Introducing Code-Switched Contexts to Cognitively-Inspired Bilingual Model Training

Zhuojing Huang **Luise Pohlmann** **Lisa Beinborn**
University of Göttingen, Germany
firstname.lastname@uni-goettingen.de

Abstract

During language acquisition, bilingual children are regularly exposed to code-switched input and use it as a cognitive scaffold to accelerate vocabulary growth and cross-linguistic syntactic mapping. In contrast, computational bilingual models are conventionally pretrained on interleaved monolingual corpora. While introducing synthetic code-switching during pretraining has become a promising strategy to enhance cross-lingual alignment and downstream performance, the structural and developmental parameters governing the success remain poorly understood. In this work, we investigate the efficiency of training with synthetic code-switched data across two typologically distinct language pairs by controlling two key variables: the structural location of code-switches and the dynamic switching rate across training stages. Our results show that training with code-switched data improves cross-lingual alignment for typologically close languages.[1]

1 Introduction

In many bilingual families, children are regularly exposed to code-switched input during language acquisition (Byers-Heinlein, 2013; Bail et al., 2015). Psycholinguistic findings suggest that child code-switching is not a result of confusion or failure (Meisel, 2019). Instead, it is seen as an attempt to communicate given limited linguistic resources (Paradis et al., 2000; Werker and Byers-Heinlein, 2008). By using code-switching as a scaffold for currently weaker language (Yow et al., 2018; Schneider et al., 2026), bilingual children actively draw on their linguistic resources from both languages to support lexical growth (Deuchar and Quay, 1994; Meisel, 2019; Gaskins et al., 2022) with the effect that they need fewer repetitions for novel word acquisition (Kaushanskaya et al., 2023). The frequency of language switches varies among bilingual children and over time, as it correlates with language proficiency (Kapantzoglou et al., 2021). With more stable proficiency in both languages, code-switching loses its role as a scaffold, and the *switch rate* decreases (Smolak et al., 2020; Kapantzoglou et al., 2021).

In contrast to random lexical substitution, the position in an utterance at which code-switching can occur is strongly constrained by structural principles. The *switch point* can be dictated by the structural frame of one of the languages (Myers-Scotton, 1993) or restricted to positions at which the word order of both languages aligns (Poplack, 1980). Bilingual children are sensitive to the code-switching patterns in the mostly spoken input they receive from their interlocutors (Comeau et al., 2003), and they may implicitly infer the syntactic constraints of both linguistic systems from these patterns of switching (Chouinard and Clark, 2003).

Although code-switched input seems beneficial for bilingual proficiency, the training data for bilingual language models usually consists of parallel corpora (e.g. Kim et al., 2019; Lin et al., 2025) or interleaved monolingual datasets (e.g. Conneau et al., 2020; Zheng et al., 2021). While this strategy is effective, it also provides an implicit signal about the language identity of each training example, resulting in cleaner language boundaries than those encountered by bilingual children. To enforce cross-lingual alignment, existing approaches typically include an explicit alignment signal by using a contrastive learning objective (Shen et al., 2025) or synthetic word-level alignments (Kautsar and Koto, 2025). These approaches primarily operate at the lexical level, failing to capture how children implicitly acquire bilingual lexicons and syntactic structure during code-switching.

Recent results indicate that code-switched data can make bilingual learning more efficient, improving cross-lingual representation alignment beyond standard parallel corpora (Wang et al., 2025),

[1] Code available at GitLab repository

Proceedings of the Second BabyLM Workshop, pages 805–822
October 29, 2026 ©2026 Association for Computational Linguistics

while also benefiting downstream instruction tuning (Asano et al., 2026) and reducing the need for massive monolingual pretraining corpora (Zeng et al., 2026). Importantly, code-switched data should not simply be a consequence of mixing languages: the direction and proportion of token insertions can substantially affect performance (Wang et al., 2025), while progressively reducing code-switched content toward monolingual input has also been shown to improve cross-language accuracy and alignment (Yoo et al., 2025).

Despite the promising results, existing work has only examined different factors of code-switched training separately, but not yet how the *switch point* and the *switch rate* jointly shape bilingual representation learning. Therefore, we hypothesize that structurally and developmentally valid code-switched input can facilitate the emergence of shared, syntax-aware bilingual representations without the need for explicit alignment. To test this hypothesis, we train four different code-switching variants across two language pairs with distinct typological distances: English–Dutch (`eng-nld`) and English–Chinese (`eng-zho`), carefully controlling both the *switch rate* and the *switch point*.

To evaluate our models under developmentally plausible constraints, we use the BabyLM training data for the multilingual track (Jumelet et al., 2026a) and the 2026 evaluation pipeline (Choshen et al., 2026), assessing the models' performance on linguistic knowledge and downstream tasks. We show that training on code-switched data generally improves cross-lingual alignment and downstream performance for `eng-nld` models. For `eng-zho` models, however, code-switched contexts negatively influence performance in Chinese. To further understand how training with code-switched data affects the learned representation, we examine the geometry of the embedding space and find that code-switching models lead to better cross-lingual alignments for `eng-nld` models.

2 Related Work

While human bilinguals naturally acquire shared semantic concepts through mixed-language input, current bilingual modeling often ignores this aspect. The following sections review the psycholinguistic principles governing human code-switching and present how these insights are currently being integrated into language modeling.

2.1 Code-Switching in Bilingual Children

During the language acquisition process, exposure to mixed-language input is a common feature of many bilingual environments (Genesee, 2008; Bail et al., 2015; Kremin et al., 2019; Huang, 2025). Despite this input variability, studies show that code-switching carries minimal processing cost in natural settings (Blanco-Elorrieta and Pylkkänen, 2017; Adamou and Shen, 2019), allowing bilingual children to achieve key linguistic milestones within the same overall time frame as monolingual peers (Hambly et al., 2013). Crucially, Bilson et al. (2015) suggest that learning a word in one language supports its acquisition in the other, presumably because the underlying semantic representation has already been established, allowing bilingual learners to map an additional lexical form onto an existing concept. As proficiency in both languages stabilizes, code-switching gradually loses its role as an essential scaffolding mechanism, leading to a corresponding increase in the monolingual utterance (Smolak et al., 2020; Kapantzoglou et al., 2021). Beyond lexical learning, code-switching during language acquisition is also governed by syntactic constraints. Myers-Scotton (1993) suggests that one "matrix" language constructs the morphosyntactic frame, restricting switches predominantly to open-class content words. Poplack (1980), on the other hand, argues that switches occur only at points where the surface word orders of both languages coincide, which opens up more switch points between closely related languages sharing similar syntactic structures.

2.2 Code-Switching in Bilingual Models

In language modeling, introducing code-switched contexts enhances cross-lingual representation alignment beyond standard parallel corpora (Wang et al., 2025), yielding benefits that extend to instruction-tuning (Asano et al., 2026) and enabling models to achieve strong performance without relying on massive monolingual pretraining corpora (Zeng et al., 2026). Code-switching has also been explored as a means of mitigating catastrophic forgetting in multilingual adaptation. By combining POS-based code-switching with a replay adapter strategy, K et al. (2026) shows that training with code-switched data can help preserve previously acquired knowledge while adapting large language models to low-resource languages. However, these advantages are heavily mediated by factors such

as data volume, language-pair similarity, and structural switching dynamics. For instance, Zeng et al. (2026) show that performance scales logarithmically with the volume of available monolingual data in English–Spanish models. Furthermore, structural mechanics such as insertion direction are critical. For example, in Chinese–English models, increasing Chinese token insertions into English yields continuous performance gains, whereas inserting English tokens into Chinese text is beneficial only up to a specific insertion percentage (Wang et al., 2025). To further examine these factors, Yoo et al. (2025) mirror the developmental reduction of human code-switching over time, showing that progressing from frequent intra- and inter-sentential switching to purely monolingual data improves cross-language accuracy and alignment.

