

Can Translation Bridge the Data Gap for Low-Resource Languages? A Case Study of Hindi and Telugu

Venkata Naga Sai Vishnu Rohit Pulipaka and Libby Barak

Montclair State University

New Jersey, USA

{pulipakav1,barakl}@montclair.edu

Abstract

Low-resource languages often lack sufficient training data for developing small language models, limiting the application of recent language acquisition benchmarks beyond English. Moreover, comparison across languages is made harder given differences in training data composition and quality. We investigate whether machine translation can be used to create BabyLM-scale training corpora for low-resource languages by translating the English BabyLM corpus into Hindi and Telugu using IndicTrans2. We train GPT-BERT and GPT-Wee models and evaluate them using perplexity and linguistic benchmarks. GPT-BERT consistently outperforms GPT-Wee across both monolingual and bilingual settings. Although perplexity is lower on in-domain test data, this pattern is consistent across English and the translated languages, suggesting that the effect reflects domain similarity rather than artifacts introduced by translation. Furthermore, the models achieve competitive performance on linguistic evaluations relative to pretrained language models. These findings suggest that translation warrants further analysis to obtain data for developing and evaluating small language models in low-resource settings.

1 Introduction

The study of language acquisition in India has drawn attention from researchers across disciplines such as psychology and education, but has been less explored from a computational perspective. Indeed, the linguistic diversity of native speakers in India presents a complex challenge, with hundreds of languages and thousands of dialects spanning five language families (Mohanty, 2006). These languages also differ substantially in their available linguistic and computational resources, and many are considered low-resource from a computational standpoint. For example, Hindi has considerably larger resources compared to Telugu, though both

are official state languages in certain regions of India (Mohanty, 2010; Philip et al., 2021). Interestingly, English remains a constant across this landscape, given its widespread use in educational settings (Mohanty, 2006).

Recently, there has been growing interest in computational modeling for Indian languages, including the development of Large Language Models (LLMs) optimized for these language families (e.g., IndicBERT, MuRIL) (Dabre et al., 2022; Khanuja et al., 2021), and workshops dedicated to the development of models and resources for these languages (Jha et al., 2024). While most work on small language models has focused on English (see (Charpentier et al., 2025; Hu et al., 2024)), several studies have focused on languages other than English, including Chinese, Japanese, Slovak, and Hebrew (Gao et al., 2025; Kriš and Šuppa, 2025; Gelboim and Sulem, 2025).

Choshen et al. (2026) added non-English datasets to the BabyLM challenge for Chinese and Dutch. These datasets rely on the availability of sufficient data in these two languages to create a complementary resource to the English data, i.e., 100M words spanning across data from diverse genres. We ask whether training data can be created artificially for low-resource languages directly from the English data.

We use IndicTrans2 (Gala et al., 2023) to translate the 100M English data provided by the BabyLM workshop into Hindi and Telugu (Choshen et al., 2026). As a preliminary analysis, we focus on Hindi and Telugu out of the many Indic languages. Hindi, an Indo-Aryan language written in Devanagari, is one of the most widely spoken languages in India, while Telugu, a Dravidian language with a distinct grammatical structure and writing system, provides a complementary typological perspective and remains significantly under-resourced in computational studies. Although we do not use additional resources to construct the

translated data itself, the low-resource status of Telugu relative to Hindi may still affect translation quality, which we account for in our evaluation. A native speaker of Hindi and Telugu evaluated a sample of the translated sentences and verified translation quality.

We train both monolingual and bilingual models for each language and each language-English pairing. The use of data in three languages that stems from the same semantic and pragmatic coverage allows us for a more direct comparison of the results across all models. We evaluate the models on their ability to represent the translated language, the original data in each language, and the developmentally grounded measures provided by BabyLM (Choshen et al., 2026). Our results show that the translation results in similar byte ratio to a small parallel corpus data in line with suggested guidelines for data creation (Choshen et al., 2026; Arnett et al., 2024). GPT-BERT achieves superior performance to GPT-Wee for both monolingual and bilingual models on perplexity and linguistic performance measures. Across all three languages, perplexity was consistently lower on the BabyLM test data drawn from the same distribution as the training corpus compared with independent evaluation data. Because this pattern was observed for English as well as the translated Hindi and Telugu corpora, it suggests that the difference is primarily driven by domain similarity rather than noise introduced by the translation process. Despite being trained on translated corpora, the models performed well on the linguistic evaluations compared with pre-trained larger language models.

