BLiSS 1.0: Evaluating Bilingual Learner Competence in Second Language Small Language Models

Yuan Gao* **Suchir Salhan*** **Andrew Caines**
Paula Buttery **Weiwei Sun**
ALTA Institute Department of Computer Science & Technology, University of Cambridge

Abstract

To bridge the gap between performance-oriented benchmarks and the evaluation of cognitively-inspired models, we introduce **BLiSS 1.0**, a Benchmark of Learner Interlingual Syntactic Structure. Our benchmark operationalizes a new paradigm of selective tolerance, testing if a model finds a naturalistic learner error more plausible than a matched, artificial error within the same sentence. Constructed from over 2.8 million naturalistic learner sentences, BLiSS provides 136,867 controlled triplets (corrected, learner, artificial) for this purpose. Experiments on a diverse suite of models demonstrate that selective tolerance is a distinct capability from standard grammaticality, with performance clustering strongly by training paradigm. This validates BLiSS as a robust tool for measuring how different training objectives impact a model's alignment with the systematic patterns of human language acquisition.

BLiSS on HuggingFace (BLiSS 1.0 Dataset and Pretrained Models)

Training Code Open-Sourced on GitHub

1 Introduction

There is a growing interest in the NLP community in developing models that are not just powerful, but also cognitively inspired—that is, models which aim to reflect the processes of human language acquisition. Current evaluation benchmarks for language models are overwhelmingly performance-oriented, centering around grammaticality tests, adherence to standard grammar, and task performance (e.g., BLiMP Warstadt et al. (2020) and GLUE Wang et al. (2018)). While these measures are informative in evaluating linguistic competence, the core question for cognitively inspired modeling is different: do our systems exhibit the kinds of behaviors that emerge in human acquisition? For models that aim to be cognitively plausible, we need a complementary, acquisition-focused perspective, one that inspects how grammar competence is organized and learned.

This evaluation gap is particularly important for models of Second Language Acquisition (SLA), which we refer to as L2LMs (Aoyama and Schneider, 2024). A central characteristic of the SLA process is the production of systematic 'errors'. These deviations are not random noise, but rather structured evidence of the learner's developing internal grammar, or "interlanguage" (Corder, 2015; Selinker, 1972). For a model to be truly 'learner-like', it must be sensitive to these specific, structured patterns observed in real human data.

To address this, we propose a new paradigm built on a key assumption: the systematicity of learner errors is tied to both the type of error and its specific locus within a sentence. This assumption, therefore, theorizes that moving an attested error to a different, albeit plausible, location renders it less naturalistic and less human-like. This approach, which uses an error's locus as a test of naturalness, is inspired by similar methodologies for evaluating complex linguistic phenomena (Sterner and Teufel, 2025). This allows us to test a model's selective tolerance: its ability to penalize a naturalistic human error less severely than a contrived artificial-locus error of the same type.

We introduce the **Benchmark of Learner Interlingual Syntactic Structure (BLiSS 1.0)**, a large-scale evaluation dataset on a model's alignment with naturalistic language learner patterns, offering a new dimension of evaluating acquisition-focused models. BLiSS is built upon three of the largest English learner corpora available: the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2014), the Write & Improve Corpus (W&I) (Nicholls et al., 2024), and

*Corresponding Authors: yg386@cam.ac.uk, sas245@cam.ac.uk

Proceedings of the First BabyLM Workshop, pages 160–174

the First Certificate in English (FCE) dataset (Yannakoudakis et al., 2011).

(1) U:DET (Unnecessary determiner)

a. There are a lot of benefits when we play sports.

b. *There are a lot of benefits when we play **the** sports.

c. **There are a lot of benefits when **the** we play sports.

The core of the BLiSS evaluation is the triplet, a controlled comparison between: a corrected sentence, the original sentence with one error from a learner, and a version with an artificially-generated error of the same error type, as shown in (1). From an initial pool of over 2.8 million raw learner sentence-corrected sentence pairs, we systematically generate a matched artificial-locus for each valid, single-edit grammatical deviation. After a rigorous multi-stage validation pipeline, BLiSS comprises 136,867 high-quality evaluation triplets. Each triplet is accompanied by rich metadata, including learner L1, proficiency level, and error type, as illustrated in Figure 1.

In this paper, we deploy BLiSS to evaluate a diverse suite of models, from large bilingual LLMs to acquisition-inspired L2LMs. Our results yield two key findings. First, we demonstrate that selective tolerance is a distinct capability from standard grammaticality; high performance on BLiMP does not guarantee high performance on BLiSS. Second, we show that model performance on BLiSS clusters strongly by training paradigm, validating it as a tool for measuring how different architectures and training objectives impact a model's alignment with the systematic patterns of human learner language.

