

Linguistically Inspired Multilingual Pre-Pretraining

Masato Mita¹ Max Müller-Eberstein^{1,2} Misaki Yamaguchi¹ Yohei Oseki¹

¹The University of Tokyo ²IT University of Copenhagen
{mita, iijimax, yamami, oseki}@g.ecc.u-tokyo.ac.jp

Abstract

Encoding Universal Grammar principles (UG) during pre-pretraining (PPT) has been shown to produce inductive biases in language models, which facilitate natural language acquisition. While UG *principles* should be language-agnostic, each language follows different typological *parameters*. To empirically validate this theory, we operationalize three levels of typological flexibility via PPT: non-typological *k*-shuffle Dyck, as well as MP-STRUCT, which encodes the Minimalist Program operations MERGE, AGREE, MOVE following fixed typological parameters. Additionally, we introduce MP-STRUCT BCC (BCC), which parameterizes typological features following the Borer-Chomsky Conjecture to allow for a more typologically flexible parametrization. Through a multilingual, subtask-level evaluation across 207 grammatical tasks for English, Dutch, and Chinese, we find that principles-driven PPT significantly improves average accuracy, but selectively degrades language-specific phenomena such as wh-island constructions. In contrast, structure-only PPT, as well as typologically flexible PPT eliminate wh-island degradation while preserving overall gains, demonstrating that typological parameterization offers a path toward multilingual syntactic priors. Exploratory model-internal analyses are consistent with these findings at the levels of output probabilities, attention structure, as well as cross-lingual representation alignment.

1 Introduction

Language models (LMs) remain far less data-efficient than human learners, particularly for syntactic phenomena that depend on hierarchical rather than linear structure (Warstadt et al., 2023). Children converge on such knowledge from comparatively little linguistic input, whereas LMs typically require orders of magnitude more text to reach comparable performance (Warstadt and Bowman,

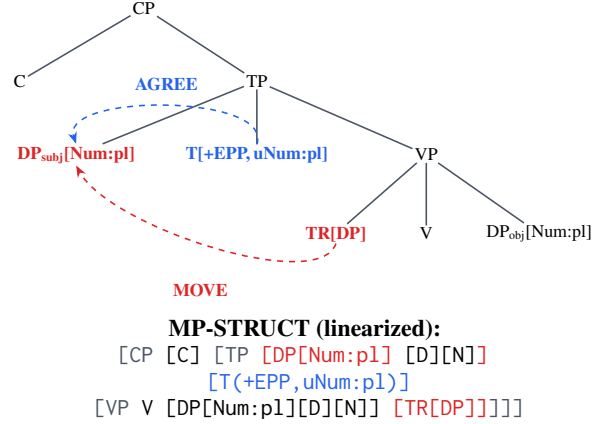


Figure 1: Example MP-STRUCT derivation, reproduced from Mita et al. (2026). **MERGE** builds the hierarchical vP/TP/CP skeleton; **AGREE** matches the subject’s Num feature to T ($uNum \leftarrow Num:pl$); **MOVE** raises the subject DP to Spec-TP (EPP), leaving a trace $TR[DP]$ at its base position. In this work, we parametrize these operations to be more typologically flexible.

2022). Under the Universal Grammar (UG) hypothesis, human language acquisition is guided by an innate bias—the Language Acquisition Device (LAD; Chomsky, 1965), grounded in UG—that constrains the learner’s hypothesis space toward natural-language-like structure from the outset. If a comparable structural bias could be injected into LMs prior to natural language training, the acquisition of more complex, hierarchical syntactic structures should be facilitated.

Pre-pretraining (PPT) instantiates this strategy: LMs are first trained on formal languages that encode abstract hierarchical structure (e.g., “{ (} [)]”), before proceeding to standard natural language pretraining, where genuine syntactic structure must be acquired—injecting a structural inductive bias for subsequent learning (Papadimitriou and Jurafsky, 2023; Hu et al., 2025). Extending prior heuristic approaches, MP-STRUCT (Mita et al., 2026) provides a PPT paradigm, which incorporates the fundamental UG-operations: MERGE

(hierarchical structure building), AGREE (feature checking), and MOVE (displacement) as shown in Figure 1. This added structure has been shown to improve the efficiency of English language acquisition. However, prior work has left open the questions of (1) which specific grammatical phenomena PPT improves or degrades, and (2) whether PPT is effective on languages other than English.

By definition, UG should be language-universal, and its inductive bias should accordingly support the acquisition of any natural language. Since MP-STRUCT is designed to operationalize UG’s core operations, its inductive bias should likewise extend to languages beyond English. However, the current implementation of MP-STRUCT contains English-specific design choices, such as obligatory wh-fronting (e.g., Chinese instead allows wh-in-situ), and head-initial word order (e.g., compared to Dutch’s underlying SOV). Such typological differences within UG are characterized by the Principles and Parameters (P&P) framework (Chomsky, 1981) as *parameters*, which delimit the range of variation possible across languages, while *principles* must be satisfied by all languages.

In this work, we provide, to the best of our knowledge, the first subtask-level study on PPT’s effects across typologically different languages, complementing prior multilingual PPT work (Papadimitriou and Jurafsky, 2020, 2023) with a fine-grained phenomenon-level analysis for additional languages, as well as methods for making PPT itself more typologically flexible following the P&P framework. Towards these goals, we provide the following contributions:

1. **Empirically differentiating P&P.** Using PPT to operationalize different levels of *principles* and *parameters* (§3), we identify that a balance between the two is crucial to achieve performance gains across BLiMP (Warstadt et al., 2020) subtasks over the non-PPT baseline, while avoiding language-specific biases, such as wh-island degradation (−10.3 pp) driven by English-centric parametrization (§4).
2. **Typologically flexible PPT.** By making the *parameters* within P&P language-adaptable and typologically coherent, we introduce the BCC framework for PPT (§3.3), which eliminates wh-island degradation while maintaining significant overall improvements, better balancing cross-lingual performance.
3. **Exploratory analysis of model-internal changes driven by PPT.** Model-internal differences suggest how PPT induces structurally-oriented attention and reshapes cross-lingual representations. BCC exhibits balanced cross-lingual convergence, whereas fixed parametrization over-converges toward English, and a lack of parametrization produces low language differentiation (§5).