The consistency of these results across the three languages suggests that the LLM-based translation might provide a sufficiently stable training data to support further exploration of using this method to procure limited training data for small language models. We conclude by outlining suggestions for further work to validate this approach across additional languages and training methods.

2 Related Work

Language acquisition in India presents a complex challenge in its diverse native languages and differences across language exposure. In addition to the numerous official languages that stem from several language families, access to English widely differs depending on social-economic status and geographical location (Parshad et al., 2016). Computational

modeling can enable analysis of language acquisition through manipulation of degree of language exposure, input properties, and language combinations. While models of English language acquisition rely on data sets of child-directed speech and story books (MacWhinney, 2000; Gerlach and Font-Clos, 2020), CHILDES only covers several Indic languages and on a much smaller scale.

While most of the work on small language models focused on models of English, several studies explored the modeling non-English learning, developing data sets for other languages, and creating standard evaluation measures (Choshen et al., 2026). One approach to data creation has been to collect data from similar sources across languages (e.g., child-directed speech, newspaper data, subtitles, etc.). Previous work suggested synthetic data generation as a way to overcome the limited data availability and evaluated the approach on English as preliminary analysis (Ulm et al., 2025; Tampier et al., 2025). Yet, synthetic data generation showed mixed results as a source for training even in high-resource languages (Poh et al., 2025; Feng et al., 2024; Haga et al., 2024). Beyond data availability, models of multilingual learners offer an additional challenge of quantifying the effect of typological differences on the training process.

In this study, we propose to use the carefully curated English BabyLM corpus as the source for generating training data in low-resource languages through machine translation, following recent advances in multilingual translation models (Gala et al., 2023; Goyal et al., 2022). Using a shared source corpus provides comparable semantic and pragmatic coverage across languages, facilitating controlled comparisons between monolingual and bilingual models, similar to multilingual evaluation benchmarks (Liang et al., 2023; Jumelet et al., 2025). This approach offers the advantage of consistent thematic and register coverage across all datasets given the shared source, while reducing the need to locate diverse content types in multiple languages independently. However, translation may introduce noise and systematic biases while reducing language-specific cultural and topical variation (Artetxe et al., 2020; Bender et al., 2021), and translation quality may vary across target languages, potentially placing lower-resource languages such as Telugu at a disadvantage relative to Hindi. Our evaluation framework provides an initial approach to evaluating these trade-offs.

3 Methods

We use the English 100M words data (BabyLM-community/BabyLM-2026-Strict) provided by the BabyLM workshop (Choshen et al., 2026). We translate each of the sub-sections of the data: CHILDES (MacWhinney, 2000), British National Corpus (BNC) (BNC Consortium, 2007), dialogue portion, Project Gutenberg (children’s stories) (Gerlach and Font-Clos, 2020), OpenSubtitles (Lison and Tiedemann, 2016), Simple English Wikipedia (Wikimedia, 2023), and Switchboard Dialog Act Corpus (Stolcke et al., 2000). We translate the data to Hindi and Telugu, independently, using IndicTrans2 (Gala et al., 2023) (the en-indic-1B checkpoint), with source language code `eng_Latn` and target language codes `hin_Deva` and `tel_Telu` respectively. Sentences are pre- and post-processed using the IndicTrans Toolkit’s IndicProcessor. We translate with a batch size of 64 for Hindi and 32 for Telugu, beam search with beam size 5, and a maximum sequence length of 256 tokens. Since the training data is a translation of the original data release, the models in all languages theoretically represent equivalent semantic input that varies in other linguistic properties such as syntactic and morphological structure. We note that the resulting Hindi translation contains 118,309,059 words and 1,381,881,024 bytes, a word-to-byte ratio of 1:11.68. The Telugu translation contains 79,179,423 words and 1,471,312,814 bytes, a word-to-byte ratio of 1:18.58. For comparison, a parallel Hindi and Telugu corpus extracted from OpenSubtitles (Lison and Tiedemann, 2016) had word-to-byte ratios of 1:11.97 and 1:17.18, respectively.