2 Related Work

2.1 Second Language Acquisition-Inspired Language Models (L2LMs)

We use L2LMs to denote cognitively inspired models of L2 acquisition (Aoyama and Schneider, 2024). Early work examined transfer—training on an L1 then an L2—and the role of typological distance (Yadavalli et al., 2023; Oba et al., 2023), while later studies add cognitive priors (e.g., alignment to learner reading times; preserving L1 knowledge to probe the Critical Period Hypothesis) and compare sequential vs. mixed L1/L2 exposure (Aoyama and Schneider, 2024; Clahsen and Felser, 2006; Constantinescu et al.; Lenneberg, 1967; Kirkpatrick et al., 2017; Arnett et al., 2025). Given heterogeneous architectures and pretraining corpora (from learner-like data to web-scale sources such as CC-100; Wenzek et al., 2020), a common benchmark tied to learner behavior is needed (Salhan et al., 2024; Arnett et al., 2025).

```
{
  "learnerID": "8421",
  "L1": "Vietnamese",
  "cefr": "C1",
  "topic": "play  sports ",
  "corrected": "There are a lot of
      benefits when we play sports .",
  "learner error": "There are a lot of
       benefits when we play the
      sports .",
  "artificial error": "There are a lot
       of benefits when the we play
      sports .",

  "errant_edits": [{
    "type": "U:DET",
    "o_str": "the",
    "c_str": ""
  }],
  "all_error_types": [
    "U:DET"
  ]
}
```

Figure 1: An example BLiSS triplet illustrating an Unnecessary Determiner (U:DET) error. The original learner sentence contains an unnecessary determiner "the", which is removed in the corrected sentence. Artificially-generated errors of the same type allow controlled evaluation of model preferences.

2.2 Learner Corpora and Error Profiling

Large-scale learner corpora provide an important empirical basis for modeling and evaluating L2 learner behavior. The Write & Improve Corpus 2024 (Nicholls et al., 2024) contains learner essays with Common European Framework of Reference for Languages (CEFR) annotations and corresponding error-labeled corrections. The essays were submitted by users of the 'Write & Improve' writing practice platform[1]. W&I uses ERRANT (Bryant et al., 2017) to annotate errors in learner essays automatically. ERRANT annotations classify errors as replacements (R), missing (M) or unnecessary (U) and assign a specific tag (e.g., M:ADJ means the text omits an adjective). A full table of ERRANT error codes are included in *Appendix C* for reference. The EF-Cambridge Open Lan-

[1] https://writeandimprove.com/

guage Database (EFCAMDAT) (Geertzen et al., 2014) offers a large collection of learner texts annotated with proficiency levels and metadata on learner nationality. Note that the proficiency levels in EFCAMDAT relate to difficulty level attained by users of the 'EF Englishtown' platform (now 'EF English Live'[2]), rather than human ratings of the texts themselves, but this information serves as a good proxy for learner proficiency. The W&I-2024 corpus has a wider range of L1s compared to other publically-available learner corpora, like the FCE subset of the Cambridge Learner Corpus (Yannakoudakis et al., 2011). There are other error-annotated English learner corpora, such as NUCLE (Dahlmeier et al., 2013), JFLEG (Napoles et al., 2017) and Lang-8 (Mizumoto et al., 2012; Tajiri et al., 2012), but are respectively age/language restricted; use fluency rewrite rather than minimal grammatical edits; and have user-generated corrections (Nicholls et al., 2024).

3 BLiSS 1.0

3.1 Motivation

The BLiSS 1.0 benchmark is a large-scale evaluation suite composed of controlled triplets designed to test a model's selective tolerance for naturalistic learner production errors. The evaluation framework for BLiSS is designed to move beyond evaluations of the formal competence of a Language Model (e.g., using broad-coverage datasets like BLiMP (Warstadt et al., 2020)) to evaluate the **alignment of a language model with second language acquisition**. BLiSS builds upon previous attempts to extend acquisition-inspired evaluation frameworks for Language Models (e.g., Evanson et al. (2023)) beyond first language acquisition.

BLiSS 1.0 focuses on naturalistic production errors in learner corpora. The BLiSS 1.0 benchmark is designed to evaluate how closely a language model's outputs align with patterns observed in second language (L2) learners, particularly in terms of grammatical errors. While it is true that individual learner errors do not imply that a majority of learners would make the same mistake in a given sentence, BLiSS focuses on systematic tendencies in learner language rather than absolute probabilities of specific errors. By aggregating errors across millions of sentence-correction pairs from multiple learner corpora, BLiSS captures the distributional patterns of learner errors that are prevalent in naturalistic L2 production. BLiSS does not encourage models to prefer errors, but rather tests alignment with learner error patterns.

This approach addresses a critical limitation of traditional LM evaluation benchmarks (e.g., BLiMP), which primarily assess formal grammatical competence. Such benchmarks assume that the model should always prefer grammatical sentences, but human learners – especially in L2 acquisition –frequently produce systematic errors that reveal underlying acquisition stages, interlanguage phenomena, or L1 transfer effects. BLiSS thus extends evaluation beyond formal competence, providing a framework to test whether models *selectively tolerate or reproduce error patterns* in ways that resemble human learners.

Concretely, we develop BLiSS to enable the study of:

1. **Error-type sensitivity:** Whether language models recognize and react differently to common L2 errors (e.g., determiner omission, verb tense errors).

2. **Position awareness:** By generating artificial errors at positions distinct from the learner's original error, we can test if language models are sensitive to the locus of grammatical deviations, not just their existence.

3. **Learner-informed evaluation:** Leveraging metadata such as L1 background and proficiency level that are available in large-scale corpora allows analysis of model behavior in the context of typologically diverse learner populations.

While BLiSS does not imply that *all learners* would produce a given error, it provides a systematically sampled and validated set of errors that represents frequent phenomena in learner production (Alexopoulou et al., 2015; Le Bruyn and Paquot, 2021; Crossley and Kyle, 2022; Alexopoulou et al., 2022). This makes BLiSS a meaningful benchmark for probing the alignment of language models with human L2 acquisition patterns, without conflating individual idiosyncrasies with population-level tendencies.