Building on these insights, we compare intra-sentential code-switching strategies within a cognitively-inspired setup for bilingual models. Using strict data ceilings as established in the BabyLM challenge (Choshen et al., 2026), we evaluate whether different code-switching variants provide consistent benefits across both typologically close (eng–nld) and typologically distant (eng–zho) language pairs.

3 Methodology

We train and evaluate one baseline and four code-switching variants for each language pair. Code-switching variants use the same architecture as the baseline, varying only *switch rate* and *switch point*. All models are trained across three random seeds.

3.1 Baseline

Our baseline models are pretrained on a balanced bilingual corpus constructed from the BabyLM 2026 multilingual data collections (Jumelet et al., 2026a; Choshen et al., 2026), summing up to 100M tokens.[2] We adopt the GPT-2-small architecture optimized with a standard causal language modeling objective. A Byte-Pair Encoding (BPE) tokenizer with a vocabulary size of $16,384$ is trained on the unmodified bilingual corpus. We train each model for 10 epochs, with a total exposure of $\sim$1B tokens. Checkpoints are saved at logarithmic intervals, namely, 1M, 10M, 100M, up to 1B words.

[2]Following the BabyLM 2026 guidelines, the Dutch and Chinese sub-corpora are scaled by byte premium factors of 1.05 and 0.93, respectively (Arnett et al., 2024).

3.2 Code-Switching Models

We train four code-switching variants that differ in *switch rate* and *switch point* as shown in Table 1. These models are optimized on the same training data as the baseline, but the languages are mixed according to the respective switching condition.[3]

Model	Switch Rate	Switch Point
Baseline	✗	✗
Static	Fixed	Random
Curriculum	Gradual	Random
POS-Static	Fixed	POS-constrained
POS-Curriculum	Gradual	POS-constrained

Table 1: Overview of code-switching variants.

Switch Rate. The *switch rate* controls the proportion of words that are switched into another language. While Static models maintain a uniform global token density, i.e., targeting 10% switches overall, Curriculum models deliver this same aggregate density through a four-phase dynamic curriculum. Inspired by the correlation between increasing language proficiency and declining code-switching rates in bilingual children (Smolak et al., 2020; Kapantzoglou et al., 2021), the curriculum progressively reduces the data proportion for code-switching across four equal word-budget phases, i.e., 90% $\rightarrow$ 70% $\rightarrow$ 40% $\rightarrow$ 0% of the data are eligible from phase 1 $\rightarrow$ 4. To match the overall 10% target density, tokens within the active switching windows are sampled at a higher rate (~20%). Documents are reshuffled within each phase to ensure variable document exposure.

Switch Point. The *switch point* controls which words are eligible for substitution. Rather than relying on unconstrained random substitution, the POS constrained models only switch at certain word classes. Our setup targets open-class words: Nouns, Verbs, and Adjectives (Urbaneja, 2020; Azuma and Meier, 1997; Prasad and Rao, 2011), motivated by switching constraints observed in human bilingual utterances (Myers-Scotton, 1993) and findings in language models (K et al., 2026). This setting aims to minimize grammatical disruption.

Bilingual Dictionary. To switch to the equivalent word in the target language, we need bilin-

[3]To match the BabyLM challenge guidelines, the word budgets for the code-switching variants are reduced by the number of tokens that are used for creating the bilingual dictionary and POS tagger.

gual dictionaries for eng–nld and eng–zho. Due to the token constraint of the BabyLM challenge (Choshen et al., 2026), we do not use pretrained models for creating the dictionaries. Instead, we extract word alignments from parallel subtitle corpora (Lison and Tiedemann, 2016) using MGIZA (Gao and Vogel, 2008), a multi-threaded implementation of GIZA++ (Och and Ney, 2003). Additionally, we train custom POS taggers on Universal Dependencies treebanks (Nivre et al., 2017)[4] using a BiLSTM architecture without pretrained embeddings. We use the bilingual dictionaries and POS taggers for controlling the *switch rate* and *switch point* of models trained with code-switched data.

4 BabyLM Challenge

We evaluate the models using the benchmark of the BabyLM Multilingual Challenge (Choshen et al., 2026), which includes tasks for English (eng), Dutch (nld), and Chinese (zho).

4.1 Benchmark Tasks

We examine both zero-shot linguistic knowledge and fine-tuned downstream performance.[5]

Zero-shot. Zero-shot evaluation measures capabilities acquired during pretraining without additional updates, focusing on grammatical competence, commonsense reasoning, discourse understanding, and semantic composition across languages. These tasks test whether models develop transferable linguistic knowledge, including sensitivity to syntactic structure, understanding of plausible relations, coherent discourse interpretation, and cross-lingual meaning representation.

Fine-tuning. Fine-tuning evaluation measures the transferability of learned representations by adapting models to downstream tasks. The tasks evaluate broader language understanding abilities, including reasoning, reading comprehension, factual knowledge retrieval, natural language inference, truthfulness, and multilingual classification.

4.2 Model Performance

In this section, we report the model performance change of the four code-switching variants compared to bilingual baselines on the BabyLM benchmark for eng–nld and eng–zho. The results shown in Table 2 are the macro-average of all the tasks, and we report the mean across three random seeds.[6]

		eng–nld		eng–zho	
Eval.	**Model**	Δeng	Δnld	Δeng	Δzho
ZS	Static	-0.08	**+0.19**	-0.52	-4.14
	Curr.	-0.32	-0.19	-0.20	-3.99
	POS-Static	**+0.38**	-0.30	-0.59	-3.66
	POS-Curr.	-0.54	-0.46	-0.68	-3.79
FT	Static	+2.11	**+5.23**	+6.09	**+0.58**
	Curr.	+1.44	+1.63	**+7.05**	-1.06
	POS-Static	-0.84	+3.31	+3.12	-1.03
	POS-Curr.	**+2.37**	+4.74	+5.73	-0.23

Table 2: Macro-averaged performance changes relative to the bilingual baseline (Δ in percentage points). The results are averaged across three random seeds. ZS: zero-shot; FT: fine-tuning. Best improvements within each language pair and evaluation setting are in **bold**.

English and Dutch. For the eng–nld models, zero-shot performance remains largely unchanged across all code-switching variants, indicating that training with code-switched data preserves general multilingual capabilities. Interestingly, the Static and POS-Static show slightly better overall grammatical knowledge compared to their Curriculum counterparts, indicating that too much code-switching at early training could be slightly harmful for the model's grammatical knowledge. For the fine-tuning tasks, we see that introducing code-switched contexts improves downstream performance for both English and Dutch, with the Static variant achieving the strongest overall results. Overall, these results indicate that training with code-switched data enhances downstream transfer without sacrificing general linguistic knowledge for this closely related language pair.

