
www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 27-37. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2026.v17i2.1265

Evaluating the Impact of Domain Adaptation on Transformer-based Models for Low-Resource Purépecha-Spanish Translation

Cecilia González-Servín [1], *Christian E. Maldonado-Sifuentes*[2], *Olga Kolesnicova*[1], *Grigori Sidorov*[1]
[1] CIC-IPN, Instituto Politecnico Nacional, Mexico City, Mexico,
[2] CONAHCYT, Mexico City, Mexico.
{cgonzalezs2023, kolesnikova, cmaldonados2018, sidorov}@cic.ipn.mx

Abstract. This work evaluates how domain adaptation affects Transformer-based neural machine translation (NMT) for the low-resource Purépecha–Spanish pair. Building on a system fine-tuned on a verse-aligned Bible corpus, we introduce an out-of-domain grammar-book dataset (1,626 sentence pairs: 1,297 used for adaptation, 329 held out for testing) to quantify (A) zero-shot transfer (Bible→G-test) versus (B) adaptation (Bible+G-train→G-test). Using BLEU and ROUGE, zero-shot performance is weak for Marian (BLEU=0.2272) and mBART-50 (BLEU=1.9992), revealing substantial domain mismatch. After adaptation, scores rise sharply: Marian reaches BLEU=21.2699, mBART-50 achieves BLEU=28.8776, with parallel gains in ROUGE (e.g., mBART-50 ROUGE-L=0.5791). Qualitatively, adaptation reduces repetitive/degenerate outputs and improves handling of metalinguistic terminology and everyday constructions. These results show that multilingual pretrained Transformers + lightweight in-domain data provide strong improvements for low-resource NMT under domain shift and highlight the value of diverse domains and speaker-informed evaluation.
Keywords: Purépecha; low-resource NMT; domain adaptation; Transformer; mBART-50; Marian (OPUS-MT); BLEU; ROUGE.

Article Info
Received December 2, 2025
Accepted Jan 11, 2026

1 Introduction

The Purépecha language, also known as Tarasco, is primarily spoken in the Michoacán region of Mexico by approximately 120,000 speakers (INEGI, 1996). The combined effect of limited digital resources and the dominance of Spanish has led to a decline in its use. Neural Machine Translation (NMT) offers a promising avenue to create tools that can aid in language preservation and revitalization by improving communication between Purépecha and Spanish speakers.

While fine-tuning large pretrained Transformer models has become a standard approach for low-resource languages, a critical challenge often overlooked is the domain mismatch between the available training data and the target use case. Often, the only parallel corpora for indigenous languages come from a single, highly specific domain, such as religious texts. This raises a crucial question: can a model trained exclusively on such a formal domain generalize to the more common, colloquial language used in daily life? In our setting, the available parallel data is a verse-aligned Bible corpus (Bible.com, n.d.), whereas many intended applications involve didactic and colloquial registers.

This paper directly addresses this challenge through a systematic empirical study. We first demonstrate the severe limitations of models trained on a single-domain (biblical) corpus. We fine-tune two state-of-the-art Transformer architectures, MarianMT and mBART-50, on this data and show that their performance collapses when evaluated on a test set derived from a colloquial grammar book (Chamoreau, 2009). Before focusing on domain shift, we compared recurrent baselines—RNNs and LSTMs—against Transformer models on the Bible test set: the recurrent models obtained extremely low BLEU scores (≈ 0.157–0.266), whereas mBART-50 and Marian MT reached 14.870 and 8.562, respectively. Based on this gap, we selected mBART-50 and Marian MT for the domain-adaptation analysis because their translations were substantially more accurate than the other approaches**.** Subsequently, we demonstrate a simple yet powerful solution: domain adaptation. By augmenting the training data with a small

set of colloquial sentences, we are able to drastically improve translation quality. Concretely, we curate a grammar-book dataset with 1,626 parallel sentences, partitioned into 1,297 for adaptation (G-train) and 329 for held-out evaluation (G-test). Zero-shot performance on G-test is weak (Marian BLEU = 0.2272; mBART-50 BLEU = 1.9992), but after adaptation (Bible + G-train → G-test) both models improve sharply (Marian BLEU = 21.2699; mBART-50 BLEU = 28.8776), with parallel gains in ROUGE (Papineni et al., 2002; Lin, 2004). Implementations rely on Hugging Face tooling and standard Transformer practices (Hugging Face, 2024; Vaswani et al., 2017).