3.2 Data: Source Corpora

The credibility and naturalistic grounding of the BLiSS benchmark stem from its foundation in

[2] https://englishlive.ef.com/

large-scale, naturalistic learner data. We aggregate sentence-correction pairs from three of the most widely-used English learner corpora, ensuring our benchmark reflects genuine learner behaviour in communicative contexts.

- The EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2014): A very large collection of over 1 million learner texts from an online English learning platform. Texts are annotated with metadata including learner nationality and proficiency levels mapped to the Common European Framework of Reference for Languages (CEFR).

- The Write & Improve (W&I) Corpus (Nicholls et al., 2024): A dataset of learner essays submitted to an online writing feedback tool. It is richly annotated with CEFR levels (A1–C2) and explicit learner L1 labels, providing high-quality metadata for fine-grained analysis.

- The First Certificate in English (FCE) Dataset (Yannakoudakis et al., 2011): A well-known subset of the Cambridge Learner Corpus containing essays from an official language proficiency exam. This provides a valuable sample of argumentative, exam-style writing from a diverse set of L1 backgrounds.

Collectively, these corpora provide a massive pool of over 2.8 million raw sentence-correction pairs, forming the empirical starting point for our triplet construction pipeline, detailed in the following section.

Corpus	# Raw Pairs
EFCAMDAT	2,711,188
W&I	63,926
FCE	52,421
Total	2,827,535

Table 1: Summary of raw single-edit sentence-correction pairs from the source corpora.

3.3 Triplet Construction Pipeline

The construction of the BLiSS dataset follows a multi-stage pipeline designed to transform raw sentence-correction pairs from the source corpora into high-quality, validated triplets. The pipeline emphasizes grammatical precision, methodological transparency, and the atomization of errors to ensure each triplet tests a single distinct linguistic phenomenon.

Grammatical Error Classification and Filtering The process begins with a comprehensive error analysis of the raw sentence pairs using the ERRANT toolkit (Bryant et al., 2017). With these annotations, we first filtered out pairs containing only non-grammatical edits, such as spelling, punctuation, or capitalization changes.

Error Atomization We then atomized sentence pairs with multiple corrections using the ERRANT annotations as a guide. Each distinct grammatical edit within a multi-error sentence was isolated to create a new single-edit pair consisting of the corrected sentence and a version with just that one specific error. This process ensures that every triplet in the final dataset is anchored to exactly one grammatical deviation, allowing for a clean and targeted evaluation.

Rule-Based Artificial Error Generation The core of the pipeline is the generation of an artificial error for each single-edit pair. This rule-based system uses linguistic analysis and morphological generation, creating a new sentence that adheres to two fundamental constraints:

1. **Error Type Consistency:** the artificial error must mirror the grammatical operation of the human error. For example, a missing determiner (M:DET) in the learner sentence prompts the generation of a new sentence where a determiner is removed.

2. **Position Divergence:** The artificial error must be introduced at a different word position than the learner error. This ensures the model is being tested on its sensitivity to the error's locus, not merely its presence.

Multi-Stage Quality Validation To ensure the integrity of BLiSS, every generated triplet was subjected to a rigorous multi-stage validation filter. A triplet was only retained if it passed all of the following checks:

1. **Morphological Correctness:** All inflected words (e.g., verbs, nouns) generated by LemmInflect[3] must be valid English forms.

[3] https://github.com/bjascob/LemmInflect

2. **Triplet Uniqueness:** The artificial error sentence must be distinct from both the corrected sentence and the original learner error sentence.

3. **Error Type Confirmation:** Finally, we used ERRANT as a verifier. The generated artificial error, when compared to the corrected sentence, must be classified by ERRANT as having the same error type as the original human error.

This stringent validation process resulted in an overall success rate of 4.8%, yielding a final dataset of 136,867 high-quality triplets. The low success rate is a direct reflection of the strictness of our quality controls, ensuring that every item in BLiSS is a valid and non-ambiguous test case. A sample of 100 triplets was also manually reviewed, confirming a grammatical and positional accuracy rate of over 95%.

3.4 Dataset Composition

Following the rigorous construction and validation pipeline, the final BLiSS dataset comprises 136,867 high-quality triplets. The composition of the dataset reflects both the diversity of the source corpora and the targeted nature of our filtering process. As shown in Table 2, the majority of the final dataset (76.7%) is derived from the large-scale EFCAMDAT corpus, supplemented by high-quality and diverse data from the W&I and FCE corpora.

Corpus	Triplets	Percentage
EFCamDat	105,034	76.7%
Write & Improve	17,380	12.7%
FCE	14,453	10.6%
Total	136,867	100%

Table 2: BLiSS Composition

Error Type Distribution The dataset provides robust coverage across a range of core grammatical error categories that are common in second language acquisition. Table 3 details the distribution of the five most frequent error types, which collectively account for over 67% of the dataset.

Learner Demographics The rich metadata from the source corpora allows for detailed analysis across learner populations. Table 4 shows the distribution of the top five L1 backgrounds in the dataset.

Error Type	Count	Percentage
M:DET	26,008	19.0%
R:NOUN:NUM	21,149	15.5%
R:PREP	18,702	13.7%
U:DET	15,708	11.5%
R:VERB:TENSE	10,599	7.7%

Table 3: Distribution of the top 5 ERRANT error types in BLiSS.