English and Chinese. For the eng–zho models, the result is more asymmetric. For the zero-shot tasks, we see that training with code-switched data is harmful for the model's linguistic knowledge, with the negative effect being particularly pronounced for Chinese. All code-switching variants outperform the bilingual baseline after fine-tuning, but the improvements are driven primarily by the English side, whereas Chinese downstream performance usually remains under the baseline. These results suggest that, for this typologically distant language pair, intra-sentence code-switching affects some aspects of Chinese knowledge acquired during pretraining.

[4] Specifically, we use English Web Treebank (EWT), Dutch LassySmall, and Chinese PUD treebanks (Silveira et al., 2014; Bouma and van Noord, 2017; McDonald et al., 2013).

[5] Detailed tasks lists are provided in Appendix F.

[6] Results for each task are presented in Appendix G.

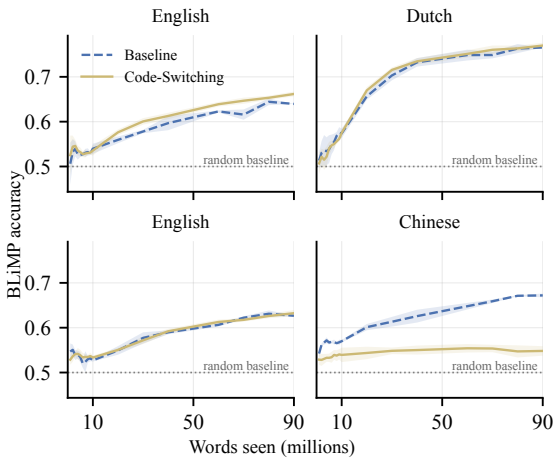

Figure 1: Early learning curve of grammatical knowledge for English–Dutch (top row) and English–Chinese (bottom row) bilingual models. Left column: English BLiMP; right column: target-language BLiMP (Dutch BLiMP for eng–nld, Chinese BLiMP for eng–zho). Each panel shows BLiMP accuracy as a function of words seen during training. The code-switching line represents the mean across four variants. Shaded bands denote standard deviation across random seeds.

4.3 Learning Curve

To inspect learning dynamics, we track learning trajectories on the BLiMP benchmark during early pretraining in Figure 1. For `eng-nld`, code-switch models show slightly faster initial progress compared to the baseline. In contrast, `eng-zho` presents a mixed picture. While English performance remains steady and largely maintains its level relative to the baseline, Chinese accuracy is negatively affected from an early stage. Rather than improving over time, Chinese performance lags behind the baseline early on and fails to recover as pretraining progresses. Overall, these trajectories show that code-switching provides an effective early learning signal for `eng-nld`, but it has an immediate negative impact on the Chinese knowledge in `eng-zho`.

5 Representational Geometry

To better understand how training with code-switched data reshapes the internal structure, we examine our models using a range of metrics comparing their global structure, relative alignment across languages, and the effect on cross-lingual transfer.

5.1 Global Structure

As a first sanity check, we measure how introducing code-switched contexts changes the representational structure of the embedding layer compared to a bilingual baseline.

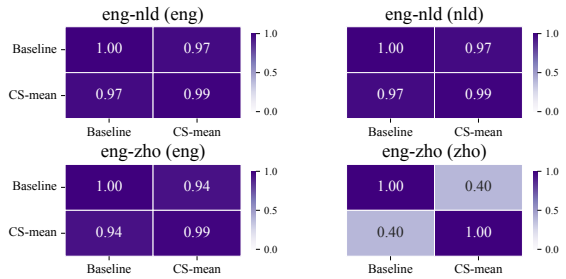

Figure 2: Similarity of the global structure of sentence representations within a language in bilingual models measured by centered kernel alignment (CKA). We compare the *Baseline* model against the mean of code-switching variants (*CS-mean*) for the English–Dutch (`eng–nld`) models in the top row and the English–Chinese (`eng–zho`) models in the bottom row. Darker cell colors indicate higher representational similarity.

Centered Kernel Alignment. We measure the similarity of sentence representations within each language using linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019). Let $X \in \mathbb{R}^{n \times d_x}$ denote the representation of n sentences in the baseline model and $Y \in \mathbb{R}^{n \times d_y}$ denote the representations of the same sentences in a code-switched model. Linear CKA is defined as

$$\mathrm{CKA}(X, Y) = \frac{\|Y^\top X\|_{\mathrm{F}}^2}{\|X^\top X\|_{\mathrm{F}} \, \|Y^\top Y\|_{\mathrm{F}}},$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. CKA quantifies the global similarity of the two representational spaces and is invariant to isotropic scaling and orthogonal transformations. Sentences are represented by the average of their token embeddings.

Results. All code-switching variants lead to highly similar representational structure. We therefore focus on the comparison to the baseline in Figure 2 and average across code-switching conditions and seeds. For the pair `eng-nld`, the global representational structure remains highly similar for both languages. In contrast, the analysis of the `eng-zho` models reveals a highly asymmetric pattern. While the structure of the English representations remains relatively close to the baseline, training with code-switched data fundamentally changes the structure of the Chinese sentence representations. Despite this strong shift relative to the baseline, the structure of code-switched variants is highly similar, indicating that the overall effect is driven primarily by the introduction of

code-switched context rather than by the specific training variant.

5.2 Cross-Lingual Alignment

After assessing what code-switching changes relative to the baseline, we now examine its effect on the alignment of the two languages within a model.

Cosine Similarity of Translations. We measure the alignment between the representation an English input $\mathbf{x}_i$ and the representation of its translation in the target language $\mathbf{y}_i$ by calculating cosine similarity:

$$\cos(\mathbf{x}_i, \mathbf{y}_i) = \frac{\mathbf{x}_i^\top \mathbf{y}_i}{\|\mathbf{x}_i\|_2 \|\mathbf{y}_i\|_2}.$$

In addition to the parallel sentence pairs, we perform the same analysis on a curated set of word pairs. We use the translation pairs described in Section 3.2 and additionally examine a list of known cognates with similar form and meaning in English and Dutch (Lefever et al., 2020).

Results. We calculate the average alignment across all input elements and report the delta over the corresponding baseline model in Figure 3. A higher score indicates that semantically equivalent inputs from the two languages are embedded closer together in the shared representation space. We observe a strong increase in alignment for the `eng-nld` models both on the sentence level (top left) and the lexical level (top right). Code-switching with a static *switch rate* produces the largest increase, but the differences between code-switching variants remain small. We visualize the lexical alignment for 5,285 `eng-nld` cognate pairs when training with the POS-Static variant in Figure 4 and compare it to the baseline.[7] It shows that both approaches closely align cognates with high orthographical overlap (*material-materiaal*, *chocolate-chocolate*), but training with code-switched data increases the alignment for more diverging pairs (*answer-antwoord*, *line-lijn*). This indicates that the code-switched context alone can facilitate better cross-lingual semantic alignment beyond surface patterns without introducing an explicit alignment objective during training.

The effect of code-switched training on cross-lingual sentence alignment is generally weaker for the language pair English–Chinese. Only POS-sensitive models slightly increase the sentence-level alignment relative to the baseline. On the lexical level, the difference in cosine similarity between the baseline and the code-switching variants is even more pronounced than for English–Dutch pairs. However, similarity increases for both translation pairs and random pairs, indicating that the changes in representational structure do not necessarily lead to better bilingual understanding.

[7] Figure 6 in Appendix E presents plots for all variants.

5.3 Relative Structure

We abstract from specific changes in lexical cross-lingual alignment by conducting a second-order representational similarity analysis.