2 Related Work

2.1 Universal Grammar and Minimalism

Under the UG/LAD hypothesis (Chomsky, 1965), human language acquisition is constrained by an innate linguistic bias, grounded in UG. P&P theory (Chomsky, 1981) decomposes UG into two kinds of constraints in order to characterize typological variation: *principles*, which every language must satisfy without exception, and *parameters*, which delimit the range of variation possible across languages, with each language fixing its own values.

We treat the core structure-building operations of the Minimalist Program (Chomsky, 1995)—MERGE, AGREE, and MOVE—as instantiating the principles of P&P, since every language is assumed to build hierarchical phrases, check feature agreement, and displace constituents in some form, without exception.¹ Parameters, in contrast, instantiate typological variation along dimensions such as wh-movement, head directionality, and richness of agreement, each admitting a small set of possible values that a given language fixes. The Borer–Chomsky Conjecture (BCC; Borer, 1984; Chomsky, 2001) further holds that cross-linguistic variation reduces to the feature composition of functional heads, providing the theoretical foundation for our parameterization approach (§3.3).

2.2 Pre-Pretraining

Pre-Pretraining (PPT; Papadimitriou and Jurafsky, 2023) injects a structural inductive bias into LMs through short-term pretraining on formal languages. Specifically, *k*-shuffle Dyck, a context-sensitive relaxation of *k*-Dyck, captures complex dependency

¹These operations reformulate, rather than replicate, the specific principles proposed in earlier P&P work (Chomsky, 1981); we adopt this operational mapping for computational tractability rather than as a claim about their historical equivalence. Note also that the *availability* of an operation (e.g., AGREE) is considered universal even where its morphological realization is minimal or absent in a given language (e.g., Chinese); we return to this distinction in §3.2.

	eng	nld	zho	MS
Wh-mvt	oblig.	oblig.	in-situ	oblig.
Word order	SVO	V2/SOV	SVO	head-init.
Agreement	limited	rich	none	limited

Table 1: Target languages and typological contrasts with MP-STRUCT (MS).

structures within the computational limits of Transformers, providing an effective structural bias (Hu et al., 2025). MP-STRUCT (Mita et al., 2026) instantiates the principles identified above (MERGE, AGREE, MOVE; §3.2) as bracketed formal languages (see Figure 1). Prior work shows that attention head subnetworks acquired during PPT persist through pretraining (Hu et al., 2025) and that low dependency resolution ambiguity determines PPT effectiveness (Mita et al., 2026). To extend PPT more effectively to multilingual settings, this work studies the gap between PPT without any UG principles (i.e., Dyck languages), and PPT with fixed English-type UG parameters (i.e., MP-STRUCT).

3 Methods

We evaluate three PPT paradigms with increasing cross-lingual flexibility: k -shuffle Dyck, MP-STRUCT, and BCC—across English, Dutch, and Chinese (Table 1). These languages differ in wh-movement (obligatory fronting in English/Dutch, Chomsky, 1977, vs. wh-in-situ in Chinese, Huang, 1982) and word order (SVO in English/Chinese vs. V2/SOV in Dutch, den Besten, 1983).

3.1 k -shuffle Dyck: Inducing General Language Structure

k -shuffle Dyck (SD; Hu et al., 2025) encodes interleaved bracket structures (e.g., “{ (} []”) with cross-serial dependencies, providing a language-agnostic structural bias. It serves as a structure-only control to validate whether any performance gains or losses on specific grammatical phenomena are a result of UG instantiation or a general characteristic of PPT. Following the main experimental settings of Hu et al. (2025), we adopt $k = 64$ with shuffling, which provides a more robust structural bias than non-shuffled, strictly nested alternatives, and a high-enough value of k as to better approximate the cross-serial dependencies found in natural language.

Head	Feature	Values	Langs
v, V	complement dir.	init./final	en,zh/nl
C	[+wh] strength	strong/weak	en,nl/zh
C	V-attraction	act./inact.	nl/en,zh
T	φ -features	rich/lim./ $\emptyset$	nl/en/zh
Top	EPP	act./inact.	zh/en,nl

Table 2: Typological parameters in BCC.

Algorithm 1 BCC Data Generation

Notation: $\mathcal{L}$: lexicon, θ : typological parameters, c : clause syntax tree, S : output sequence, $\oplus$: string concatenation, $L_{\max}$: sequence length limit.

```

1: Input: lexicon  $\mathcal{L}$ , target length  $L_{\max}$ 
2: Output: token sequence  $S$ 
3:  $S \leftarrow \emptyset$ 
4: while  $|S| < L_{\max}$  do
5:   Step 1: Clause-level Parameter Sampling
6:    $\theta \leftarrow \text{SAMPLEPROFILE}(\{\text{en}, \text{nl}, \text{zh}\})$ 
7:   // Determines the cross-linguistic syntactic variation for the current clause
8:   Step 2: Clause Generation and Linearization
9:    $c \leftarrow \text{GENERATECLAUSE}(\mathcal{L}, \theta)$ 
10:  // Applies MERGE, AGREE, MOVE following parameters  $\theta$ 
11:   $c_{\text{lin}} \leftarrow \text{LINEARIZE}(c)$ 
12:  Step 3: Concatenation (Code-switching Simulation)
13:   $S \leftarrow S \oplus c_{\text{lin}}$ 
14:  // Appends the linearized clause to simulate typological code-switching
15: end while
16: return  $S$ 