We train monolingual and bilingual models using the code provided by the BabyLM workshop (Choshen et al., 2026), including GPT-Wee (based on GPT-2) and GPT-BERT (Bunzeck and Zarri  , 2023; Charpentier and Samuel, 2024a). The monolingual models are trained on data from a single language. The English model is trained in the full 100M word data, and the Hindi and Telugu models are trained on the full translation. To analyze the effect of language learning setting on the resulting model, we create bilingual models for each language pair: we combine translated data from CHILDES and Gutenberg with the original English data for the remaining subsets (BNC, OpenSubtitles, Simple Wikipedia, and Switchboard). This selection is intended to approximate naturalistic Indic-English bilingual exposure, in which the na-

tive/Indic language predominates in the home setting (CHILDES for child-directed language and Gutenberg for book reading) and English predominates in more formal or educational contexts (e.g., Wikipedia) (Mohanty, 2006).

We train each GPT-Wee model 2 times for each setup with different seeds, for a total of 10 models. The models are trained for 10 epochs with a learning rate of $5e-5$ and a warmup ratio of 0.01. Batch size is 4 for monolingual runs and 16 for bilingual runs. Total training steps are determined by dataset size, and no fixed gradient accumulation.

We train each model 2 times for each setup with different seeds, for a total of 10 GPT-BERT models. The parameters for training are a batch size of 32768 (global; 128 per GPU), max steps of 15625, 250 warmup steps, and a learning rate of $1e-2$. Gradient accumulation is dynamic rather than fixed. Validation is every 1000 steps for bilingual runs, while monolingual runs disable validation.

We evaluate the models with two complementary goals: (1) their ability to model natural language as reflected in corpus data and (2) their linguistic knowledge. For language modeling, we measure perplexity on two evaluation sets. The first is the BabyLM test set, which follows the same data distribution as the training corpus and therefore provides an in-domain evaluation of how well the models learn each language despite their typological differences. The second consists of parallel sentence pairs extracted from OpenSubtitles that do not overlap with any of the BabyLM subtitle data. Although this dataset is drawn from the same subtitle domain as 31% of the BabyLM training data, it represents an independent sample with no overlap with BabyLM data and therefore evaluates the models’ ability to generalize beyond the specific benchmark data. Moreover, the Hindi and Telugu models trained and tested on BabyLM data are both using the translation method, while the OpenSubtitles data measures their ability to generalize to data that is not obtained through the same translation methodology. Using parallel data ensures that all languages are evaluated on equivalent content, allowing differences in perplexity to reflect language modeling performance rather than differences in sentence meaning. Together, these two evaluations assess both in-domain language modeling and generalization to unseen subtitle data.

In addition, we analyze the linguistic abilities of the models using the measures made available by BabyLM that include benchmark for either Hindi,

Telugu or both. As such, we evaluate the monolingual and bilingual models using MultiBLiMP (Hindi only), which measures grammatical and syntactic knowledge through minimal sentence pairs; SIB200 (Hindi and Telugu), which evaluates document classification and semantic understanding across languages; and MuBench (Hindi and Telugu), which assesses multilingual language understanding on a range of downstream NLP tasks.

Our training and evaluation are primarily based on the BabyLM code and datasets. The additional code used for translation¹, and the translated data² are publicly available in the corresponding repositories.

4 Results

We evaluate the monolingual and bilingual language models using complementary intrinsic and extrinsic benchmarks designed to assess both language modeling performance and cross-lingual generalization. We first present perplexity results over all models and data sources, followed by an evaluation on linguistic abilities using three benchmarks following the BabyLM evaluation guidelines.