The significant representation of typologically diverse languages such as Chinese, Japanese, and Arabic makes the benchmark particularly powerful for investigating L1 transfer effects.

L1 Background	Count	Percentage
Chinese	23,771	17.4%
Japanese	14,478	10.6%
Italian	11,918	8.7%
French	11,486	8.4%
Arabic	9,484	6.9%

Table 4: Distribution of the top 5 L1 backgrounds in BLiSS.

In terms of learner proficiency, BLiSS spans a wide range of the CEFR scale, from beginner (A1) to advanced (C2), as detailed in Figure 2. The dataset has substantial representation between the beginner and intermediate(A1 - B2) levels but significantly less at higher levels with only 25 triplets at the C2 level. This broad distribution is a key strength, enabling the study of how model behavior might differ when evaluated on errors typical of different proficiency levels.

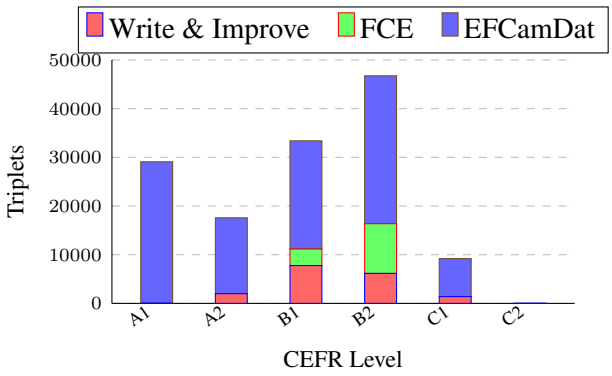

Figure 2: Distribution of CEFR proficiency levels in BLiSS by corpus (stacked triplet counts).

4 Evaluation

The core objective is to quantitatively measure a model's alignment to the naturalistic production errors produced by second language (L2) learners of English. To evaluate a model's selective tolerance, we introduce a set of complementary metrics that capture different aspects of its behavior. The **Learner Preference (LP)** metric provides a simple metric that measures whether the model prefers a human learner sentence over the corrected version, though a high LP could reflect either accurate simulation of learner tendencies or poor grammatical knowledge. To directly probe selective tolerance, **Human vs. Artificial Preference (HAP)** measures whether the model favors naturalistic learner errors over contrived, artificial errors, while **HAP-τ** is a stricter version that ensures the model's preference is meaningful and not just due to numerical noise. Finally, the **Strict Order (SO)** metric captures the most stringent behavior, requiring the model to rank all three sentences in the hypothesized order—corrected first, learner second, artificial last—indicating a balance between grammatical competence and nuanced sensitivity to L2 error patterns. Together, these metrics provide a multi-faceted view of whether a language model can recognize correct grammar, differentiates between plausible and implausible errors, and exhibits robust, cognitively plausible error sensitivity.

A model's preference for a sentence is quantified using token-normalized surprisal, measured in Bits Per Token (BPT), where low BPT indicates high plausibility under the model's learned distribution and high BPT signals a grammatical deviation. By computing BPT scores for each sentence in a BLiSS triplet—including the corrected sentence, the human learner error, and an artificially generated error—we can evaluate not only whether a model recognizes correct grammar, but also whether it differentiates between naturalistic learner errors and contrived mistakes. These BPT scores underpin the three evaluation metrics in the BLiSS framework.

We recommend that **each metric should be reported separately**, as they provide complementary insights: LEARNER PREFERENCE (LP) captures general grammatical preference, HUMAN V ARTIFICIAL PREFERENCE (HAP and HAP-τ) metrics assess selective tolerance, and STRICT ORDER (SO) evaluates the full hypothesized ranking. Combining these metrics into a single score would obscure these distinctions and reduce the interpretability of a language model's behavior on L2 error patterns.

4.1 Scoring Signal

We quantify a model's preference for a given sentence s by its token-normalized surprisal, measured in Bits Per Token (BPT). This is calculated as the negative log-likelihood of the sentence, normalized by the number of tokens.

$$BPT(s) = -\frac{1}{|s|}\sum_{t=1}^{|s|} log_2 p(w_t|w_{<t})$$

where $|s|$ is the number of tokens in the sentence and $p(w_t|w_{<t})$ is the probability assigned by the model to token w_t given the proceeding context.

From a cognitive perspective, surprisal is often used as a proxy for processing effort. A sentence that aligns with a model's learned grammatical and statistical patterns will have low surprisal (low BPT), indicating it is highly plausible under the model's distribution. Conversely, a sentence with a grammatical deviation will have high surprisal (high BPT). This allows us to use BPT as a 'plausibility score' to measure the model's preference for each of the three sentences in a BLiSS triplet.

For each item in the BLiSS dataset, we apply this scoring signal to three sentences in the triplet, which we will formally denote as: s_{corr} (corrected), s_{lrn} (learner), and s_{art} (artificial).

4.2 Evaluation Metrics

We present a suite of metrics designed to provide a multi-faceted view of a model's behavior. These metrics are organized around two key concepts: a baseline measure of simple learner preference and our primary measures of selective tolerance.

Baseline Metric: Learner Preference (LP) We provide this metric as a minimal-pair evaluation. We define Learner Preference (LP) as the proportion of items where the model prefers the learner sentence over the corrected version: $BPT(s_{lrn}) < BPT(s_{corr})$. The motivation for LP is that for certain applications, such as simulating learner output, a model might be intentionally designed to reproduce learner errors. However, LP is inherently ambiguous, as a high score could also simply reflect poor grammatical knowledge. We therefore use it as a diagnostic baseline.