```

3.2 MP-STRUCT: Operationalizing Universal Grammar Principles

MP-STRUCT (MS; Mita et al., 2026) generates syntactic derivations encoding Minimalist Program operations (Figure 1): non-terminal labels (CP, TP, vP), functional heads (C, T, v), and φ -features are combined using the principle operations MERGE (hierarchical structure building; Chomsky, 1995), AGREE (feature checking; Chomsky, 2000), and MOVE (displacement; Chomsky, 1981, 1995), as well as movement traces (TRACE). The resulting derivations are linearized into bracketed sequences (e.g., [CP MERGE [C [+wh]] [TP ...]]) with language-independent abstract lexical tokens, so the LM learns syntactic operations before exposure to natural language. Beyond the principle operations (MERGE, AGREE, MOVE), the design includes *parameters* that govern how these operations are applied. Following its original work, we fix these parameters to follow English-like syntactic rules: (**wh-movement**) wh-elements obligatorily move to Spec-CP (English-type fronting); wh-in-situ is never generated; (**head directional-**

ity) always head-initial; V2 and head-final orders are not generated; (**agreement features**) partially encoded; rich agreement systems and the absence of agreement are not represented.

3.3 BCC: Parameterizing Typological Features

To address MS’s English-centric parameter settings, we propose MP-STRUCT BCC (BCC), which introduces explicit typological parameters following the Borer–Chomsky Conjecture (Borer, 1984; Chomsky, 2001). For each generated clause, functional head features are sampled from typological profiles (Table 2). We introduce five parameters: First, **complement directionality** on the v and V heads determines whether verbs precede (English, Chinese) or follow their objects (Dutch). Second, **[+wh] strength** on the C head dictates whether wh-movement is obligatory (English, Dutch) or whether wh-in-situ is allowed (Chinese). Third, **V-attraction** on the C head controls verb-second (V2) word order, which is active in Dutch but inactive in English and Chinese. Fourth, **φ -features** on the T head specify the richness of morphosyntactic agreement, ranging from rich (Dutch) to limited (English) or absent (Chinese). Finally, **EPP** on the Top head regulates topic-prominence, which is active in Chinese but inactive in English and Dutch.

Each PPT sequence concatenates multiple independent clauses whose typological parameters are dynamically sampled from plausible language profiles (English, Dutch, Chinese; Algorithm 1). Since lexical items and structural operations are identical to MS, any observed differences are attributable solely to typological parameterization.

4 Experiments

4.1 Setup

Training. Following the BabyLM 2026 Multilingual track (Jumelet et al., 2025; Choshen et al., 2026), we train GPT-2-small (Radford et al., 2019; 124M parameters) on the official en_nld_zho_equal pre-training data (100M words $\times$ 10 epochs). PPT conditions observe 500 initial steps of synthetic language pre-pretraining ($\text{lr} = 5 \times 10^{-4}$, warmup = 100). Four conditions are tested with three random seeds each: **Baseline (BL)** without PPT; **Shuffle Dyck (SD)** with k -shuffle Dyck ($k=64$) PPT (§3.1); **MP-STRUCT (MS)** with English-type parameter-fixed PPT (§3.2); **BCC** with typological parameterization (§3.3). All

	eng	nld	zho	avg
BL	.700 $\pm$.001	.790 $\pm$.005	.752 $\pm$.004	.747 $\pm$.003
SD	.707 $\pm$.003	.792 $\pm$.001	.762 $\pm$.002	.754 $\pm$.002*
MS	.710 $\pm$.005	.794 $\pm$.005	.759 $\pm$.008	.755 $\pm$.006*
BCC	.708 $\pm$.007	.795 $\pm$.003	.760 $\pm$.004**	.754 $\pm$.005*

Table 3: Group-level BLiMP accuracy across all 4 conditions (mean $\pm$ std, 3 seeds). Bold: best per column. * $p < .05$; ** $p < .01$ vs BL (paired t , $n=3$). Paired mean differences vs BL: SD +0.6 pp; MS +0.7 pp; BCC +0.7 pp (95% CIs overlap but all exclude zero).

conditions share identical model initializations, natural-language data ordering, and training schedule per seed; only the PPT phase differs.

Evaluation. All models are evaluated on BabyLM 2026’s multilingual BLiMP setup: 67 English subtasks from BabyLM-BLiMP (Warstadt et al., 2020); 22 Dutch subtasks from BLiMP-NL (Suijkerbuijk et al., 2025); and 118 Chinese subtasks from ZhoBLiMP (Liu et al., 2026)—totaling 207 evaluated subtasks. Each subtask uses minimal pairs to test grammaticality judgments. Benchmark accuracy is reported as mean $\pm$ stddev (3 seeds), with paired t -tests and 95% CIs.² should be interpreted cautiously as descriptive summaries rather than population-level inferential evidence. For win/loss rates across subtask-level results, each subtask’s 3-seed mean difference Δ_{mean} is classified as Win ($\Delta \geq 2$ pp), Loss ($\Delta \leq -2$ pp), or Tie, with 95% binomial CIs (Wilson score). The 2 pp threshold exceeds typical cross-seed variance (mean within-condition std ≈ 0.4 pp), and we find conclusions to remain stable across ± 1 to ± 5 pp thresholds (Table 7, Appendix A).

4.2 Benchmark-level Results

All PPT conditions significantly improve the 3-language average over BL (Table 3). MS shows the largest English improvement, while SD and BCC show stronger gains on Chinese. BCC is **not strictly superior to MS** ($p = .217$ for the BCC–MS difference), with each condition winning on different subtasks.