4.1 Perplexity

We evaluate monolingual language models trained on BabyLM data translated into Hindi and Telugu using IndicTrans2, comparing two model architectures: GPT-Wee and GPT-BERT (Bunzeck and Zarriß, 2023; Charpentier and Samuel, 2024a). We evaluate on two data sources: (1) held-out BabyLM test data translated into the target language using the same IndicTrans2 procedure as the training data, providing an in-distribution evaluation; and (2) OpenSubtitles (OS), a naturally occurring parallel corpus used to assess generalization beyond the training distribution. To assess the value of translated child-directed speech relative to native-language data, we additionally train GPT-BERT models on non-translated Hindi and Telugu data drawn from CC-100 (Wenzek et al., 2020), a web crawl corpus. For each language, we sample CC-100 data to match the byte size of the corresponding translated training corpus, ensuring a controlled comparison between translated and non-translated conditions.

Table 1 presents perplexity results across all models and evaluation sets. Importantly, all models are trained and tested on parallel data making the comparison more streamlined for semantic content. GPT-BERT outperforms GPT-Wee across all three languages on both evaluation sets, consistent with previous finding (Charpentier and Samuel, 2024b). For GPT-Wee, Hindi perplexity is lower than Telugu on both sets, while GPT-BERT shows the reverse pattern, achieving lower perplexity for Telugu than Hindi, though on a smaller scale. This crossover suggests that the two architectures interact differently with the typological properties of each language, or with the translation quality of each language pair, and warrants further investigation.

The Hindi and Telugu GPT-BERT models achieve markedly low perplexity scores (4.60 and 3.68 on BabyLM test data, respectively), substantially lower than the English GPT-BERT model (119.98). While this may partly reflect that the translated test data shares structural and lexical patterns with the training data more tightly than the English test set does, the magnitude of the difference suggests that tokenization may also play a role: subword segmentation for morphologically rich languages like Hindi and Telugu can produce longer token sequences that reduce per-token uncertainty independently of linguistic knowledge. We return to this point in the Discussion and flag it as a confound in interpreting absolute perplexity values across languages.

The increase in perplexity from BabyLM test data to the OpenSubtitles corpus is substantially larger for GPT-Wee in the translated languages, and particularly dramatic for Telugu (134.25 to 1343.92), indicating that the smaller model learns representations that do not generalize beyond the training distribution. GPT-BERT shows more stable generalization: for English, it achieves lower perplexity on OpenSubtitles than on the BabyLM test set (114.69 vs. 119.98). Hindi and Telugu GPT-BERT models show a larger increase from test to OS-data than English, but remain far more stable than GPT-Wee, suggesting that the translated corpora provide a reasonably stable training signal despite originating from machine translation.

Comparing translated and non-translated training conditions, the translated GPT-BERT models outperform their non-translated counterparts on BabyLM test data, as expected given the in-distribution nature of that evaluation. For both

¹<https://github.com/NLPlabMSU/babylm-translation-scripts>

²https://osf.io/973gp/?view_only=b255c4a0243341f498b6b243ddf528f6

Model	Test	OS-data
EN-GPT-Wee	140.5	292.35
HI-GPT-Wee	85.41	707.13
TE-GPT-Wee	134.25	1343.92
EN-GPT-BERT	119.98	114.69
HI-GPT-BERT	4.6	20.79
TE-GPT-BERT	3.68	17.62
NT-HI-GPT-BERT	11.17	8.36
NT-TE-GPT-BERT	8.53	12.13

Table 1: Results for the models trained on a single language (showing the mean across runs), GPT-Wee using BabyLM data (top panel), GPT-BERT using BabyLM data (middle panel), and GPT-BERT on non-translated data (bottom panel). Each column corresponds to testing on the data of the language used to train the model.

Hindi and Telugu, naturally occurring web data generalizes better to naturally occurring evaluation text. NT-HI-GPT-BERT outperforms HI-GPT-BERT on the OpenSubtitles data (8.36 vs. 20.79) and NT-TE-GPT-BERT outperforms the translated model on OpenSubtitles (12.13 vs. 17.62). This pattern may reflect translation quality, typological distance from English, or the adequacy of the CC-100 data as a representative corpus for each language, and represents an important direction for future work. Nevertheless, the difference in out-of-domain generalization between translated and non-translated models is modest, particularly for Telugu, suggesting that translated child-directed speech remains a competitive training source even beyond its native distribution — an encouraging result given the significantly lower cost of data creation via translation compared with sourcing naturally occurring corpora for low-resource languages.

