	1.000	0.933	0.497	0.393	0.733	0.823
fci_renhe_suoyou	1.000	1.000	1.000	0.993	1.000	0.993	1.000
intransitive_double_obj	0.917	0.877	0.893	0.843	0.817	0.763	0.867
intransitive_no_obj	0.707	0.700	0.703	0.623	0.653	0.680	0.690
left_adverbial_b	0.970	0.970	0.933	0.920	0.913	0.940	0.970
left_adverbial_d	0.960	0.927	0.873	0.917	0.907	0.933	0.970
left_adverbial_e	0.833	0.883	0.793	0.743	0.790	0.627	0.890
left_adverbial_negation	0.780	0.663	0.637	0.680	0.723	0.637	0.727
left_dou	1.000	1.000	0.973	0.963	0.993	0.987	1.000
modal_raising_hui	0.943	1.000	0.823	0.887	0.900	0.900	0.943
modal_raising_topicalization	0.890	0.797	0.820	0.780	0.753	0.813	0.933
nominal_definite_men	0.997	0.987	0.983	0.973	0.963	0.987	0.997
nominal_modal_insertion	0.733	0.753	0.650	0.693	0.843	0.657	0.833
noun_adjective_shi	0.630	0.593	0.683	0.580	0.620	0.670	0.630