4.3 Subtask-level Results

Roughly half of subtasks show no meaningful change (Table 4). Among those with $|\Delta| \geq 2$ pp,

²Due to the sample size $n=3$, all effect sizes and significance values

	#Sub	W	L	T	W/(W+L)
<i>SD vs BL</i>					
eng	67	18	10	39	64.3%
nld	22	2	3	17	40.0%
zho	118	37	23	58	61.7%
Total	207	57	36	114	61.3%
<i>MS vs BL</i>					
eng	67	21	5	41	80.8%
nld	22	2	2	18	50.0%
zho	118	36	26	56	58.1%
Total	207	59	33	115	64.1%
<i>BCC vs BL</i>					
eng	67	17	8	42	68.0%
nld	22	4	2	16	66.7%
zho	118	35	28	55	55.6%
Total	207	56	38	113	59.6%

Table 4: Win/Loss analysis ($|\Delta_{\text{mean}}| \geq 2$ pp, 3-seed means). SD and MS both improve overall, but with different per-language profiles. BCC shows the most balanced per-language win rates. 95% CIs: SD [51.1–70.6%]; MS [53.9–73.2%]; BCC [49.5–68.9%].

all three PPT conditions favor wins. **PPT acts selectively on specific phenomena:** MS achieves the highest English win rate (80.8%) but the lowest Dutch rate (50.0%), hinting at an English-specific bias outlined below. BCC shows the most balanced performance profile across languages (66.7% median vs. 61.7% for BL and 58.1% for MS), consistent with its typologically diverse parametrization.

Language-bias in MS. Larger improvements involve language-specific structural phenomena: in Chinese, BEI construction and quantifier scope; in English, *wh_obj_gap* and *adjunct_island* (Table 5). All PPT conditions improve on these subtasks, with SD showing the largest *npi_renhe* gain (+22.1 pp), thus reflecting the structural inductive bias operating universally.

However, a striking dissociation emerges in wh-related subtasks (Table 6). MS improves accuracy where wh-movement occurs but degrades *wh_island* substantially. Critically, **SD shows no wh-island degradation**. This dissociation suggests that the degradation is not caused by PPT *per se*, but is consistent with MP-STRUCT’s obligatory wh-fronting parameter (this being one parameter that differs from MS to BCC; §6). A parallel pattern appears in Dutch, where MS degrades *wh_movement_restrictions* while SD shows no change (Table 8), though the Dutch subtask differs in item construction from the English *wh_island*, so this cross-lingual parallel is suggestive rather than exact. Chinese question-formation subtasks

Subtask	Lang	SD Δ	MS Δ	BCC Δ
<i>Top improvements</i>				
BEI_deletion	zho	+11.6	+28.4	+28.9
superlative_quant_2	zho	+16.2	+26.3	+12.3
npi_renhe_wh_subj	zho	+22.1	+8.3	+17.8
question_V_not_VP_1	zho	+9.0	+8.7	+15.4
<i>Key regressions</i>				
wh_island	eng	−0.3	−10.3	+0.9
q_daodi_A_not_A	zho	−11.4	−17.7	−5.8
renhe_no_sup_neg	zho	+4.2	+14.8	−10.4

Table 5: Selected top changes vs. BL (Δ in pp). Subtasks are selected post-hoc by largest $|\Delta|$ across any condition (top improvements) and by relevance to wh-phenomena (key regressions); no a priori selection was applied. SD shows the largest *npi_renhe* improvement (+22.1) but no *wh_island* degradation (−0.3), consistent with parameter-free structural bias.

Subtask	Structure	BL	SD Δ	MS Δ	BCC Δ
wh_obj_gap	wh-mvt (gap)	.708	+6.7	+3.5	+7.5
adjunct_island	adjunct extr.	.752	+2.6	+5.7	+4.7
wh_island	wh prohibited	.672	−0.3	−10.3	+0.9
wh_vs_that	wh vs. that	.251	+4.9	+4.5	+1.3

Table 6: Wh-related subtask effects across all conditions (Δ from BL in pp). SD shows no *wh_island* degradation (−0.3); MS severely degrades where wh-movement is prohibited (−10.3); BCC successfully eliminates this degradation (+0.9).

(31 subtasks including A-not-A, *nandao*, topicalization) show no systematic MS degradation, consistent with wh-in-situ being unaffected by the obligatory-fronting bias.

Table 8 (Appendix B) summarizes this cross-lingual pattern. Structural bias improves performance across all languages for both SD and MS (UG *principles*), whereas the MOVE-derived bias specific to MS is skewed toward English (*parameters*). Within the tested typological space, the P&P distinction thus maps onto computational PPT.

BCC eliminates wh-island degradation (Table 6) while maintaining overall gains (Table 3). SD avoids degradation by not encoding wh-movement; BCC avoids it by encoding diverse typological settings. BCC also yields the largest *wh_obj_gap* improvement among the three PPT conditions, suggesting that diverse movement patterns promote a more general wh-gap detection mechanism. However, BCC, MS, and SD produce different subtask-level trade-offs, so parameterization alone does not yield a strictly superior design.

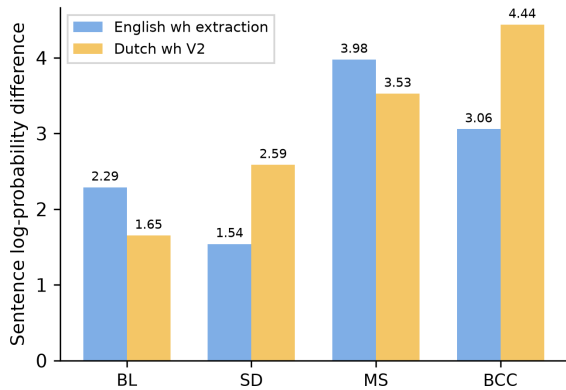


Figure 2: Wh-preference margin (correct/incorrect log-probability difference). MS is largest on English wh (3.98); BCC is largest on Dutch wh V2 (4.44). SD shows the lowest English margin (1.54), consistent with its lack of wh-movement encoding. Chinese is omitted because wh-in-situ does not involve overt wh-extraction.