Representational Similarity Analysis. We calculate the Spearman correlation over the cosine similarities between English representations $X^{(A)}$ and target language representations $X^{(B)}$ for translation pairs. While CKA compares feature spaces via linear transformations, RSA evaluates whether two models preserve similar relative distances across lexical translation pairs.

Results. We observe a similar pattern as for the CKA analysis on sentence embeddings. Figure 5 shows that token-level relations remain similar when comparing the code-switching variant to the baseline for all `eng-nld` models. This indicates that training on code-switched data causes only limited reorganization of the embedding geometry. In contrast, we observe a substantially larger divergence for the `eng-zho` models, suggesting that code-switching has a much stronger effect for this language pair that does not share lexical overlap. However, different code-switching variants remain highly similar to one another, indicating that they converge toward a shared lexical geometry despite differences in training strategies.

Overall, training with code-switched data produces a more pronounced representational shift for `eng-zho` than for `eng-nld`, on both global structure and relational alignment.

5.4 Cross-Lingual Transfer

We now examine how the observed changes in cross-lingual alignment induced by training on code-switched data affect the performance on tasks requiring bilingual knowledge.

Translation Retrieval. We test whether the increased alignment on translation pairs suffices to effectively distinguish them from unrelated sentences. We use a similar procedure as Lample et al.

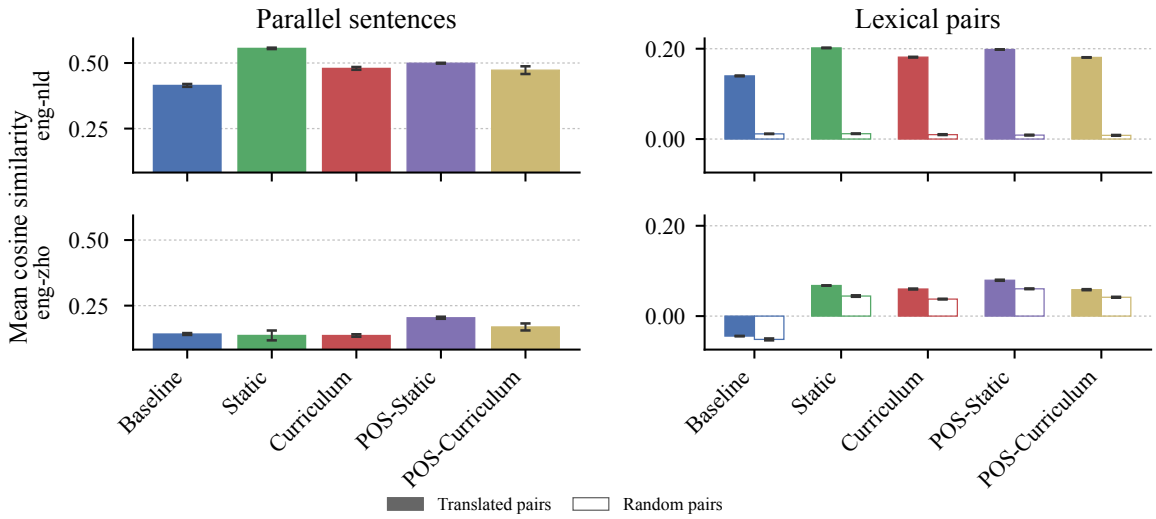

Figure 3: Cross-lingual embedding alignment across models for English–Dutch (eng–nld; top row) and English–Chinese (eng–zho; bottom row). The left column displays the mean cosine similarity for aligned parallel sentences, while the right column shows lexical alignment for curated translation pairs (filled bars) versus randomly matched word pairs (outlined bars).

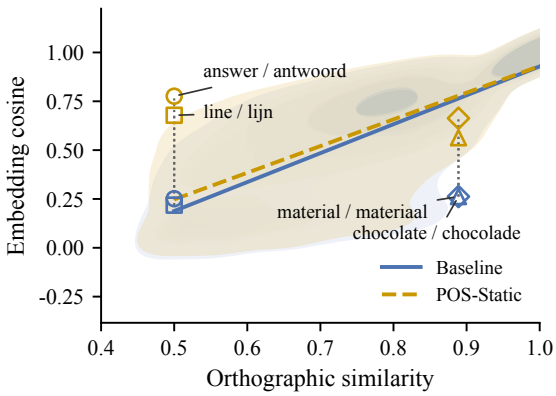

Figure 4: Relationship between orthographic similarity and word embedding cosine similarity for English–Dutch cognates ($n = 5,285$ pairs). Contours indicate kernel density estimates. Dotted connectors link the same cognate pair between the *Baseline* and *POS-Static* setups, illustrating shifts in embedding similarity at equivalent levels of orthographic overlap (e.g., *answer/antwoord*, *line/lijn*), where ideal cross-lingual representations maintain high cosine similarity regardless of distance. Pearson correlations are $r = 0.75$ for Baseline and $r = 0.72$ for POS-Static.

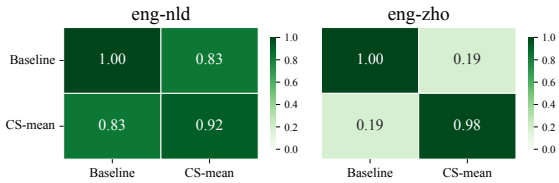

Figure 5: Lexical-level Representational Similarity Analysis (RSA) heatmaps for English—Dutch (eng–nld) and English–Chinese (eng–zho). Panels compare cross-model representational similarity between the *Baseline* model and the average of code-switching variants (*CS-mean*) over paired-token embeddings. Darker cell colors indicate higher representational similarity.

(2018) and rank 1,000 candidate sentence representations according to their cosine similarity to the representation of the source sentence.[8] The translation retrieval performance is reported using Precision@1 (P@1), which measures the percentage of the correct translation that appears among the top-1 retrieved sentences. Higher P@1 indicates stronger cross-lingual alignment.

Results. We report retrieval performance as P@1 over all candidate sentences in Table 3 and see a clear difference between the two language pairs. For eng–nld, training with code-switched data improves translation retrieval over the bilingual baseline. This indicates that code-switching increases not only the absolute similarity of translation pairs as shown in Figure 3, but also their relative distance to other candidate sentences. In contrast, the eng–zho results show a weaker retrieval pattern. All code-switching variants further decrease the already low retrieval precision. While retrieval success remains above the chance level, it indicates that code-switching does not reliably position Chinese sentences closer to their English translations relative to other candidate sentences.

[8] Sentences are represented by the average of their token embeddings.

Language pair	Model	eng → X	X → eng
eng–nld	Baseline	13.9	14.9
	Static	30.5	32.0
	Curr.	20.2	22.1
	POS-Static	21.5	22.4
	POS-Curr.	17.2	19.4
eng–zho	Baseline	5.6	6.9
	Static	2.7	3.0
	Curr.	2.7	3.1
	POS-Static	2.8	3.1
	POS-Curr.	2.6	2.9

Table 3: Sentence-level translation retrieval measured by P@1 (%), showing the percentage of the correct translation that appears among the top-1 retrieved sentences. The results are averaged across three random seeds. The P@1 for random retrieval is 0.1%.