Continued on next page

Table 10 – continued

Phenomenon	base	base bf	A	B	C	FiLM	switch
noun_phrase_conjunction_jian	0.983	0.973	0.970	0.970	0.983	0.993	0.993
npi_renhe_A_not_A_question	0.043	0.097	0.137	0.330	0.233	0.087	0.027
npi_renhe_conditional	0.047	0.073	0.017	0.003	0.227	0.013	0.017
npi_renhe_neg_scope_locP	0.890	0.933	0.437	0.680	0.727	0.700	0.617
npi_renhe_neg_scope_subj	0.997	1.000	0.957	0.953	0.990	0.920	0.997
npi_renhe_wh_question_obj	0.883	0.990	0.847	0.880	0.827	0.957	0.950
npi_renhe_wh_question_subj	0.737	0.760	0.570	0.713	0.707	0.727	0.860
passive_agent_deletion_long_left	0.250	0.370	0.693	0.540	0.850	0.687	0.283
passive_agent_deletion_long_right_a	0.003	0.000	0.250	0.287	0.293	0.200	0.013
passive_agent_deletion_long_right_b	0.000	0.000	0.013	0.013	0.027	0.017	0.000
passive_agent_deletion_short	0.000	0.003	0.100	0.103	0.130	0.440	0.000
passive_body_part	0.863	0.907	0.857	0.920	0.873	0.913	0.817
passive_intransitive	0.890	0.910	0.820	0.867	0.843	0.840	0.897
passive_no_adj	0.943	0.920	0.927	0.903	0.943	0.900	0.943
passive_suo	0.207	0.467	0.427	0.457	0.517	0.487	0.410
plural_cardinal_men_a	1.000	1.000	1.000	1.000	1.000	1.000	1.000
plural_cardinal_men_b	1.000	1.000	1.000	1.000	1.000	1.000	1.000
preposition_deletion	0.967	0.910	0.847	0.817	0.843	0.853	0.927
preposition_insertion	0.987	0.977	0.890	0.943	0.927	0.953	0.987
principle_A_c_command	0.417	0.567	0.490	0.440	0.430	0.450	0.373
principle_A_c_command_number	0.197	0.100	0.030	0.377	0.230	0.227	0.063
principle_A_domain	0.433	0.427	0.373	0.413	0.313	0.447	0.390
principle_A_domain_number	0.987	0.990	0.990	0.853	0.963	0.947	0.980
question_A_not_A	1.000	0.983	0.910	0.910	0.960	0.950	0.983
question_A_not_A_daodi_a	0.937	0.967	0.903	0.943	0.943	0.803	0.993
question_A_not_A_daodi_b	0.940	0.963	0.907	0.923	0.893	0.820	0.980
question_A_not_A_indirect	1.000	1.000	1.000	0.993	0.990	1.000	1.000
question_V_not_VP_1	0.920	0.810	0.713	0.800	0.847	0.480	0.857
question_V_not_VP_2	0.943	0.560	0.873	0.740	0.657	0.440	1.000
question_daodi_nandao_1	1.000	0.970	0.863	0.937	0.977	0.937	1.000
question_daodi_nandao_2	0.707	0.940	0.740	0.590	0.687	0.783	0.577
question_daodi_nandao_A_not_A_intran	0.580	0.973	0.593	0.737	0.830	0.723	0.530
question_daodi_nandao_A_not_A_tran	0.803	0.790	0.717	0.793	0.857	0.767	0.677
question_daodi_negation	0.863	0.943	0.927	0.610	0.690	0.770	0.777
question_nandao_negation	1.000	1.000	0.997	0.993	1.000	0.997	0.997
question_nandao_raising_1_a	1.000	1.000	1.000	1.000	0.993	1.000	1.000
question_nandao_raising_1_b	1.000	1.000	1.000	0.973	0.997	0.997	1.000
question_nandao_raising_2	0.980	0.930	0.920	0.910	0.943	0.880	0.877
question_nandao_raising_3	0.997	0.997	0.893	0.920	0.923	0.940	1.000
question_nandao_scope_1	0.913	0.757	0.920	0.917	0.800	0.957	0.660
question_nandao_scope_2	0.887	0.853	0.747	0.887	0.793	0.773	0.860
question_particle_daodi_choice_intran	0.943	0.923	0.983	0.963	0.993	0.957	0.987
question_particle_daodi_choice_tran	0.993	1.000	0.970	0.970	0.993	0.993	1.000
question_particle_nandao	0.777	0.600	0.627	0.537	0.570	0.760	0.943
relative_operator_intepretation	0.290	0.287	0.553	0.643	0.337	0.250	0.313
relative_operator_who	0.990	1.000	0.993	0.993	0.983	0.940	0.990
relativization_movement_no_gap	1.000	1.000	1.000	0.990	0.997	0.987	1.000
relativization_movement_when_where	0.997	0.993	0.957	0.970	0.957	0.960	0.993
renhe_no_episodic_sentences	0.697	0.917	0.550	0.840	0.893	0.630	0.887
renhe_no_superordinate_negation	0.980	0.813	0.953	0.777	0.970	0.763	0.940
renhe_non_factive_verb	0.643	0.347	0.627	0.470	0.593	0.520	0.513
right_yijing_a	0.943	0.983	0.863	0.940	0.890	0.827	0.993
right_yijing_b	0.637	0.773	0.660	0.710	0.507	0.630	0.857
singular_PN_but_plural_pron	0.953	0.943	0.957	0.977	0.973	0.933	0.940
superlative_quantifiers_1	0.383	0.180	0.097	0.037	0.130	0.557	0.360
superlative_quantifiers_2	0.967	0.907	0.723	0.917	0.683	0.783	0.983
topicalization_OSV	0.970	0.953	0.907	0.907	0.823	0.923	0.957
topicalization_OSV_mei	0.970	0.977	0.860	0.923	0.890	0.870	0.970
topicalization_SOV	0.903	0.920	0.900	0.887	0.843	0.913	0.933
topicalization_SOV_mei	0.973	0.940	0.893	0.943	0.843	0.860	0.943
verb_negation_particle	0.980	0.957	0.943	0.813	0.887	0.897	0.967
verb_phrase_left_adverbial	0.957	0.987	0.890	0.810	0.850	0.917	0.947
verb_phrase_left_negation	0.883	0.947	0.747	0.740	0.590	0.777	0.947
ya_insertion	1.000	0.987	0.920	0.883	0.697	0.930	0.967
you_quantifier_adj	0.977	0.973	0.977	0.947	0.993	0.953	0.983
you_yige	0.953	0.923	0.817	0.923	0.807	0.790	0.963