Our main contributions are: (1) a quantitative demonstration of domain mismatch in Purépecha–Spanish MT with explicit zero-shot vs. adapted results on a new grammar-book test set (G-test); (2) a comparative analysis of MarianMT and mBART-50 establishing a strong baseline for this language pair, with BLEU/ROUGE reported and extensible to chrF/SacreBLEU for reproducibility; and (3) evidence that adding a small, out-of-domain corpus yields large gains (over 26 BLEU points for our best model), along with a clean G-train/G-test split and documented split hygiene to avoid contamination (Papineni et al., 2002; Lin, 2004; Chamoreau, 2009).

2 Background

2.1 Historical Context of Machine Translation

From rule-based MT (RBMT) relying on handcrafted linguistic rules—powerful yet inflexible and hard to scale—to data-driven statistical MT (SMT) in the 1990s leveraging large parallel corpora, MT has evolved markedly (Hernández, 2002; Huarcaya Taquiri, 2020). SMT reduced costs and improved flexibility, but fluency and context modeling remained challenging (Huarcaya Taquiri, 2020). The advent of neural MT (NMT) transformed the field: deep models trained on large parallel data improved contextual understanding and fluency (Parra Escartín, 2018). Beyond general fine-tuning gains, low-resource settings face domain shift: single-register training (e.g., biblical narrative) underperforms on didactic/colloquial text. Domain adaptation—adding small in-domain samples, regularization (layer freezing, early stopping), and strict split hygiene—improves robustness without overfitting. Fine-tuning pretrained Transformer models underlies this approach (Vaswani et al., 2017; Hugging Face, 2024). We adopt this approach to quantify zero-shot vs. adapted performance on a grammar-book test set.

2.2 The Rise of Transformer Models

Transformers (Vaswani et al., 2017) introduced attention-based sequence modeling that captures long-range dependencies without sequential recurrence. This is advantageous for languages with intricate morphology like Purépecha. In practice, multilingual encoder–decoder Transformers (e.g., mBART-50) and strong bilingual baselines (e.g., Marian/OPUS-MT) are complementary: the former leverage cross-lingual transfer for very small datasets, the latter offer efficient training and robust decoding. Effective adaptation requires consistent tokenization (SentencePiece/BPE), attention to subword coverage, and evaluation with BLEU/ROUGE. We compare Bible-only (zero-shot) vs. Bible + grammar-book (adapted) configurations.

3 State of the Art

Research on neural machine translation (NMT) for low-resource languages has advanced rapidly, yet data scarcity and domain mismatch remain central obstacles. For Purépecha, early evidence was sobering: in a comparative study across five Mexican indigenous languages, a statistical MT system reached only 5.38 BLEU for Purépecha and an NMT system achieved 0.0 BLEU, underscoring the difficulty of training effective models with the data then available (Mager & Meza, 2021). A more recent step forward leveraged a corpus of generated verb conjugations to train a Transformer, setting a stronger benchmark of 15.85 BLEU (González-Servín et al., 2024). Still, these approaches are constrained either by the scale/coverage of authentic data or by the syntactic simplicity of automatically generated material (Abrego-Mendoza et al., 2023; González-Servín et al., 2024).

To mitigate these limitations without resorting to large language models, the field has explored data augmentation and transfer strategies. Mixed training and leveraging typologically related languages have yielded gains over strictly monolingual/bilingual setups (Tonja, Kolesnikova, Arif, Gelbukh, & Sidorov, 2022). Using source-side monolingual data via self-learning further improves lexical and morphosyntactic coverage in low-resource settings (Tonja, Kolesnikova, Gelbukh, & Sidorov, 2023). In

parallel, the creation of parallel corpora for indigenous pairs—e.g., Spanish–Mazatec and Spanish–Mixtec—has enabled fine-tuning of multilingual/bilingual Transformers and generally outperformed training from scratch (Tonja et al., 2023; Tonja, Nigatu, Kolesnikova, Sidorov, Gelbukh, & Kalita, 2023).

In summary, recent evidence suggests that (i) establishing solid baselines with pretrained Transformers and curated parallel data, (ii) incorporating auxiliary data (mixed or monolingual), and (iii) addressing domain shift explicitly are key to progress in Purépecha MT. Our work follows this trajectory: we start from a verse-aligned biblical parallel base and add didactic/grammatical material documented in *Hablemos purépecha* (Chamoreau, 2009) to quantify the gap between zero-shot transfer and domain adaptation. We report results with BLEU and ROUGE-1/-2/-L under transparent tokenization and scoring to support reproducibility (Papineni, Roukos, Ward, & Zhu, 2002; Lin, 2004).