5 Exploratory Model-internal Analysis

To better understand how the different PPT paradigms affect model-internal processing of multilingual inputs relative to the non-PPT baseline, we next analyze differences at the level of output probabilities (§5.1), intermediate parameters (§5.2), and representational alignment (§5.3).

5.1 Wh-preference Margins

For wh-movement, Figure 2 revealed significant differences depending on the PPT paradigm. Measuring the magnitude of this effect via the log-probability margin between correct and incorrect wh-extraction items, we find that MS shows the highest margin for correct versus incorrect English cases (3.98), consistent with English-type specialization. SD shows the *lowest* English margin (1.54)—which places it below even BL (2.29)—consistent with its lack of any explicit MOVE operation. Nonetheless, SD suffers no wh-island degradation, suggesting that a weaker but unbiased wh-representation is preferable to a strong but English-skewed one. BCC has an English margin of 3.06, placing it closest to MS, while also exhibiting higher margins for other languages, including the highest Dutch wh V2 margin (4.44), reflecting **more balanced cross-lingual wh-preferences**.

5.2 Attention Head Specialization

At the level of model parameters, we classify each attention head as either **structurally-oriented (S)** or **proximity-oriented (P)** based on its correlation

with dependency structures versus linear token distance, following the intuition of Hu et al. (2025). Concretely, for each head we compute the Pearson correlation between its attention weights and a linear-proximity matrix $\frac{1}{1+|i-j|}$ over three English sentences containing relative clauses and PP modifiers. Heads with $r < 0.5$ are classified as S, the remainder as P.³ The S/P ratio = $\frac{|S|}{|S|+|P|}$ quantifies the strength of the structural bias induced by each PPT paradigm.

All PPT models exceed BL’s S/P ratio (0.847). SD achieves the highest ratio (0.997), followed by MS (0.975) and BCC (0.925). Since the classification is based on English stimuli, the ordering also reflects how strongly each paradigm specializes toward English-like structural patterns: SD’s near-ceiling ratio indicates an extreme structural bias with little proximity-based attention, while BCC’s comparatively moderate ratio is consistent with its typologically diverse training producing more balanced attention patterns that are not narrowly tuned to any single language’s dependency structure.

5.3 Cross-Lingual Representation Alignment

At a broader multilingual level, we next investigate how PPT affects a model’s overall cross-lingual representational alignment.

Embedding Alignment. To measure embedding alignment across languages, we first group token embeddings by language. Concretely, for each token in the models’ vocabulary, we trace which language split of the pre-training data (en_nld_zho_equal) it originates from, assigning tokens to multiple languages if they occur in more than one split (e.g., digits). Since the tokenizer is trained on the same corpus, this allows us to group token IDs and their corresponding embeddings by language. Similarly, this means that if a token only occurs in one language (e.g., most Chinese characters), their embeddings will also only be directly affected by pre-training data in that language.

The resulting groupings in Figure 3 show that PPT substantially increases cross-lingual embedding similarity, with MS and BCC exhibiting the highest gains. These alignment increases are particularly pronounced for non-English and multilingual embeddings, jumping from 0.00–0.05 (BL) to 0.10–0.18 (PPT).

³Note that low proximity-oriented correlation may also reflect diffuse, non-local attention; the S/P ratio should therefore be interpreted as a coarse proxy for structural biases.

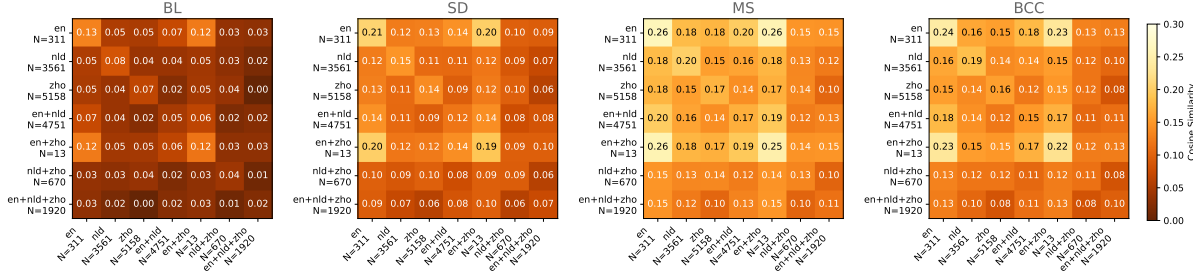


Figure 3: Mean cosine similarities across token embeddings from different language groups (multi-language groups if a token occurs in more than one language data subset). N indicates the number of embeddings per group, summing to the total vocabulary size of 16,384. PPT models show higher cross-lingual similarity than the baseline.

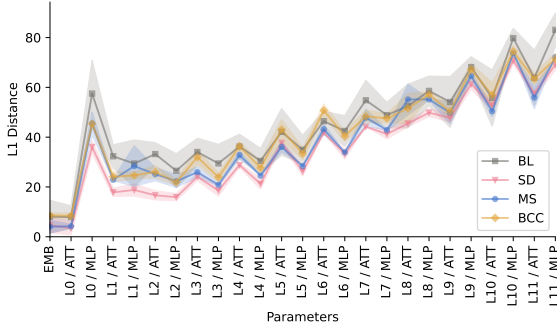


Figure 4: Cross-lingual L1 residual distance across embeddings (EMB), as well as layer-wise attention (ATT) and feed-forward blocks (MLP).