6 Discussion

Our complementary analysis of performance aspects and representational structure provides a coherent picture of how training with code-switched data impacts bilingual language models. Our contrasting findings for `eng-nld` and `eng-zho` models suggest that the benefit of intra-sentence code-switching depends on how easily the two languages can be integrated within the same training context. For the typologically close `eng-nld` pair, the bilingual baseline already leads to a reasonably aligned representation space, and code-switching further improves both the accuracy on linguistic minimal pairs and downstream performance. The strongest results are obtained with the unconstrained Static variant. This may indicate that further constraining the *switch point* using linguistic constraints is not necessary if the surface structure of the two languages is highly similar (as is the case for Dutch and English). These findings are compatible with Poplack's (1980) Equivalence Constraint, under which switching is more likely at points where the surface structures of the two languages can be combined without violating either grammar.

For the `eng-zho` models, training with code-switched data fails to improve meaningful bilingual alignment and also leads to weaker grammatical knowledge in Chinese. At the same time, some code-switched models produce higher absolute similarity between `eng-zho` translation pairs than the bilingual baseline. This suggests that the problem is not simply that the two languages become less similar in the representation space. Rather, the increase in similarity does not translate into better relative alignment, as non-translation pairs may become similarly close. The CKA and RSA results provide further evidence that code-switching substantially reorganizes the Chinese representation space while leaving the English space comparatively stable. These findings suggest that intra-sentence code-switching may alter the organization of Chinese representations without producing the translation-relevant structure needed for effective cross-lingual transfer.

The difficulty of the `eng-zho` setting may partly arise from the substantially different linguistic and orthographic properties of the two languages, which provide fewer straightforward *switch points* for intra-sentence integration. However, this interpretation should be considered alongside the training constraints of the BabyLM challenge (Choshen et al., 2026). The limited token budget requires the POS tags and bilingual dictionary to be self-trained from relatively little data, making the resources used to identify and construct switching points relatively noisy. Such noise may have a greater impact when the languages are more distant, where reliable lexical and structural correspondences are already harder to establish. Thus, the observed difference between `eng-nld` and `eng-zho` may reflect an interaction between properties of the language pair and the quality of the resources available for code-switching, rather than typological distance alone.

7 Conclusion

We tested four cognitively-inspired code-switching variants and examined their effect on bilingual language models for two language pairs. Our results show that training with code-switched data can improve cross-lingual transfer and representational alignment, but its effectiveness seems to depend on the compatibility of the languages.

For the closely related `eng-nld` pair, training with code-switched data generally improves linguistic proficiency, downstream performance, and internal cross-lingual alignment compared to the bilingual baseline. The effects of manipulating the *switch rate* and the*switch point* are relatively small, with the Static variant showing slightly better overall performance. This suggests that the introduction of code-switching itself is more influential than the precise choice of when or how frequently language switching occurs for `eng-nld`. The results for `eng-zho`, however, show that these benefits do not necessarily generalize to all languages. Code-switching variants perform worse than the bilin-

gual baseline on sentence-level translation retrieval and on grammatical tasks, especially for Chinese. Within the eng−zho code-switching variants, POS-Static tends to perform marginally better than the ones with random *switch point* and the ones with high *switch rate*, suggesting that constraining the switching location and reducing switched content at early training stages may better preserve grammatical structure when code-switching between more distant languages.

Overall, our findings suggest that training with code-switched data is a promising mechanism for improving cross-lingual transfer, particularly for languages that have high lexical overlap. Future work could therefore investigate whether typological distance remains a significant factor when larger token budgets, higher-quality POS annotations, and more reliable bilingual dictionaries are available.

Limitations

Although training with code-switched data appears to support cross-lingual transfer, particularly for eng−nld even under the limited-token constraints, several limitations remain regarding the cognitive plausibility of these models.

Importantly, our approach does not strictly mirror the cognitive and interactional realities of bilingual child language acquisition. In naturalistic settings, parent-child code-switching is highly contextualized and communicatively driven; for instance, bilingual parents frequently utilize one language to instruct an infant about a novel object or concept associated with another language (García, 1980; Bail et al., 2015). Our word- and phrase-level substitution framework lacks this semantic and situational grounding, as it relies on a blind stochastic process rather than communicative intent. In addition, the code-switched data we use in this study is synthetic. While we try to introduce high-quality code-switched context, the data are inherently noisy and occasionally invalid, as they are generated via a probabilistic pipeline applied to translated parallel corpora. Another limitation in cognitive plausibility stems from the modeling design: our current pipeline exposes the model to a fixed architectural design from the onset of training, while in human language acquisition, children dynamically expand their vocabulary size, lexical processing speed, and syntactic complexity over time. Incorporating a cognitively plausible architecture remains an important direction for future work.

Acknowledgments

We thank the reviewers for the thoughtful and detailed feedback. We thank Irantzu Elcoro Alberdi for the insightful discussion of the literature on code-switching. This research is partially supported by the zukunft.niedersachsen program of the VolkswagenStiftung (L.B., Z.H., L.P.) and by a VENI grant (Vl.Veni.211C.039) from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) (L.B.).

Author Contributions

Z.H.: Conceptualization, Analysis, Literature review, Visualization, Writing—original draft, Writing—review & editing. **L.P.:** Conceptualization, Literature review, Writing—original draft (parts of Introduction and Related Work). **L.B.:** Supervision, Conceptualization, Analysis, Writing—review & editing.

References

Evangelia Adamou and Xingjia Rachel Shen. 2019. There are no language switching costs when codeswitching is frequent. *International Journal of Bilingualism*, 23(1):53–70.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.

Shunta Asano, Jeonghun Baek, and Toshihiko Yamasaki. 2026. Beyond bilingual transfer: Multilingual code-switching in instruction tuning. *Preprint*, arXiv:2605.29414.

Shoji Azuma and Richard P. Meier. 1997. Open class and closed class: Sentence-imitation experiments on intrasentential code-switching. *Applied Psycholinguistics*, 18(3):257–276.

Amelie Bail, Giovanna Morini, and Rochelle S. Newman. 2015. Look at the gato! Code-switching in speech to toddlers. *Journal of Child Language*, 42(5):1073–1101.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Bilson, Hanako Yoshida, Crystal D. Tran, Elizabeth A. Woods, and Thomas T. Hills. 2015. Semantic facilitation in bilingual first language acquisition. *Cognition*, 140:122–134.

Esti Blanco-Elorrieta and Liina Pylkkänen. 2017. Bilingual Language Switching in the Laboratory versus in the Wild: The Spatiotemporal Dynamics of Adaptive Language Control. *The Journal of Neuroscience*, 37(37):9022–9036.

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Krista Byers-Heinlein. 2013. Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition*, 16(1):32–48.

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.

Michelle M. Chouinard and Eve V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(3):637–669.

Liane Comeau, Fred Genesee, and Lindsay Lapaquette. 2003. The Modeling Hypothesis and child bilingual codemixing. *International Journal of Bilingualism*, 7(2):113–126.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Margaret Deuchar and Suzanne Quay. 1994. Language choice and code-switching in a young bilingual child. In *Proceedings of the 10th Annual Sociolinguistics Symposium*, Lancaster, England, United Kingdom. 18-page conference paper.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Eugene E. García. 1980. The function of language switching during bilingual mother-child interactions. *Journal of Multilingual and Multicultural Development*, 1(3):243–252.

Dorota Gaskins, Antje Endesfelder Quick, Anna Verschik, and Ad Backus. 2022. Usage-based approaches to child code-switching: State of the art and ways forward. *Cognitive Development*, 64:101269.

Fred Genesee. 2008. Bilingual First Language Acquisition: Evidence from Montreal. *Diversité urbaine*, pages 9–26.