Selective Tolerance Metrics To overcome the ambiguity of LP, our primary metrics are designed to probe a model's selective tolerance directly. The desired behavior for a cognitively plausible model is twofold: it should, first and foremost, still recognize and prefer correct grammar, yet it should also differentiate between the plausibility of different types of errors. Specifically, it should find a naturalistic, systematic learner error to be more plausible (less surprising) than a contrived, artificial error. According to this principle, the ideal ordering of preferences for any triplet should be the corrected sentence, followed by the learner sentence, and finally the artificial sentence. Following this, we present three primary metrics that quantify a model's adherence to this behavior.

1. **SO (Strict Order):** This is the most stringent metric. It measures the proportion of the items where the model's preferences follow the full, hypothesized order of plausibility: $BPT(s_{corr}) < BPT(s_{lrn}) < BPT(s_{art})$. A high SO score is the strongest evidence that a model successfully balances grammatical competence with a nuanced sensitivity to interlanguage.

2. **HAP (Human vs. Artificial Preference):** This metric isolates the central test of selective tolerance by measuring the proportion of items where the model simply prefers the human error over the artificial one: $BPT(s_{\text{lrn}}) < BPT(s_{\text{art}})$. HAP allows us to credit a model for correctly distinguishing between the two error types.

3. **HAP-τ (Robust HAP):** A stricter version of HAP, this metric requires the BPT difference between the artificial and learner sentences to exceed a small positive buffer τ: $BPT(s_{\text{art}}) - BPT(s_{\text{lrn}}) > \tau$. This ensures the model's preference is confident and meaningful, rather than an artifact of numerical noise.

5 Models

We evaluate a diverse range of models on the BLiSS benchmark. The models are grouped into four distinct families, ordered by their increasing degree of specialization for second language acquisition (SLA). This progression allows us to systematically investigate how training data, architecture, and SLA-inspired objectives influence a model's capacity for selective tolerance.

Standard Bilingual LLMs This family serves as our baseline, representing powerful, general-purpose models that have not been specifically designed to model learner language or the acquisition process. These are large language models trained on massive corpora of standard, native-speaker text in two languages. Their training objective is to model fluent, grammatical language, not the intermediate stages of learning. We include Bilingual-GPT-NeoX-4B[4] (Japanese–English) (Zhao et al.; Sawada et al., 2024), CroissantLLM[5] (Faysse et al., 2024) (French–English), and MAP-Neo-7B[6] (Zhang et al., 2024) (Chinese–English).

Bilingual BabyLMs This family represents models that are 'acquisition-inspired' in their data scale but are not explicitly designed for SLA. These are smaller models trained from scratch on developmentally plausible, child-directed speech (CDS) in two languages. While they model the acquisition of language, they are primarily simultaneous bilingual first language acquisition (BFLA), not successive L2 learning. We evaluate publicly released models (Jumelet et al., forthcoming)[7] trained from scratch on 10M words of CDS in English plus one other language (Persian, German, Indonesian, Japanese, Dutch, or Chinese).

Acquisition-Inspired L2 Models This family includes models that explicitly incorporate principles from SLA research into their design. They are designed to simulate the process of an L1 speaker learning an L2, often through sequential training regimes or other architectural priors that model transfer. SLABERT (Yadavalli et al., 2023) follows the *Test for Inductive Bias via Language Model Transfer* (TILT; Pauls and Klein, 2012): pretrain on age-ordered CDS in L1 (French, Polish, Indonesian, Japanese), then fine-tune on English adult-directed speech with all parameters frozen except embeddings. B-GPT (Arnett et al., 2025) is trained with *sequential* exposure (L1 then L2) or *simultaneous* exposure (L1+L2 mixed), here evaluated only for English L2 with Dutch or Spanish L1.

Learner-Trained Models This final family represents models that are directly exposed to learner language during training. Instead of learning from

[4] https://huggingface.co/rinna/bilingual-gpt-neox-4b

[5] https://huggingface.co/croissantllm/CroissantLLMBase

[6] https://map-neo.github.io/

[7] https://huggingface.co/BabyLM-community

native text and hoping learner-like patterns emerge, we train these models on the same kind of data used in our benchmark. We train several GPT-2 medium models from scratch on learner-produced English essays from the Cambridge Learner Corpus (CLC) and EFCAMDAT. To ensure fairness in evaluation, these models are only evaluated on the W&I slice of BLiSS.

6 Results

Table 5 presents BLiSS scores for all evaluated models. Our analysis shows three primary findings that validate the BLiSS benchmark as a tool for measuring a distinct, acquisition-related dimension of model behavior.

An analysis of the model families in Table 5 reveals distinct performance profiles. The Bilingual LLMs and B-GPT models emerge as the strongest performers on our primary selective tolerance metrics. Both families form tight clusters with high HAP scores (≈66-67%) and, notably, the highest Strict Order (SO) scores (≈55-57%). This indicates a robust ability to correctly rank the full triplet.

The Bilingual BabyLM models also perform significantly above chance, but with lower SO scores (≈35-44%), suggesting a weaker, though still present, signal of selective tolerance. A consistent and important trend among these three families is a statistically significant increase in performance on their respective L1 data slices, providing strong evidence that they have internalized L1-dependent transfer patterns and validating BLiSS as a tool for probing these fine-grained behaviors.