Overall, these results suggest that IndicTrans2-based translation of child-directed speech offers a promising path toward cognitively plausible training data for low-resource languages, with translated data providing advantages in Hindi that do not fully extend to Telugu, underscoring the importance of evaluating this approach across multiple language pairs before drawing general conclusions.

We next train bilingual models on either English and Hindi or English and Telugu. Since English is commonly used in education settings in India, we use the data from CHILDES and Gutenberg in the translated form, and the original English for the other sources, which is roughly an even split between the two languages for each model. Table 2

presents the perplexity results for the bilingual models with the addition of two multilingual models: Llama 3.2 1B, and Sarvam-2b (Dubey et al., 2024; Sarvam AI, 2024). The pre-trained models provide a reference point for model performance using large data, which is non-translated and covers a wider linguistic scope, and training protocols that have been optimized for Indic languages. We run the same evaluation methods using the pre-trained models to allow more direct comparison.

The bilingual experiments further demonstrate the utility of translated data for multilingual language modeling. Because the bilingual models were trained on a fixed input size and similar semantic scope, the observed performance can be more directly interpreted as model sensitivity to language composition rather than properties of the input in each language. GPT-BERT maintained low perplexity on both English and the translated language, while exhibiting remarkably similar behavior for English–Hindi and English–Telugu training. Despite the substantially different resource availability of Hindi and Telugu, the bilingual models achieved comparable performance, suggesting that the translated corpora provide a stable basis for cross-lingual learning.

Though the translated data offers similar semantic content across all languages and models, we recognize that the results from monolingual and bilingual models may be influenced by the tokenizer training. We return to this issue in the Discussion when interpreting the cross-lingual results and their implications.

Finally, compared with larger multilingual models, the bilingual models show a mixed pattern. Llama 3.2 1B achieves lower perplexity on English, consistent with the benefits of large-scale pretraining. However, the gap is smaller for Hindi and Telugu, where the bilingual GPT-BERT models achieve comparable or lower perplexity than Sarvam-2B despite being substantially smaller. These results suggest that carefully constructed translated training data can enable small models to learn effective representations for targeted languages, although larger models maintain an advantage in broader language modeling settings.

4.2 Linguistic Knowledge

We next evaluate whether the representations learned from translated training data extend beyond perplexity-based evaluation. Specifically, we consider the multilingual benchmarks available

Model	English		Hindi		Telugu	
	Test	OS-data	Test	OS-data	Test	OS-data
EN-HI-GPT-Wee	233.69	448.11	93.05	879.17	—	—
EN-TE-GPT-Wee	199.26	365.27	—	—	270.56	3937.21
EN-HI-GPT-BERT	82.55	188.63	8.37	30.67	—	—
EN-TE-GPT-BERT	84.49	187.90	—	—	4.25	19.10
Llama 3.2 1B	60.45	46.98	24.97	21.07	3.70	4.41
Sarvam-2B	186.58	76.14	210.51	229.16	205.81	2350.40

Table 2: Results for the models trained on two languages, and general-purpose baselines (showing the mean across runs). Each column group corresponds to testing on English, Hindi, or Telugu data, regardless of which language pair the model was trained on.

Model	English			Hindi			Telugu	
	MBLIMP	SIB200	MuBench	MBLIMP	SIB200	MuBench	SIB200	MuBench
EN-GPT-Wee	0.7469	0.3284	0.2762	-	-	-	-	-
HI-GPT-Wee	-	-	-	0.9261	0.2353	0.2750	-	-
TE-GPT-Wee	-	-	-	-	-	-	0.2108	0.2832
EN-GPT-BERT	0.7978	0.2451	0.2823	-	-	-	-	-
HI-GPT-BERT	-	-	-	0.9399	0.1961	0.2809	-	-
TE-GPT-BERT	-	-	-	-	-	-	0.1471	0.2822
EN-HI-GPT-Wee	0.7070	0.2647	0.2827	0.9150	0.1324	0.2729	-	-
EN-TE-GPT-Wee	0.7136	0.2794	0.2803	-	-	-	0.2304	0.2765
EN-HI-GPT-BERT	0.7868	0.2500	0.2768	0.9136	0.2304	0.282	-	-
EN-TE-GPT-BERT	0.7991	0.2500	0.2709	-	-	-	0.1961	0.2813
Sarvam-2B	0.8006	0.4706	0.2833	0.9772	0.4510	0.2854	0.3824	0.2768
Llama 3.2 1B	0.8188	0.6275	0.2786	0.9772	0.6078	0.2796	0.43	0.277