Helen Hambly, Yvonne Wren, Sharynne McLeod, and Sue Roulstone. 2013. The influence of bilingualism on speech production: A systematic review. *International Journal of Language & Communication Disorders*, 48(1):1–24.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbinliu, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. 2026. MuBench: Assessment of multilingual capabilities of large language models across 61 languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 16163–16192, San Diego, California, United States. Association for Computational Linguistics.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Adrian Florea, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, Hinrich Schütze, and Nima Mesgarani. 2025. XCOMPS: A multilingual benchmark of conceptual minimal pairs. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 75–81, Vienna, Austria. Association for Computational Linguistics.

Qiaoya Huang. 2025. Bilingual children's code-switching: Types, approaches, and functions. *International Journal of Social Science and Education Research*, 7(1):38–46.

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. 2026a. BabyBabelLM: A multilingual benchmark of developmentally plausible training data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3297–3329, Rabat, Morocco. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026b. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Santosh Srinath K, Mudit Somani, Varun Reddy Padala, Prajna Devi Upadhyay, and Abhijit Das. 2026. Continual-learning for modelling low-resource languages from large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6258–6275, Rabat, Morocco. Association for Computational Linguistics.

Maria Kapantzoglou, Julie Esparza Brown, Lauren M. Cycyk, and Gerasimos Fergadiotis. 2021. Code-switching and language proficiency in bilingual children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, 64(5):1605–1620.

Margarita Kaushanskaya, Kimberly Crespo, and Anne Neveu. 2023. Does code-switching influence novel word learning? *Developmental Science*, 26(2):e13292.

Muhammad Dehan Al Kautsar and Fajri Koto. 2025. Parallel tokenizers: Rethinking vocabulary design for cross-lingual transfer. *Preprint*, arXiv:2510.06128.

Yunsu Kim, Hendrik Rosendahl, Nick Rossenbach, Jan Rosendahl, Shahram Khadivi, and Hermann Ney. 2019. Learning bilingual sentence embeddings via autoencoding and computing similarities with a multilayer perceptron. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 61–71, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Lena V. Kremin, Adriel John Orena, Linda Polka, and Krista Byers-Heinlein. 2019. Investigating naturalistic code-switching directed towards infants. *The Journal of the Acoustical Society of America*, 145(3_Supplement):1766–1766.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Els Lefever, Sofie Labat, and Pranaydeep Singh. 2020. Identifying cognates in English-Dutch and French-Dutch by means of orthographic information and cross-lingual word embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4096–4101, Marseille, France. European Language Resources Association.

Peiqin Lin, Andre Martins, and Hinrich Schuetze. 2025. A recipe of parallel corpora exploitation for multilingual large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4038–4050, Albuquerque, New Mexico. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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. 2025. A systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Jürgen M. Meisel. 2019. *Keeping Languages Apart: Mixing, Interference and Interaction of Languages*, page 93–122. Cambridge University Press.

Carol Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University PressOxford.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 71–76, Bergen, Norway. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Johanne Paradis, Elena Nicoladis, and Fred Genesee. 2000. Early emergence of structural constraints on code-mixing: evidence from french–english bilingual children. *Bilingualism: Language and Cognition*, 3(3):245–261.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching1. *Linguistics*, 18(7-8).

B. A. Mahalakshmi Prasad and Prema K. S. Rao. 2011. Code Switching in the Lexical Corpora of Children. *Theory and Practice in Language Studies*, 1(5):471–477.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Imanol Schlag, Marzieh Fadaee, Sara Hooker, Antoine Bosselut, and 1 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *ICLR*.

Julie M. Schneider, Giovanna Morini, Veronica De Jesus, and Aquiles Iglesias. 2026. Development of intrasentential code-switching patterns in spanish-speaking english learners. *Child Development*, 97(4):1013–1027.

Yingli Shen, Wen Lai, Shuo Wang, Ge Gao, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2025. From unaligned to aligned: Scaling multilingual LLMs with multi-way parallel corpora. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7357–7379, Suzhou, China. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Erin Smolak, Stephanie De Anda, Bianka Enriquez, Diane Poulin-Dubois, and Margaret Friend. 2020. Code-switching in young bilingual toddlers: A longitudinal, cross-language investigation. *Bilingualism: Language and Cognition*, 23(3):500–518.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Desirée Ramírez Urbaneja. 2020. "¿Tú tienes una little pumpkin?" Mixed noun phrases in Spanish-English bilingual children and adults. *International Journal of Bilingualism*, 24(4):824–839. _eprint: https://doi.org/10.1177/1367006919888580.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11032–11046, Vienna, Austria. Association for Computational Linguistics.

Janet F. Werker and Krista Byers-Heinlein. 2008. Bilingualism in infancy: first steps in perception and comprehension. *Trends in Cognitive Sciences*, 12(4):144–151.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2025. Code-Switching Curriculum Learning for Multilingual Transfer in LLMs. *arXiv preprint*. ArXiv:2411.02460 [cs.CL].

W. Quin Yow, Jessica S. H. Tan, and Suzanne Flynn. 2018. Code-switching as a marker of linguistic competence in bilingual children. *Bilingualism: Language and Cognition*, 21(5):1075–1090.

Linda Zeng, Steven Y. Feng, and Michael C. Frank. 2026. Bringing Up a Bilingual BabyLM: Investigating Multilingual Language Acquisition Using Small-Scale Models. *arXiv preprint*. ArXiv:2603.29552 [cs.CL].

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pre-training. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.

A Building the Bilingual Dictionaries

The bilingual dictionaries utilized in our code-switching pipeline are built from parallel subtitle corpora from OpenSubtitles (Lison and Tiedemann, 2016), obtained via the OPUS platform (Tiedemann, 2012). Prior to statistical alignment, all sentence pairs were filtered to enforce a maximum length constraint of 80 tokens per side and a length ratio of $\leq$ 9:1. The specific dataset parameters, release configurations, and resulting sentence pair counts utilized during the alignment process are detailed in Table 4.

Languages	Release	Aligned Sentence Pairs
`eng-nld`	v2018	325,874
`eng-zho`	v2024	504,020

Table 4: Parallel Corpus Configurations for Bilingual Lexicon Extraction

To learn word-level translation mappings, we perform unsupervised word alignment using MGIZA (Gao and Vogel, 2008), a multi-threaded extension of GIZA++ (Och and Ney, 2003). The preprocessing pipeline first generates internal source vocabularies and sentence co-occurrence structures via `snt2cooc`, while `mkcls` (Och, 1999) clusters words into statistical frequency classes. Word alignments are trained using the Expectation–Maximization (EM) algorithm (Brown et al., 1993). To bootstrap baseline lexical translation probabilities, the optimization routine begins with 5 iterations of IBM Model 1. The alignment pipeline then transitions to 3 iterations each of IBM Models 3 and 4 to account for word-order reordering (distortion) and one-to-many word mappings (fertility). Models are trained symmetrically in both directions (source $\rightarrow$ target and target $\rightarrow$ source). Finally, the resulting directional alignments are merged using the standard `grow-diag-final-and` symmetrization heuristic (Koehn et al., 2003) to extract high-confidence word and multi-word alignment links.