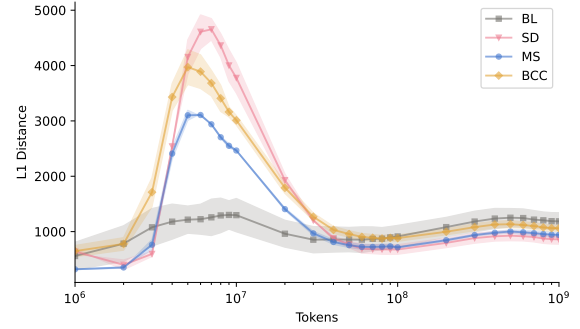


Figure 5: Cross-lingual total L1 residual distance across training (indicated by the number of observed tokens).

Residual Distance. As a proxy for cross-lingual alignment across model components, we measure L1 distances between language-specific representations in each model’s residual stream (details in Appendix D) using 1k parallel sentences from FLORES (NLLB Team, 2024). Figure 4 shows that PPT models exhibit lower residual distance (i.e., higher cross-lingual alignment) in the earlier half of layers (L0–L5), before following a trajectory of increasing separation similar to the baseline.

Interestingly, PPT models initially show significantly *larger* distances after PPT (~ 10 M tokens), but this reverses beyond 100M tokens (Figure 5). A Wilcoxon signed-rank test at $p < 0.05$ with Holm-Bonferroni correction on the final ordering of residual distances, $\text{SD} \approx \text{MS} < \text{BCC} < \text{BL}$, shows that SD and MS exhibit significantly lower typological diversity (i.e., convergent representations) compared to BL’s higher cross-lingual separation. Meanwhile, BCC’s intermediate level of alignment reflects typologically balanced representations that resist under and over-convergence.

6 Discussion

6.1 Principles vs. Parameters in PPT

In the P&P framework (§2), *principles* are constraints every language must satisfy, while *parameters* are constraints that delimit the range of possible variation across languages. Within the tested typological space, the P&P distinction between “principles” and “parameters” maps onto PPT. Language-agnostic structural effects improve performance consistently across languages (SD and MS; §4.2), while parameter-level MOVE effects are specific to MS and skewed toward English (§4.3). Since SD encodes hierarchical structure without any typological parameters, yet produces cross-lingual improvements without language bias, this isolates the source of the bias in MS’ parameter settings rather than in PPT itself.

Formal language design is thus **not theoretically neutral**: parameter settings affect the induced bias, and should be treated as explicit design variables. BCC addresses this issue by sampling parameters from plausible typological profiles rather than fixing them, eliminating wh-island degradation while preserving overall cross-lingual gains (§4.3).

Ruling Out Alternative Explanations. To establish the parameter fixation hypothesis, we consider four alternative explanations for the wh-island degradation. First, the degradation is unlikely due to insufficient PPT steps, as other wh-improvements (wh_obj_gap, adjunct_island) clearly emerge within the same 500-step budget. Second, it cannot be attributed solely to the intrinsic complexity of island structures. BL already achieves substantial accuracy on wh_island (Table 6), indicating that these constraints are learnable from natural language alone without PPT. Third, the hypothesis that PPT in general disrupts island learning is inconsistent with the SD control, which undergoes the same 500 steps of structural PPT yet shows no wh-island degradation (Table 6). Finally, the absence of island structures in the formal language is complementary rather than alternative to our hypothesis: the “always-move” and “never prohibit” biases may jointly contribute to the degradation. That is, while English itself exhibits both obligatory wh-fronting and strict island constraints, MS’s formal language encodes only the former without the latter, creating an asymmetric bias that favors extraction without learning when to block it. The converging evidence from SD and BCC is consistent with parameter fixation as the primary source of the observed bias, though full causal identification requires targeted single-parameter ablation.

6.2 Subtask-level Multilingual Evaluation Is Essential

Behind the modest benchmark-level improvement (§4.2) lie phenomenon-specific subtask-level changes spanning a much wider range (§4.3, Table 5), including the wh-island degradation, which would have been invisible without subtask-level analysis. These results highlight the need to balance language-agnostic SD and language-specific features of MS, while reinforcing the practical implication that typologically plausible inductive biases (BCC) can lead to more balanced cross-lingual performance, while universal structural biases (SD) are preferable when the target language characteristics are unknown. Future work should likely explore how many parameter profiles can be included in BCC-style PPT, to test the limits of inducing biases for many more languages simultaneously. To surface these language-specific differences during evaluation, we thus recommend: (1) standardized reporting of subtask-level Win/Loss distributions

and cross-lingual comparisons; (2) explicit documentation of typological parameter settings in formal language design; (3) inclusion of typologically diverse evaluation languages in PPT studies.

6.3 Model-internals Reflect PPT Paradigms

The exploratory analyses in §5 suggest that each PPT paradigm’s characteristics are reflected in the model-internal differences induced by each method: parametrization choices surface in output probabilities, PPT reshapes the embedding space that later training builds on, and these early effects persist through subsequent learning phases.

Probabilities Reflect Parametrization. Across our different levels of typological parametrization, §5.1 showed how SD’s low English wh-margin reflects its typologically unbiased parametrization, while MS’ fixed English-type parameters manifest as an excessive English wh-preference. BCC, by contrast, maintains balanced wh-preferences across languages, matching its superior performance on wh-island resolution. Practically, the fact that SD leads to more balanced results than MS suggests that no typological bias (i.e., randomly sampled parametrization) may be preferable when the target typology is unknown.

PPT Improves Embedding Alignment. Beyond output probabilities, PPT also leaves a mark on the model’s internal representations. Cross-lingual representational alignment has proven to be crucial for balanced multilingual performance (Basirat, 2025; Körner et al., 2026), yet in practice, multilingual embeddings are often skewed towards English (Wendler et al., 2024; Harrasse et al., 2025). We observe this same skew: despite our relatively small three-language setup and equal token ratios across languages in the pre-training data, multilingual embeddings still show slightly higher alignment with English (§5.3). It is therefore notable that PPT substantially increases cross-lingual embedding similarity for non-English embeddings compared to standard pre-training on the same data (from 0.00 to 0.18), even though it only encodes abstract structural information. This indicates that even a small amount of PPT promotes closer cross-lingual alignment of the embedding space that forms later in training. Encoding UG-principles appears to further aid this process, as both MS and BCC show additional alignment gains compared to SD alone.