In contrast, the SLABERT and Learner-Trained models show a different and less successful profile. Their very high Learner Preference (LP) scores (often >50%) are coupled with poor performance on our primary selective tolerance metrics, particularly Strict Order. This suggests that their training may have made them indiscriminately accepting of learner-like forms, hindering their ability to distinguish between plausible human errors and implausible artificial ones.

6.1 BLiSS vs. BLiMP

To visualize the relationship between a model's BLiSS and BLiMP scores, Figure 3 plots the HAP score against the BLiMP score for each evaluated mode, colour-coded by the model family. The plot demonstrates several key insights into the nature of the BLiSS benchmark and the capabilities of

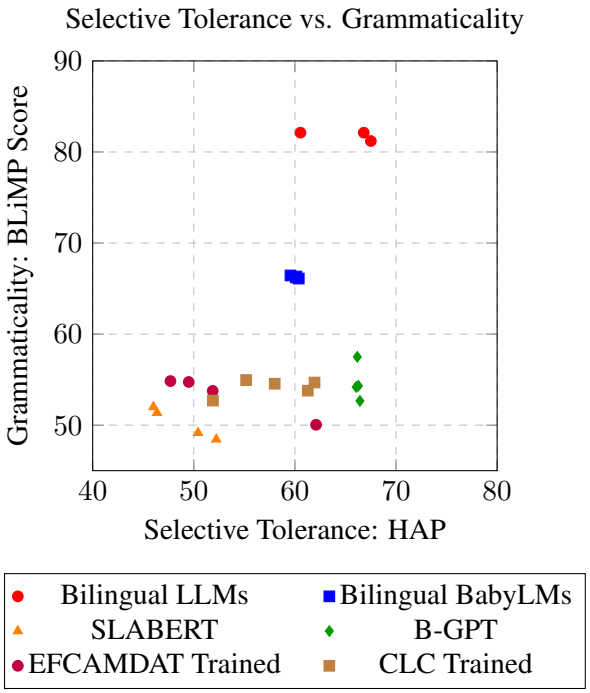

Figure 3: Selective tolerance (HAP score) versus grammaticality (BLiMP score) across all evaluated models. Each point represents a model, colour-coded by its training family.

different model architectures.

A striking observation is that models from the same training family form tight clusters. For example, the large Bilingual LLMs occupy a distinct region in the top-right of the plot, while the B-GPT and SLABERT models form their own clear groups. This consistency is a powerful validation of our methodology; it suggests that BLiSS is successfully capturing a stable signal that is reflective of the underlying training paradigm, rather than just idiosyncratic model behavior. The two learner-trained families (EFCAMDAT and CLC) show slightly more internal variance, which is expected, as the primary differentiating factor within those families is the training data.

Another pattern we observe from the plot is the clear lack of a strong positive correlation between the two metrics. High performance on BLiMP does not guarantee high performance on BLiSS and vice-versa. The large Bilingual LLMs, for instance, excel at both. However, other models achieve strong selective tolerance without top-tier grammaticality. The B-GPT models are a prime example.

This demonstrates that BLiSS offers a complementary, second dimension for language model evaluation. It measures a distinct capability that

Model	BLiSS							BLiMP
	HAP		HAP@τ		SO		LP	
	Overall	L1	Overall	L1	Overall	L1		
Bilingual LLMs								
CroissantLLM	67.51*	81.76**	57.42*	71.70**	57.64*	54.09	12.84	81.20
Neox-4B	60.55*	78.26**	42.98*	62.32**	35.15*	1.88	16.27	82.12
MAP-Neo-7B	66.81*	77.12	58.14*	72.03**	56.05*	45.76**	14.14	82.12
Bilingual BabyLM models								
BBLM–DE	60.15*	76.92**	50.59*	66.67**	43.73*	33.33	18.80	66.32
BBLM–ZH	59.56*	72.88**	49.57*	66.95**	34.93*	38.14	20.44	66.44
BBLM–ID	60.08*	66.6	50.41*	62.96	43.66*	37.04	28.57	66.22
BBLM–FR	60.38*	79.25**	50.70*	67.92**	43.93*	44.03	13.71	66.10
SLABERT								
SLABERT–JP	50.42*	47.83	31.40*	27.54	16.58*	15.94	63.46	49.16
SLABERT–FR	52.22*	52.20	34.50*	33.96	16.20*	15.09	57.47	48.44
SLABERT–ID	46.36*	38.89	30.91*	25.93	15.40*	12.96	57.14	51.36
SLABERT–PL	46.01*	54.92**	32.99*	38.52	14.07*	18.85	57.57	52.00
B-GPT								
B-GPT-ES-SIM	66.43*	77.14**	56.48 *	62.14**	54.57*	50.00	12.99	52.66
B-GPT-ES-SEQ	66.06*	74.29**	56.15*	61.43**	55.06	48.57**	12.17	54.19
EFCAMDAT Trained								
LM–EF	51.87*	47.94**	36.94*	33.78**	15.23*	12.66**	39.51	53.76
Noise–EF	47.69*	46.26	28.47*	29.67	11.40*	11.33	41.95	54.84
Contr–EF	62.09*	62.24	48.02*	41.74	23.70*	18.03	69.51	50.04
Compl–EF	49.50*	56.23	44.75*	45.80	21.35*	19.54	40.73	54.74

Table 5: BLiSS metrics (HAP, HAP@τ, Strict Order, LP) with L1 and Overall subcolumns, alongside BLiMP grammaticality accuracy. An asterisk (*) indicates performance significantly above the 50% chance baseline ($p < 0.05$), while a double asterisk (**) on L1 scores indicates a statistically significant difference between the L1-specific and overall performance. See full result table in A.

is not captured by standard grammaticality benchmark alone.