Table 3: Linguistic knowledge results (BLiMP/MultiBLiMP, SIB-200, MuBench) across English, Hindi, and Telugu (showing the mean across runs).

for Hindi and Telugu: MultiBLiMP (Hindi only), SIB-200 (Hindi and Telugu), and MuBench (Hindi and Telugu). As for the bilingual models, we run the same evaluation methods using the pre-trained models to allow more direct comparison.

Table 3 presents the results for all monolingual and bilingual models across the three languages. Note that each model is only evaluated on languages included in its training data, e.g., the bilingual English-Hindi model is not tested in Telugu. In addition, Telugu is not tested on MultiBLiMP since the benchmark does not include this language. Unlike the results for perplexity measures, the two models, GPT-Wee and GPT-BERT, score mostly similarly across the languages and measures without one model being consistently superior to the other across the board.

Across linguistic evaluations, the effect of model scale depends strongly on the type of capability being measured. Small models trained on translated BabyLM-scale corpora achieve strong performance on MultiBLiMP, indicating that translated data can support learning of syntactic patterns in both Hindi and English. Unlike the perplexity results, Telugu

performance on SIB-200 is considerably lower, suggesting that lower language modeling perplexity does not necessarily translate into stronger syntactic generalization. The larger multilingual models substantially outperform small models on SIB-200, suggesting that broader semantic classification benefits from additional parameters and training resources. Interestingly, the gap between the small models and larger multilingual models is considerably smaller for Telugu than for English and Hindi, suggesting that model scale interacts differently with language-specific factors, including training data availability and typological differences. In contrast to MultiBLiMP and SIB-200, MuBench showed little variation across models regardless of architecture, language, or model scale. This may suggest that the benchmark is less sensitive to differences introduced by translated training data or multilingual modeling in our study. While this uniformity could reflect a genuine insensitivity of the benchmark to differences in training data composition, it may alternatively indicate that all models — regardless of size or training language — converge on a similar performance floor for the task

types included in MuBench, which would itself be a finding worth investigating in future work.

Overall, the bilingual models show only modest degradation relative to monolingual models, indicating that multilingual training with translated corpora preserves linguistic competence while enabling cross-lingual modeling. Some English–Hindi and English–Telugu combinations outperform their monolingual counterparts and others showing slight degradations. The absence of a consistent trend suggests that the interaction between the paired languages, rather than the translated data itself, plays a larger role in shaping the performance.

5 Discussion

This work explores the use of LLM-based translation to overcome limitations in training data availability for low-resource language contexts. Using the English BabyLM corpus as the source for all translated datasets provides consistent semantic content and comparable distributions across data types and modalities, enabling controlled cross-lingual comparisons. At the same time, this design may bias the learned representations toward English and reduce language-specific cultural and linguistic variation. Furthermore, translation quality may differ across target languages, potentially placing lower-resource languages such as Telugu at a disadvantage relative to Hindi. Nevertheless, the potential benefits of generating training data for languages with limited existing resources motivate this approach. Our evaluation framework is designed to account for these trade-offs by examining language modeling, linguistic competence, and multilingual transfer across both monolingual and bilingual settings.

Overall, our results suggest that the translated Hindi and Telugu corpora preserve sufficient linguistic information to support the training of small language models. Both GPT-Wee and GPT-BERT successfully learned the translated corpora, with GPT-BERT consistently achieving lower perplexity and stronger linguistic performance across monolingual and bilingual settings. While perplexity was substantially lower on the held-out BabyLM test sets than on independent evaluation data, this pattern was observed consistently across English, Hindi, and Telugu, suggesting that it primarily reflects domain similarity rather than artifacts introduced by translation.