Early PPT Affects Later Learning Phases.

This improved alignment, however, does not emerge immediately. §5.3 shows that PPT models have lower cross-lingual residual stream distance at the end of pretraining, but that the training dynamics leading to this result are non-trivial: all PPT models initially show dramatically larger cross-lingual distance than BL in the early phases of pre-training, right after PPT (Figure 5). Beyond this phase, cross-lingual distance falls and remains consistently below the baseline from around 100M tokens onwards. Notably, this stage of pre-training coincides with a previously reported “critical learning phase” (Müller-Eberstein et al., 2023; Chen et al., 2024), during which LMs acquire most of their fundamental syntactic capabilities. Together with the high level of structural specialization of attention heads in PPT models (§5.2), this indicates that the inductive biases introduced by PPT establish a structure that shapes training dynamics well beyond their initial phase.

7 Conclusion

Through the first subtask-level multilingual analysis of UG-inspired PPT across English, Dutch, and Chinese, we find evidence that MP-STRUCT’s inductive bias based on UG principles is *partially universal but contains language-specific bias*. Structure-derived effects are cross-linguistically consistent, whereas parameter-derived effects—most prominently obligatory wh-fronting—are associated with an English-centric bias that degrades wh-restriction constructions in both English and Dutch. BCC-based parameterization eliminates this degradation while preserving overall gains. Exploratory analyses of model internals are consistent with the hypothesis that PPT’s benefits are mechanistically decomposable: structural attention bias (present in all PPT conditions including *k*-shuffle Dyck) is separable from cross-lingual alignment (requiring language-specific tokens), with each paradigm’s parametrization choices surfacing in both output probabilities and representational geometry. Within the tested typological space, the P&P distinction between principles and parameters thus maps onto computational PPT, indicating that for genuinely universal PPT, principled typological parameter design and subtask-level multilingual evaluation are essential.

Limitations

Typological Diversity. Only three languages are tested. SOV languages (e.g., Japanese), polysynthetic languages, and other typological profiles require further investigation. BLiMP phenomenon coverage is also benchmark-dependent; while Dutch *wh_movement_restrictions* tests subjacency-type constraints on wh-extraction (§4.3), it differs from the English *wh_island* subtask in item construction and coverage, so the cross-lingual parallel is suggestive rather than exact. Currently, dedicated island constraint benchmarks for Dutch and Chinese do not exist, limiting the scope of direct comparison.

Experiment Scale. Three seeds provide limited statistical power for individual subtask significance testing. We use only GPT-2 small (124M parameters) with 100M words of training data; whether the observed effects persist, diminish, or amplify with larger models and more data is untested, and small-scale artifacts cannot be ruled out. PPT uses 500 steps, and operates under BabyLM constraints; optimal step counts remain unexplored.

Single-factor Ablations. The attribution of wh-island degradation to wh-fronting fixation is supported by converging evidence (SD control + BCC elimination) but not by targeted single-parameter ablation; BCC changes five parameters simultaneously, so the individual contribution of each parameter remains unquantified. BCC also introduces increased structural diversity alongside typological parameterization; a random-parameterization baseline would help disentangle these factors. Additionally, while BCC uses the same vocabulary and operations as MS, its clause-level parameter sampling may alter the marginal distributions of token frequencies, sequence lengths, and tree depths, meaning that some observed differences could partly reflect distributional rather than purely typological changes. The strongest bias evidence comes from English wh-islands with supporting evidence from Dutch wh-movement restrictions; broader claims about “language bias” should be interpreted with this scope in mind. Internal analyses are exploratory seed-0 observations without cross-seed replication. BCC is not strictly superior to MS, indicating that parameterization alone does not yield optimal PPT design.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24H00087, JST BOOST Grant Number JPMJBY24B2, and JST CREST Grant Number JPMJCR2565. MME was supported by the Carlsberg Foundation, grant CF-25-0624. MY was supported by the Nippon Foundation HUMAI Program.

References

- Ali Basirat. 2025. [Multilingual learning strategies in multilingual large language models](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 507–518, Suzhuo, China. Association for Computational Linguistics.
- Hagit Borer. 1984. *Parametric Syntax: Case Studies in Semitic and Romance Languages*. Foris, Dordrecht.
- Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. [Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.
- Noam Chomsky. 1977. On wh-movement. In Peter Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*, pages 71–132. Academic Press, New York.
- Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.
- Noam Chomsky. 1995. *The Minimalist Program*. MIT Press, Cambridge, MA.
- Noam Chomsky. 2000. Minimalist inquiries: The framework. In Roger Martin, David Michaels, and Juan Uriagereka, editors, *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, pages 89–155. MIT Press, Cambridge, MA.
- Noam Chomsky. 2001. Derivation by phase. In Michael Kenstowicz, editor, *Ken Hale: A Life in Language*, pages 1–52. MIT Press, Cambridge, MA.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. 2026. [BabyLM turns 4 and goes multilingual: Call for papers for the 2026 babyLM workshop](#). *Preprint*, arXiv:2602.20092.
- Hans den Besten. 1983. [On the interaction of root transformations and lexical deletive rules](#). In Werner Abraham, editor, *On the Formal Syntax of the Westgermania*, pages 47–131. John Benjamins, Amsterdam.
- Abir Harrasse, Florent Draye, Punya Syon Pandey, Zhi-jing Jin, and Bernhard Schölkopf. 2025. [Tracing multilingual representations in llms with cross-layer transcoders](#). *Preprint*, arXiv:2511.10840.
- Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. [Between circuits and Chomsky: Pre-pretraining on formal languages impacts linguistic biases](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709, Vienna, Austria. Association for Computational Linguistics.
- C.T. James Huang. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Ph.D. thesis, Massachusetts Institute of Technology.
- Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrana, Francesca Padovani, Francois Meyer, Hai Hu, Julien Etxaniz, Laurent Prévot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, and 7 others. 2025. [BabyBabelLM: A multilingual benchmark of developmentally plausible training data](#). *Preprint*, arXiv:2510.10159.
- Felicia Körner, Max Müller-Eberstein, Anna Korhonen, and Barbara Plank. 2026. [When meanings meet: Investigating the emergence and quality of shared concept spaces during multilingual language model training](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3149–3169, Rabat, Morocco. Association for Computational Linguistics.
- Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2026. [A systematic assessment of language models with linguistic minimal pairs in chinese](#). *Transactions of the Association for Computational Linguistics*, 14:755–771.
- Masato Mita, Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2026. [Language acquisition device in large language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19564–19577, San Diego, California, United States. Association for Computational Linguistics.
- Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. [Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.
- NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Isabel Papadimitriou and Dan Jurafsky. 2020. [Learning Music Helps You Read: Using transfer to study linguistic structure in language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2023. [Injecting structural hints: Using language models to study inductive biases in language learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8402–8413, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. [BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation](#). *Computational Linguistics*, 51(4):1267–1301.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *Algebraic structures in natural language*, 17.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. [Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus](#). *Preprint*, arXiv:2301.11796.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do llamas work in English? on the latent language of multilingual transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