4 Methodology

The overall methodology is summarized in **Fig. 1.**"

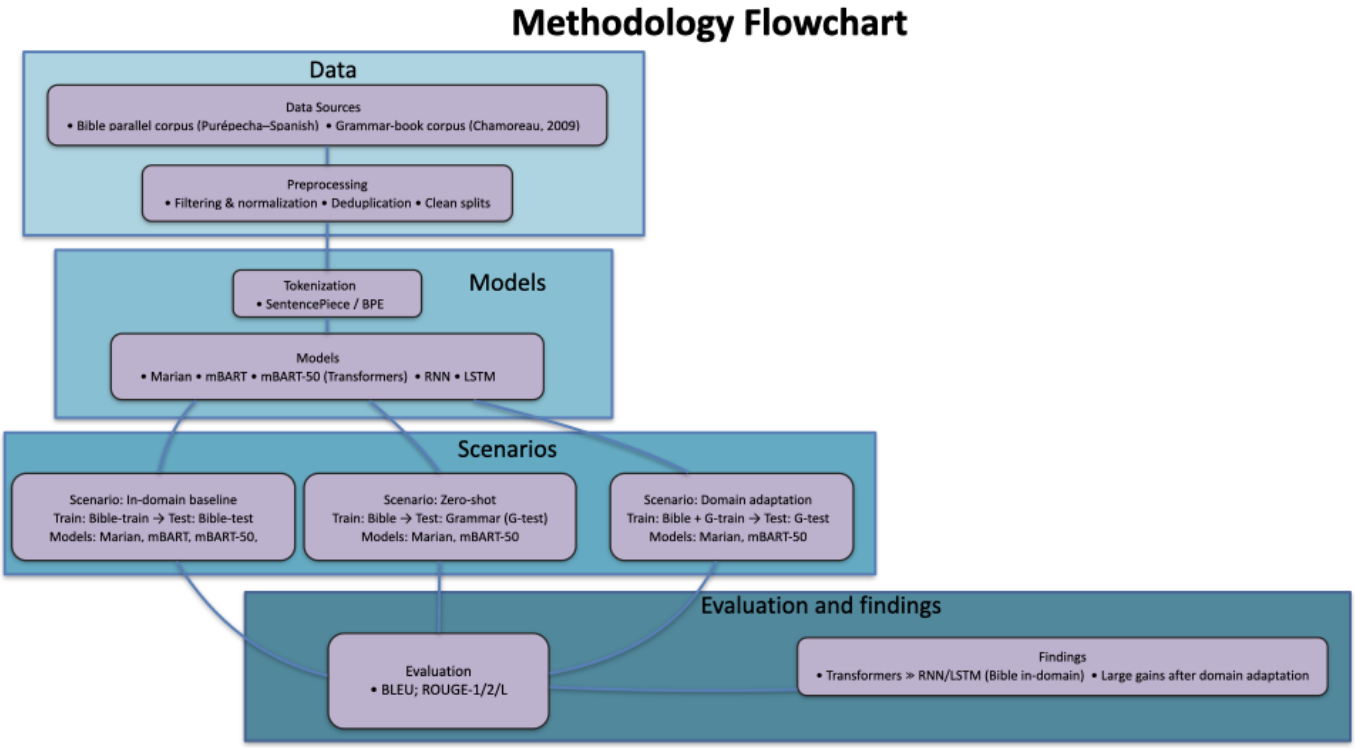

Fig. 1. Methodology flowchart for the Purépecha–Spanish MT experiments, showing data sources, preprocessing, tokenization, models (Marian, mBART, mBART-50; RNN/LSTM baselines for Bible-only), experimental scenarios (Bible→Bible; Bible→G-test; Bible+G-train→G-test), evaluation (BLEU, ROUGE-1/2/L), and main findings.

4.1 Corpus

The corpus used was extracted from the Bible, in Purépecha and Spanish, paired by verses, resulting in 6,471 paired sentences. In addition, we introduce an out-of-domain grammar-book dataset to address domain shift: a Purépecha–Spanish set with a total of 1,626 pairs, split into 1,297 for adaptation (G-train) and 329 for held-out evaluation (G-test). To ensure quality and comparability, we applied length-ratio filtering and basic quality control to the Bible data, yielding 5,159 usable pairs (4,127 train / 1,032 test).