One result that warrants further attention is the crossover between GPT-Wee and GPT-BERT in their relative performance on Hindi versus Telugu. GPT-Wee achieves lower perplexity on Hindi than Telugu, while GPT-BERT shows the reverse pattern. We hypothesize that this interaction reflects a combination of translation quality differences. Telugu, being lower-resource may result in noisier translations that benefit more from the greater capacity of GPT-BERT and tokenization effects, whereby the more agglutinative morphology of Telugu produces longer subword sequences that interact differently with each model’s training objective. Disentangling these factors would require controlled experiments varying translation quality and tokenizer design independently, which we leave for future work.

The downstream evaluations reveal a more nuanced picture. Models trained on translated corpora achieved strong performance on MultiBLiMP, indicating that translated data preserve the syntactic regularities required for grammatical generalization. In contrast, larger multilingual models maintained a substantial advantage on SIB-200, particularly for English and Hindi. Interestingly, the performance gap was considerably smaller for Telugu. This pattern suggests that balanced language representation during training may be particularly beneficial for modeling low-resource languages, although further work is needed to disentangle the effects of language balance, model scale, and training data composition.

This work also opens several directions for future research. While the translated corpora provide a controlled framework for comparing language learning across English, Hindi, and Telugu, the approach should be validated on a broader range of evaluation data, including additional domains, registers, and English varieties spoken by Hindi speakers. Such evaluations would provide a more comprehensive assessment of out-of-domain generalization and help disentangle the effects of translation quality, domain similarity, and language contact. Future work should also examine the role of tokenization, as byte-pair encoding (BPE) may be less suitable for morphologically rich or agglutinative languages, potentially affecting both language modeling and downstream performance.

An important consideration when training models on translated data is the possible presence of systematic linguistic properties that distinguish translated language from naturally produced language also known as translationese. Previous work

has demonstrated that translationese can be readily identified and can introduce systematic differences in lexical, syntactic, and stylistic properties (Kunz, 2026; Vanmassenhove et al., 2021; Rabinovich et al., 2016; Rabinovich and Wintner, 2015; Koppel and Ordan, 2011). While our evaluation mostly focused on linguistic abilities, which are of interest in the context of language acquisition and learning, further work will evaluate the use of translation as training data as reflected in more culture-dependent benchmarks and downstream tasks.

While additional evaluation is needed to validate this approach across languages and benchmarks, the translated BabyLM framework also provides a controlled setting for investigating more nuanced questions of bilingual language acquisition. Unlike naturally occurring corpora, the translated data allow systematic manipulation of the linguistic environment, including the proportion of English and translated input (e.g., child-directed vs. educational content in each language), the degree of code-switching, and the balance of language exposure during training. Such controlled rearing conditions make it possible to examine how different patterns of bilingual experience influence language learning and to compare the resulting developmental trajectories with those observed in native and heritage bilingual speakers. More broadly, this work demonstrates that machine-translated developmental corpora can serve not only as a practical resource for low-resource language modeling, but also as an experimental framework for studying multilingual language acquisition under carefully controlled conditions.

Limitations

We acknowledge several limitations of our study. First, due to computational resource constraints, we trained and evaluated a limited number of model configurations. Additional runs would allow for a more robust estimation of variance and a better understanding of the stability of the observed differences across languages and model architectures. Similarly, exploring a broader range of model choice for translation and target languages would provide a more complete picture of how translated corpora influence multilingual learning. Additionally, our evaluation is limited by the availability of linguistic benchmarks for the target languages. While MultiBLiMP provides a valuable measure of syntactic knowledge, it is not currently

available for Telugu.

Although translation provides a controlled approach for constructing comparable training corpora across languages, it may also introduce biases toward the source language and reduce naturally occurring linguistic and cultural variation. Future work should examine additional translation methods, native-language resources, and alternative evaluation settings to better characterize the trade-offs between controlled multilingual training data and naturally occurring language input.

Finally, as noted in Section 4, subword tokenization using byte-pair encoding (BPE) may disproportionately affect perplexity comparisons across languages. Because BPE segments morphologically rich languages such as Hindi and Telugu into more subword units per word than English, the effective prediction task differs in granularity across languages, potentially lowering per-token perplexity independently of linguistic knowledge. Models such as Sarvam-2B are trained with tokenizers specifically optimized for Indic scripts, which may account for some of the differences observed in cross-model comparisons. For this reason, perplexity values should be interpreted within the same language and model family rather than across them, and future work should evaluate tokenization-controlled perplexity measures or character-level evaluation alternatives.