7 Conclusion

As the BabyLM Challenge extends cognitively-inspired language modeling beyond English, there are methodological challenges in evaluating the formal competence of BabyLM-inspired L2LMs that are modeling second language or bilingual acquisition. To address this, we introduced BLiSS, a large-scale benchmark built on a new paradigm of selective tolerance. By evaluating models on controlled triplets (corrected, learner error, artificial error), BLiSS measures a model's ability to distinguish naturalistic human errors from contrived ones, disentangling sensitivity to learner patterns from general grammatical competence. Our experiments demonstrate that selective tolerance is a distinct capability from standard grammaticality, with performance clustering strongly by training paradigm and revealing sensitivity to L1-specific transfer effects. We hope that BLiSS will serve as both a benchmark and a research catalyst for developing L2 language models that better reflect the diversity and systematicity of human language acquisition.

Limitations

Several limitations should be considered when interpreting our results. First, BLiSS relies on

sentence-level corrections from learner corpora, which may not capture all aspects of learner language development. The benchmark focuses on grammatical and lexical errors but does not assess discourse-level phenomena, pragmatic competence, or other dimensions of L2 proficiency that extend beyond sentence boundaries. This imbalance may affect the reliability of conclusions about advanced learner behavior and limits our ability to study developmental trajectories at higher proficiency levels.

Specific L1 backgrounds and grammatical error types that were already infrequent in the source corpora become even more sparse in the final dataset. The low success rate of our generation process (4.8%) means that only the most common and structurally regular phenomena are represented at scale. This may limit the statistical power for fine-grained analyses on these lower-frequency L1-error combinations and means that our results are most representative of common error patterns.

Acknowledgments

With thanks to Laura Barbenel for proof-reading this manuscript. We thank the anonymous reviewers for their useful feedback and suggestions, which greatly improved the manuscript. This paper reports on work supported by Cambridge University Press & Assessment.

References

Theodora Alexopoulou, Jeroen Geertzen, Anna Korhonen, and Detmar Meurers. 2015. Exploring big educational learner corpora for sla research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, 1(1):96–129.

Theodora Alexopoulou, Detmar Meurers, and Akira Murakami. 2022. Big data in sla: Advances in methodology and analysis. In *The Routledge handbook of second language acquisition and technology*, pages 92–106. Routledge.

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling Nonnative Sentence Processing with L2 Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940. Association for Computational Linguistics.

Catherine Arnett, Tyler A. Chang, James A. Michaelov, and Benjamin Bergen. 2025. On the acquisition of shared grammatical representations in bilingual language models. In *Annual Meeting of the Association for Computational Linguistics*.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805. Association for Computational Linguistics.

Harald Clahsen and Claudia Felser. 2006. Continuity and shallow structures in language processing. *Applied Psycholinguistics*, 27(1):107–126.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. Investigating Critical Period Effects in Language Acquisition through Neural Language Models. 13:96–120.

Stephen Pit Corder. 2015. The significance of learners' errors. In *Error analysis*, pages 19–27. Routledge.

Scott A Crossley and Kristopher Kyle. 2022. 34 managing second language acquisition data with natural. *The Open Handbook of Linguistic Data Management*, page 411.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Linnea Evanson, Yair Lakretz, and Jean-Remi King. 2023. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Henrique Martins, Antoni Bigata Casademunt, François Yvon, André Martins, Gautier Viaud, C'eline Hudelot, and Pierre Colombo. 2024. CroissantLLM: A truly bilingual French-English language model. *ArXiv*, abs/2402.00786.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic linguistic annotation oflarge scale L2 databases: The EF-Cambridge Open Language Database(EFCamDat).

Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrama, Francesca Padovani, Francois Meyer, Hai Hu, Julen Etxaniz, Laurent Prevot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, and 7 others. forthcoming. BabyBabelLM: A multilingual benchmark of developmentally plausible training data. *Forthcoming*. Under review. Download PDF available online.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu,

Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Bert Le Bruyn and Magali Paquot. 2021. *Learner corpus research meets second language acquisition*. Cambridge University Press.

Eric H. Lenneberg. 1967. The biological foundations of language. *Hospital Practice*, 2(12):59–67.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872, Mumbai, India. The COLING 2012 Organizing Committee.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 174–188, Miami, FL, USA. Association for Computational Linguistics.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905. https://arxiv.org/abs/2404.01657.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

Igor Sterner and Simone Teufel. 2025. Minimal pair-based evaluation of code-switching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18575–18598, Vienna, Austria. Association for Computational Linguistics.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, and 26 others.

2024. MAP-Neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv: 2405.19327*.

Tianyu Zhao, Toshiaki Wakatsuki, Akio Kaga, Koh Mitsuda, and Kei Sawada. rinna/bilingual-gpt-neox-4b.