English–Dutch Lexicon Extraction Prior to statistical alignment, the English and Dutch text is standardized by lowercasing and applying whitespace tokenization. Once the unified bidirectional alignments are produced by the Pharaoh-formatted alignment map, a unigram extraction layer processes the aligned sentences to count every valid one-to-one mapping across the corpus. To eliminate structural alignment noise, spurious translations, and low-frequency accidental links, we enforce a strict minimum co-occurrence threshold where a word translation pair must appear ≥ 5 times in the corpus to be retained. Punctuation marks are automatically stripped, and the final filtered lexicon maps individual source unigrams directly to their most probable target counterparts.

English–Chinese Lexicon Extraction Prior to running the alignment models, the text is processed without relying on any pretrained segmenters or tokenizers: the English data is lowercased and whitespace-tokenized, whereas the Chinese sentences are split strictly at the character level, treating each individual character as an autonomous token. From the unified bidirectional alignments, the lexicon is constructed via a dual-layer extraction architecture to properly handle structural differences:

1. **Unigram Layer:** Captures and counts every direct one-to-one alignment link across the corpus (e.g., mapping individual English words to single Chinese characters).

2. **Phrase Layer:** Extracted via standard Moses-style phrase extraction to recover compound multi-character concepts. For each alignment anchor, the algorithm identifies all contiguous English and Chinese spans (up to a maximum boundary constraint of 5 English tokens and 6 Chinese characters), where every internal alignment link is strictly consistent, and no links cross outside the candidate span boundaries.

Once multi-word English expressions and multi-character Chinese compounds are recovered, contiguous Chinese characters within an extracted phrase are joined into a unified, non-whitespace string (e.g., fusing character-level outputs to form a multi-character word). The unigram and phrase layers are subsequently merged into a master dictionary. Mappings are filtered out if they have fewer than a minimum co-occurrence frequency threshold ($\geq$ 2 occurrences), punctuation marks are stripped, and English-only function-word phrases are discarded. Finally, a subsumption pruning pass removes shorter dictionary entries if they are strictly dominated by a longer phrase that possesses an equal or higher co-occurrence frequency.

Runtime Dictionary Utilization Each generated entry in both lexicons tracks total occurrence

counts alongside the empirical translation probability $P(\text{target} \mid \text{source})$, representing how frequently a specific translation appears among all candidate options for a given source term. At runtime, the data pipeline queries these ranked candidates directly, choosing the highest-probability option.

B POS tagger Trainer

In this work, we train part-of-speech (POS) taggers for English, Chinese, and Dutch using the Universal Dependencies treebanks (Nivre et al., 2017). Each dataset is provided in CoNLL-U format and is split into standard training, development, and test partitions.

We employ a BiLSTM-CRF architecture implemented in PyTorch. The model is trained from scratch without the use of pretrained word embeddings. Instead, it learns both word-level and character-level representations directly from the training data. Each token is represented as the concatenation of a learned word embedding and a character-level representation obtained via a convolutional neural network over characters. This allows the model to better handle rare and out-of-vocabulary words. The sequence of token representations is then processed by a bidirectional LSTM, which captures contextual information from both the left and right contexts of each word. The resulting contextualized representations are passed through a linear layer to produce emission scores over POS tags. A conditional random field (CRF) layer is used on top to model dependencies between adjacent tags and to decode the globally optimal tag sequence.

For data preparation, sentences are extracted from the CoNLL-U files as sequences of tokens paired with their gold POS tags. We construct vocabularies for words, characters, and POS tags based on the training set. During inference, unseen words are mapped to a special `UNK` token.

The model is trained using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 32, and gradient clipping to improve training stability. The training objective is the negative log-likelihood of the correct tag sequence under the CRF layer. After each epoch, performance is evaluated on the development set using token-level accuracy. The best-performing model on the development set is saved separately for each language. Early stopping is applied if development accuracy does not improve for three consecutive epochs.

At evaluation time, the saved model checkpoints are used to compute performance on both the development and test sets. We report overall token-level accuracy as well as per-tag accuracy for major POS categories. The models achieve approximately 82% accuracy on English, 85% on Chinese, and 83% on Dutch test sets. Performance is generally strongest on frequent and syntactically regular tags, while lower performance is observed on rare categories such as interjections and foreign words.

Once trained, the models assign a POS tag to each token in an input sentence. For example, in English, *orange* may be labeled as an ADJ and *cat* as a NOUN. However, in Chinese, as a word can be formed by several characters, the POS tagger can assign one tag to a compound word. For example, 橘猫 (orange cat) can be labeled as NOUN.

C Data Licensing

The training data used for our models follow the licensing conditions described by Jumelet et al. (2026a), who report that, during their data collection effort, they verified that all data were released with licenses permitting academic research, such as Creative Commons licenses or Public Domain.

For the construction of the bilingual dictionaries, we used the OpenSubtitles2016 corpus (Lison and Tiedemann, 2016), which is distributed through the OPUS collection and was made freely available to the research community. For training the POS tagger, we used the Universal Dependencies (UD) English Web Treebank (EWT) (Silveira et al., 2014), Dutch LassySmall (Bouma and van Noord, 2017), and Chinese PUD treebanks (McDonald et al., 2013). The English-EWT and Dutch-LassySmall treebanks are licensed under CC BY-SA 4.0, while the Chinese-PUD treebank is licensed under CC BY-SA 3.0. We used these resources in accordance with their respective licensing terms.

Finally, the Dutch–English cognate dataset used for cross-lingual alignment is released under the Creative Commons Attribution–NonCommercial (CC BY-NC) license, which permits sharing and adaptation with appropriate attribution but prohibits commercial use.

D Hyperparameter

The hyperparameters of model training are described in Table 5.

Parameter	Value
Model architecture	12-layer GPT-2 with hidden size 768 and 12 attention heads.
Tokenizer	16,384-token shared bilingual BPE vocabulary.
Optimizer	AdamW.
Learning Rate	5×10^{-5}. Cosine decay with 1% linear warmup.
Batch size	16.
Sequence length	512 tokens.
Randomization	Independent random seeds (13, 42, 123) for model initialization and code-switch decisions.

Table 5: Training hyperparameters and configuration.

E Cognates Alignment

Figure 6 shows the Pearson correlation between orthographic similarity and word embedding cosine similarity for English–Dutch cognates. The figure shows the results for the bilingual baseline and all four code-switching variants.

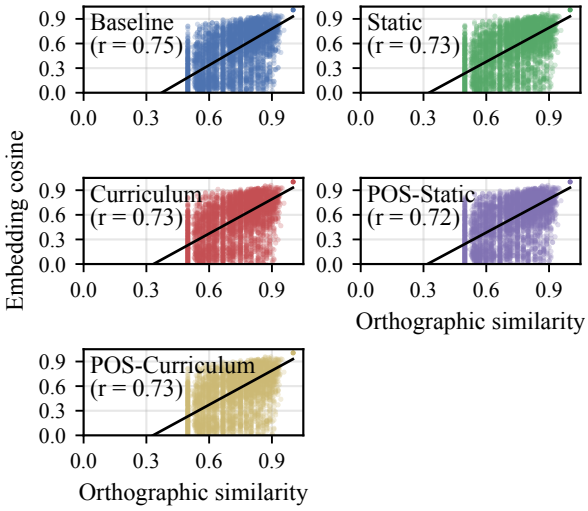

Figure 6: Relationship between orthographic similarity and word embedding cosine similarity for English–Dutch cognates ($n = 5,285$ pairs). Lower Pearson correlation indicates better cross-lingual alignment, especially for cognates with less similar surface forms.

F BabyLM Benchmark

Table 6 and Table 7 list all the tasks being evaluated for `eng-nld` and `eng-zho` models.