References

- 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.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- BNC Consortium. 2007. The british national corpus, xml edition.
- Bastian Bunzeck and Sina Zarrieß. 2023. [GPT-wee: How small can a small language model really get?](#)

- In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore. Association for Computational Linguistics.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, and 1 others. 2025. BabyLM turns 3: Call for papers for the 2025 babyLM workshop. *arXiv preprint arXiv:2502.10645*.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024a. **GPT or BERT: why not both?** In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024b. **GPT or BERT: why not both?** In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.
- 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. *arXiv preprint arXiv:2602.20092*.
- Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. **Is child-directed speech effective training data for language models?** In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. **Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages**. *Transactions on Machine Learning Research*.
- Yuan Gao, Suchir Salhan, Andrew Caines, Paula Buttery, and Weiwei Sun. 2025. Bliss: Evaluating bilingual learner competence in second language small language models. In *Proceedings of the First BabyLM Workshop*, pages 160–174.
- Anita Gelboim and Elior Sulem. 2025. Tafberta: Learning grammatical rules from small-scale language acquisition data in hebrew. In *Proceedings of the First BabyLM Workshop*, pages 76–90.
- Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenber corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. **BabyLM challenge: Exploring the effect of variation sets on language model training efficiency**. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 252–261, Miami, FL, USA. Association for Computational Linguistics.
- Michael Y Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second babyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21.
- Girish Nath Jha, Sobha L., Kalika Bali, and Atul Kr. Ojha, editors. 2024. *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*. ELRA and ICCL, Torino, Italia.
- Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. **Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs**. *Preprint*, arXiv:2504.02768.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.
- Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326.
- L’uboš Kriš and Marek Šuppa. 2025. SlovkbabyLM: Replication of the babyLM and sample-efficient pre-training for a low-resource language. In *Proceedings of the First BabyLM Workshop*, pages 301–312.

- Jenny Kunz. 2026. [Preferences for idiomatic language are acquired slowly — and forgotten quickly: A case study on swedish](#). *Transactions of the Association for Computational Linguistics*, 14:1266–1285.
- Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. [XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. 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).
- Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.
- Ajit K Mohanty. 2006. Multilingualism of the unequals and predicaments of education in india: Mother tongue or other tongue. *Imagining multilingual schools*, 262:283.
- Ajit K Mohanty. 2010. Languages, inequality and marginalization: implications of the double divide in indian multilingualism. *International Journal of the Sociology of Language*, 2010(205).
- Rana D Parshad, Suman Bhowmick, Vineeta Chand, Nitu Kumari, and Neha Sinha. 2016. What is india speaking? exploring the “hinglish” invasion. *Physica A: Statistical Mechanics and its Applications*, 449:375–389.
- Jerin Philip, Shashank Siripragada, Vinay P Namboodiri, and CV Jawahar. 2021. Revisiting low resource status of indian languages in machine translation. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 178–187.
- Whitney Poh, Michael Tombolini, and Libby Barak. 2025. What did you say? generating child-directed speech questions to train llms. In *Proceedings of the First BabyLM Workshop*, pages 237–245.
- Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1881.
- Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.
- Sarvam AI. 2024. Sarvam 2b: A small language model for indian languages. <https://huggingface.co/sarvamai/sarvam-2b-v0.5>.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.
- Alexander Tampier, Lukas Thoma, Loris Schoenegger, and Benjamin Roth. 2025. Recombitext: Compositional data augmentation for enhancing llm pre-training datasets in low-resource scenarios. In *Proceedings of the First BabyLM Workshop*, pages 548–565.
- Jannek Ulm, Kevin Du, and Vésteinn Snæbjarnarson. 2025. Contrastive decoding for synthetic data generation in low-resource language modeling. In *Proceedings of the First BabyLM Workshop*, pages 29–41.
- Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume*, pages 2203–2213.
- 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.
- Wikimedia. 2023. Simple english wikipedia dump. <https://dumps.wikimedia.org/simplewiki/20230301/>. Accessed: 2023-07-31.