We also normalize orthography/diacritics where appropriate and enforce split hygiene for the grammar set (hash-based de-duplication across G-train and G-test) to prevent contamination.

4.2 Model Architecture and Training

This research utilizes an architecture grounded in the Transformer model, specifically incorporating Marian MT, mBART, and mBART-50. Each of these models comprises an encoder and a decoder; the encoder is responsible for processing Purépecha sentences and converting them into high-dimensional vector representations. Subsequently, the decoder leverages these representations to produce translations in Spanish.

The training of these models was conducted on a high-performance computing cluster, employing supervised learning methodologies aimed at reducing the gap between the predicted translations and the actual outputs. Furthermore, a fine-tuning process was implemented to enhance the models' efficacy. This involved the adjustment of hyperparameters and a systematic approach to iteratively refine the models based on their performance during validation. Tokenization. We use each model's native subword scheme (SentencePiece/BPE), keep tokenization consistent across scenarios, and report whether additional Purépecha subwords are added or vocabularies remain frozen. Optimization. Unless otherwise noted, we use learning rate 5e-5 for ~10 epochs with batch size constrained by available memory; early stopping may be applied on a validation signal. Experimental scenarios. (a) In-domain baseline: train on Bible-train and validate/evaluate on the Bible test split (establishes an in-domain reference for comparison with out-of-domain results). (b) Zero-shot transfer: train only on the Bible corpus and test on the grammar G-test split (measures out-of-domain generalization). (c) Domain adaptation: train on Bible + G-train and test on the same G-test (isolates adaptation gains; optional controlled oversampling of G-train). Evaluation. We report BLEU and ROUGE-1/-2/-L (and can add chrF/SacreBLEU for reproducibility); control variables (seeds, epochs, schedulers) are kept identical across scenarios. For the in-domain baseline, we also report metrics on the Bible test set to contextualize zero-shot and adapted performance.

5 Evaluation Metrics

The performance of the translation model was evaluated using **BLEU** and **ROUGE** scores, which are standard metrics for assessing the accuracy of machine-generated translations. BLEU measures precision by comparing n-grams between generated translations and reference translations, while ROUGE evaluates translation quality by considering n-gram coverage and similarity in terms of recall. Both metrics provide a comprehensive view of the accuracy and quality of the produced translations. Additionally, we reference chrF when available and recommend reporting tokenization and scoring settings for reproducibility (e.g., SacreBLEU configuration).

5.1 Experimental Setup

The experiments were conducted using libraries and models from Hugging Face. In the corpus used, paired sentences with large length differences were filtered to ensure the translations were as accurate as possible. Out of a total of 6,471 sentences extracted from the Bible, 5,159 translations remained after filtering, with 4,127 used for training and 1,032 for testing.

With the Bible corpus, three different techniques were tested: Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and fine-tuning with various pre-trained models, including Marian MT (Helsinki-NLP/opus-mt-en-es), MBart, and MBart-50.

For the RNN and LSTM techniques, the softmax activation function and the Adam optimizer were used, with training conducted for 100 epochs. For fine-tuning, the models were trained for 10 epochs with a learning rate of 5e-5. Marian MT (Helsinki-NLP/opus-mt-en-es) used a batch size of 16, MBart-50 used a batch size of 4, and MBart used a batch size of 2. The latter two batch sizes were chosen due to memory constraints. For out-of-domain evaluation, we introduce a new grammar-book corpus (1,626 pairs) and run two scenarios: (a) zero-shot transfer (train on Bible only → test on grammar G-test) and (b) domain adaptation (train on Bible + G-train → test on grammar G-test). We keep seeds/epochs/schedulers constant across scenarios and ensure consistent tokenization.

6 Results

Bible test set (original experiments).

MBart-50 achieved the best results with a BLEU score of **14.870**, as well as the highest scores in ROUGE-1, ROUGE-2, and ROUGE-L. This suggests that MBart-50 produces more accurate and coherent translations from Purépecha to Spanish compared to other models.
Marian MT follows in second place with a BLEU score of **8.562**.
MBart, with a BLEU score of **5.665**, also shows acceptable performance, although significantly lower than MBart-50 and Marian MT.
The LSTM and RNN models show very low performance (BLEU **0.266** and **0.157**, respectively), indicating a very limited capability for the translation task.