A Win/Loss Threshold Sensitivity

Table 7 tests whether the Win/Loss conclusions in §4.3 depend on the 2 pp threshold used to classify subtasks. For SD and MS, the W/(W+L) ratio remains within 3 pp of the ± 2 pp value across all thresholds, confirming that the overall positive effect is robust. MS is particularly stable (64.1% at ± 2 –5 pp), suggesting that its wins tend to be large in magnitude. BCC shows a slight downward trend at the most stringent threshold (± 5 pp: 56.0%), indicating that a larger share of its wins are moderate

in size; this is consistent with BCC’s balanced but individually smaller per-subtask gains compared to MS’s more concentrated improvements.

Threshold	W	L	T	W/(W+L)
<i>SD vs BL</i>				
± 1 pp	89	55	63	61.8%
± 2 pp	57	36	114	61.3%
± 3 pp	44	31	132	58.7%
± 5 pp	28	17	162	62.2%
<i>MS vs BL</i>				
± 1 pp	86	51	70	62.8%
± 2 pp	59	33	115	64.1%
± 3 pp	41	23	143	64.1%
± 5 pp	25	14	168	64.1%
<i>BCC vs BL</i>				
± 1 pp	82	51	74	61.7%
± 2 pp	56	38	113	59.6%
± 3 pp	44	29	134	60.3%
± 5 pp	28	22	157	56.0%

Table 7: Win/Loss sensitivity to threshold choice. The Win/(Win+Loss) ratio remains stable across thresholds from ± 1 to ± 5 pp for all three PPT conditions, indicating that the conclusions are not sensitive to the specific 2 pp threshold.

B Effect Patterns by Language

Table 8 summarizes how PPT effects decompose across languages. Three patterns emerge: (1) structural improvements from MERGE-derived bias are universal across all conditions and languages; (2) wh-restriction degradation is specific to MS and observed in both English (−10.3 pp) and Dutch (−3.4 pp), consistent with its fixed obligatory-fronting parameter; (3) BCC eliminates this degradation (eng +0.9 pp) while maintaining balanced group-level gains across all three languages (+0.5–0.8 pp), though Dutch wh_movement_restrictions shows a modest decline (−2.1 pp). SD and BCC converge in showing no wh-island degradation despite differing mechanisms: SD avoids it by not encoding wh-movement at all, while BCC avoids it by encoding typologically diverse movement patterns.

C MS-BCC Generator Details

The MS-BCC generator samples typological parameters per clause following the profiles in Table 2, then applies MERGE, AGREE, and MOVE with the sampled settings to produce linearized sequences. Clauses are concatenated with code-switching across language profiles (Algorithm 1).

Effect	Cond.	eng	nld	zho
Structural $\uparrow$ (universal)	SD, MS, BCC	✓	✓	✓
Wh-detection $\uparrow$ (partial)	MS, BCC	✓	Lim.	N/A
Wh-restrict. $\downarrow$ (MS-spec.)	MS	-10.3	-3.4	—
Wh-restrict. (no degrad.)	SD	-0.3	-0.2	—
Wh-restrict. (no degrad.)	BCC	+0.9	-2.1	—
Group Δ (pp)	SD	+0.7	+0.2	+0.9
Group Δ (pp)	MS	+1.1	+0.4	+0.7
Group Δ (pp)	BCC	+0.8	+0.5	+0.8

Table 8: Effect patterns by language and condition. Structural improvement is universal across all PPT conditions. Wh-restriction degradation is MS-specific; SD and BCC show no systematic degradation. BCC maintains balanced group-level gains across all languages.

The vocabulary consists of the same abstract lexical tokens used in MP-STRUCT (**w1**, **w2**, etc.), ensuring that any performance differences between MS and BCC are attributable to parametrization rather than lexical content.

D Residual Stream Distances

To compute the representational distance between our three target languages (English, Dutch, Chinese), we extract the residual stream at 25 locations in the model: embeddings (EMB), attention blocks (ATT), and feed-forward components (MLP). As a source of parallel data, which allows us to control for language independently of semantic content, we use 1k sentences per language from the development data of the FLORES corpus (NLLB Team, 2024). Each sentence representation is computed from the mean over its tokens. We then compute the L1 distance between matching sentences across all language pairs (excluding self-similarity), taking the mean over all sentence distances.