A Full Evaluation Results

Model	BLiSS							BLiMP
	HAP		HAP@τ		SO		LP	
	Overall	L1	Overall	L1	Overall	L1		
Bilingual LLMs								
CroissantLLM	67.51*	81.76**	57.42*	71.70**	57.64*	54.09	12.84	81.20
Neox-4B	60.55*	78.26**	42.98*	62.32**	35.15*	1.88	16.27	82.12
MAP-Neo-7B	66.81*	77.12	58.14*	72.03**	56.05*	45.76**	14.14	82.12
Bilingual BabyLM models								
BBLM–DE	60.15*	76.92**	50.59*	66.67**	43.73*	33.33	18.80	66.32
BBLM–ZH	59.56*	72.88**	49.57*	66.95**	34.93*	38.14	20.44	66.44
BBLM–ID	60.08*	66.6	50.41*	62.96	43.66*	37.04	28.57	66.22
BBLM–FR	60.38*	79.25**	50.70*	67.92**	43.93*	44.03	13.71	66.10
SLABERT								
SLABERT–JP	50.42*	47.83	31.40*	27.54	16.58*	15.94	63.46	49.16
SLABERT–FR	52.22*	52.20	34.50*	33.96	16.20*	15.09	57.47	48.44
SLABERT–ID	46.36*	38.89	30.91*	25.93	15.40*	12.96	57.14	51.36
SLABERT–PL	46.01*	54.92**	32.99*	38.52	14.07*	18.85	57.57	52.00
B-GPT								
B-GPT-ES-SIM	66.43*	77.14**	56.48 *	62.14**	54.57*	50.00	12.99	52.66
B-GPT-ES-SEQ	66.06*	74.29**	56.15*	61.43**	55.06	48.57**	12.17	54.19
EFCAMDAT Trained								
LM–EF	51.87*	47.94**	36.94*	33.78**	15.23*	12.66**	39.51	53.76
Noise–EF	47.69*	46.26	28.47*	29.67	11.40*	11.33	41.95	54.84
Contr–EF	62.09*	62.24	48.02*	41.74	23.70*	18.03	69.51	50.04
Compl–EF	49.50*	56.23	44.75*	45.80	21.35*	19.54	40.73	54.74
CLC Trained								
CLC–A1	61.94*	-	51.94*	-	28.86*	-	35.28	54.68
CLC–A2	58.00*	-	46.94*	-	23.74*	-	39.27	54.55
CLC–B1	61.27*	-	50.75*	-	29.62*	-	29.62	53.79
CLC–B2	55.18*	-	43.47*	-	21.93*	-	41.68	54.95
CLC–C1	51.89*	-	38.65*	-	18.48*	-	46.02	52.70

Table 6: BLiSS metrics (HAP, HAP@τ, Strict Order, LP) with L1 and Overall subcolumns, alongside BLiMP grammaticality accuracy. An asterisk (*) indicates performance significantly above the 50% chance baseline ($p < 0.05$), while a double asterisk (**) on L1 scores indicates a statistically significant difference between the L1-specific and overall performance.

B Learner-Trained Model Details

All models were trained using the HuggingFace `Trainer` API with the following configuration. Training ran for 10 epochs for CLC-trained models and 5 epochs for EFCAMDAT-trained models.

Parameter	**Value**
Seed	42
Block size	1024 tokens
Per-device batch size	2
Gradient acc. steps	8
Effective batch size	16
Learning rate	5×10^{-5}
Weight decay	0.1
Warmup steps	500
Logging steps	50
Max steps	-1 (full epochs)
Scheduler	cosine
Optimiser	AdamW
Mixed precision	`fp16`
Gradient checkpointing	Enabled
Save strategy	End of each epoch

Table 7: Training hyperparameters.

C ERRANT Annotation Scheme

All learner sentences in BLiSS are automatically annotated with ERRANT v3.0.0 to obtain token-level error labels.

Table 8: Complete list of valid error code combinations

Operation Tier	Type	Missing	Unnecessary	Replacement
Token Tier	Adjective	M:ADJ	U:ADJ	R:ADJ
	Adverb	M:ADV	U:ADV	R:ADV
	Conjunction	M:CONJ	U:CONJ	R:CONJ
	Determiner	M:DET	U:DET	R:DET
	Noun	M:NOUN	U:NOUN	R:NOUN
	Particle	M:PART	U:PART	R:PART
	Preposition	M:PREP	U:PREP	R:PREP
	Pronoun	M:PRON	U:PRON	R:PRON
	Punctuation	M:PUNCT	U:PUNCT	R:PUNCT
	Verb	M:VERB	U:VERB	R:VERB
	Other	M:CONTR	U:CONTR	R:CONTR
	Morphology	-	-	R:MORPH
	Orthography	-	-	R:ORTH
	Other	M:OTHER	U:OTHER	R:OTHER
	Spelling	-	-	R:SPELL
	Word Order	-	-	R:WO
Morphology Tier	Adjective Form	-	-	R:ADJ:FORM
	Noun Inflection	-	-	R:NOUN:INFL
	Noun Number	-	-	R:NOUN:NUM
	Noun Possessive	M:NOUN:POSS	U:NOUN:POSS	R:NOUN:POSS
	Verb Form	M:VERB:FORM	U:VERB:FORM	R:VERB:FORM
	Verb Inflection	-	-	R:VERB:INFL
	Verb Agreement	-	-	R:VERB:SVA
	Verb Tense	M:VERB:TENSE	U:VERB:TENSE	R:VERB:TENSE