Evaluation Dimension	eng–nld Tasks	eng–zho Tasks
Grammatical competence	blimp_babylm_filtered[1], multiblimp_eng[2], blimp_nl[3], multiblimp_nld[2]	blimp_babylm_filtered[1], multiblimp_eng[2], zhoblimp[4]
Commonsense reasoning	hellaswag_en_mubench[5], winogrande_en_mubench[5], hellaswag_nl_mubench[5], winogrande_nl_mubench[5]	hellaswag_en_mubench[5], winogrande_en_mubench[5], hellaswag_zh_mubench[5], winogrande_zh_mubench[5]
Discourse understanding	xstorycloze_en_mubench[5], xstorycloze_nl_mubench[5]	xstorycloze_en_mubench[5], xstorycloze_zh_mubench[5]
Compositional semantics	xcomps_nl[6]	xcomps_zh[6]

Table 6: Zero-shot evaluation benchmarks grouped by evaluation dimension. [1](Choshen et al., 2026); [2](Jumelet et al., 2026b); [3](Suijkerbuijk et al., 2025); [4](Liu et al., 2025); [5](Han et al., 2026); [6](He et al., 2025);

Evaluation Dimension	eng–nld Tasks	eng–zho Tasks
Science and general reasoning	arc_mubench[1]	arc_mubench[1]
Reading comprehension	belebele[2]	belebele[2]
Factual knowledge	bmlama_mubench[1]	bmlama_mubench[1]
Natural language inference	mnli_mubench[1], xnli (eng)[3], mnli_mubench (nld)[1]	mnli_mubench[1], xnli[3]
Topic classification	sib200[4]	sib200[4]
Truthfulness and calibration	truthfulqa_mubench[1]	truthfulqa_mubench[1]
Language-specific understanding	include (nld)[5]	include (zho)[5]

Table 7: Fine-tuning evaluation benchmarks grouped by evaluation dimension. [1](Han et al., 2026); [2](Bandarkar et al., 2024); [3](Conneau et al., 2018); [4](Adelani et al., 2024); [5](Romanou et al., 2024);

G Per-Category Results

Table 8 and Table 9 present the results of all models across each task.

Task	Setting	Baseline	Static	Curriculum	POS-Static	POS-Curriculum
Zero-Shot						
BLiMP	eng	72.6	**73.1**	72.6	72.7	72.4
MultiBLiMP	eng	87.7	**88.8**	87.9	88.1	88.3
Hellaswag	eng	26.5	26.5	26.5	26.4	**26.6**
Winogrande	eng	50.8	50.3	50.3	**51.1**	50.1
XStoryCloze	eng	49.1	49.0	48.6	**49.2**	48.6
PIQA parallel	eng	22.8	20.4	20.1	**23.3**	19.7
PIQA nonparallel	eng	49.5	51.0	**51.3**	50.7	50.0
BLiMP	nld	80.8	80.9	80.6	**81.1**	80.1
MultiBLiMP	nld	93.8	**94.0**	93.5	93.5	93.1
Hellaswag	nld	26.5	26.5	**26.6**	26.3	**26.6**
Winogrande	nld	49.8	**50.2**	49.6	49.8	**50.2**
XCOMPS	nld	53.8	54.0	**54.3**	53.9	**54.3**
XStoryCloze	nld	48.4	**48.6**	47.7	48.2	48.3
PIQA parallel	nld	22.7	22.3	22.3	18.8	**23.6**
PIQA nonparallel	nld	55.0	55.7	54.7	**56.7**	50.7
Fine-Tuned						
ARC	eng	**25.6**	24.7	25.1	25.1	24.4
ARC	nld	**24.9**	23.8	**24.9**	**24.9**	24.1
Belebele	eng	25.0	23.9	23.9	23.9	**25.6**
Belebele	nld	23.9	**28.2**	25.0	25.0	28.1
BM-LAMA	eng	9.7	10.2	11.3	9.7	**11.8**
BM-LAMA	nld	11.5	**12.0**	10.3	10.2	10.0
MNLI	eng	**50.3**	**50.3**	48.5	49.2	49.7
MNLI	nld	49.6	49.8	48.9	51.0	**51.3**
SIB-200	eng	59.8	**78.3**	76.2	58.8	77.7
SIB-200	nld	39.8	72.2	55.8	56.7	**72.7**
TruthfulQA	eng	**27.7**	**27.7**	24.7	25.6	25.4
TruthfulQA	nld	26.2	22.3	23.2	**27.1**	25.0
XNLI	eng	47.3	45.2	45.9	47.2	**47.4**
INCLUDE	nld	29.5	**33.6**	28.6	**33.6**	27.2

Table 8: Evaluation results for `eng-nld`. The results are averaged across three random seeds.

Task	Setting	Baseline	Static	Curriculum	POS-Static	POS-Curriculum
Zero-Shot						
BLiMP filtered	eng	71.3	**71.8**	71.7	70.0	71.5
MultiBLiMP	eng	86.0	**87.2**	86.9	85.2	86.5
Hellaswag	eng	**26.5**	**26.5**	26.4	26.2	26.4
Winogrande	eng	50.9	50.6	50.8	48.9	**51.1**
XStoryCloze	eng	**50.2**	48.8	49.0	47.5	48.6
PIQA parallel	eng	25.6	25.9	27.2	**28.8**	26.9
PIQA nonparallel	eng	**52.3**	49.3	50.0	51.3	48.0
Hellaswag	zho	**26.9**	25.9	25.9	25.8	26.0
Winogrande	zho	49.3	51.0	50.6	50.8	**51.1**
XCOMPS	zho	**53.4**	50.3	50.3	50.8	50.7
XStoryCloze	zho	**48.8**	46.2	46.3	46.6	46.6
ZhoBLiMP	zho	**76.0**	54.9	55.1	54.5	56.0
Hanzi structure	zho	**56.2**	50.1	50.6	50.6	50.9
Hanzi pinyin	zho	49.2	50.9	**51.4**	50.0	51.0
PIQA parallel	zho	**21.8**	19.7	18.8	20.7	17.5
PIQA nonparallel	zho	**50.5**	46.0	47.3	49.3	48.3
Fine-Tuned						
ARC	eng	24.8	23.8	**26.2**	24.4	24.5
ARC	zho	24.6	24.2	**25.5**	25.1	25.1
Belebele	eng	25.0	25.9	25.0	**26.1**	**26.1**
Belebele	zho	25.0	22.5	24.2	23.9	**26.9**
BM-LAMA	eng	10.1	10.3	10.9	10.2	**11.9**
BM-LAMA	zho	9.5	10.2	10.1	**10.9**	10.5
MNLI	eng	46.4	49.8	**56.0**	49.4	50.3
MNLI	zho	**45.8**	38.8	42.2	39.1	41.9
SIB-200	eng	36.8	**76.8**	76.2	56.8	74.7
SIB-200	zho	40.0	**51.3**	29.3	37.0	36.8
TruthfulQA	eng	21.7	**26.5**	24.4	23.2	25.0
TruthfulQA	zho	22.3	25.3	**27.7**	**27.7**	23.5
XNLI	eng	45.2	45.6	**47.7**	44.6	43.3
XNLI	zho	45.1	**46.1**	45.6	45.4	45.1
INCLUDE	zho	26.8	27.1	26.5	22.0	**28.3**

Table 9: Evaluation results for eng−zho. The results are averaged across three random seeds.