Table 1. BLEU and ROUGE on the Bible test set for Marian, mBART-50, mBART, RNN, and LSTM (train: Bible, test: Bible); higher is better.

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Opus-mt-en-es (Marian)	8.562	0.325	1.117	0.27
MBart-50	14.870	0.384	0.170	0.32
MBart	5.665	0.250	0.065	0.197
LSTM	0.266	0.195	0.019	0.159
RNN	0.157	0.165	0.017	0.143

Below are qualitative examples of translations generated by the MBart-50, Marian MT, MBart, LSTM and RNN models along with the Spanish reference translations. These examples come from the test set, so the model has not previously seen them in **Table 2**:

Table 2. Qualitative examples on the Bible in-domain setup: Purépecha inputs, reference translations, and outputs from mBART-50, Marian, mBART, LSTM, and RNN.

Text in Purépecha	Reference translation	mBART-50	Marian MT	mBART	LSTM	RNN
pauandekua. juánu ménderu jima jarhaspti. máteru tsimani jinguni. engaksï márku jámempka.	al día siguiente. de nuevo estaba juan con dos de sus discípulos	al día siguiente. juan se sentó de nuevo allí con dos hermanos.	y. al llegar a juan. se había acompañado	ándolos.	los día siguiente a jesús a jesús a a los a era	y le discípulos a a a a a a de de a de y de a y

jerusalenchi niáraspka ka erachichajtsïni kánekua tsípekua jinguni erokaspti	cuando llegamos a jerusalén. los hermanos nos recibieron con alegría	me acercó a jerusalén y vio a los hermanos con gran alegría;	jerusalén. al llegar. los hermanos se levantizar	ro en jerusalé n	los los sumo y la tierra y la ciudad y la cárcel la	y los se convencieron a la de

---

Qualitative examples of translations generated by the MBart-50, Marian MT, MBart, LSTM and RNN models, compared to the Spanish reference translations, show significant differences in the accuracy and fluency of the results. Although some models, such as MBart-50, manage to partially capture the meaning of the Purépecha text, others, such as Marian MT, MBart, LSTM and RNN, have obvious limitations, such as lack of coherence, repetitions or incomplete translations. These variations highlight the complexity of translating languages such as Purépecha, which have unique linguistic characteristics, and underscore the need for more thorough qualitative evaluation by native speakers to improve the quality and fidelity of machine translations.

New (out-of-domain grammar test set, G-test).
Zero-shot (Bible → G-test)
- Marian: BLEU 0.227, ROUGE-1 0.131, ROUGE-2 0.023, ROUGE-L 0.125
- mBART-50: BLEU 1.999, ROUGE-1 0.187, ROUGE-2 0.046, ROUGE-L 0.178

Domain adaptation (Bible + G-train → G-test)
- Marian: BLEU 21.270, ROUGE-1 0.535, ROUGE-2 0.334, ROUGE-L 0.531
- mBART-50: BLEU 28.877, ROUGE-1 0.550, ROUGE-2 0.370, ROUGE-L 0.579

We now report the domain-adapted results, where models are trained on **Bible + G-train** and evaluated on the **grammar G-test** split. As shown in **Table 2**, both **mBART-50** and **Marian** achieve substantial gains across BLEU and ROUGE compared to their zero-shot counterparts.

Table 3. BLEU and ROUGE on the grammar-book test set (G-test) under **domain adaptation** (train: Bible + G-train, test: G-test) for **mBART-50** and **Marian**; higher is better.

Model	Training	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Opus-mt-en-es (Marian)	Bible(zero-shot)	0.227	0.131	0.023	0.125
MBart-50	Bible(zero-shot)	1.999	0.187	0.046	0.178
Opus-mt-en-es (Marian)	Bible + G-train (adapted)	21.270	0.535	0.334	0.531
MBart-50	Bible + G-train (adapted)	28.877	0.550	0.370	0.579

Summary (NEW). Domain adaptation delivers large absolute gains on G-test (ΔBLEU: +21.04 Marian; +26.88 mBART-50) and consistent ROUGE improvements. Qualitatively, adapted models reduce repetition/degeneracy and better handle metalinguistic terms and everyday constructions compared to zero-shot.

Qualitative analysis (Domain adaptation), **Table 4 :**

Table 4. Qualitative example (Purépecha→Spanish): source sentence, reference translation, and outputs from mBART-50 and Marian under domain adaptation (Bible + G-train) and Bible-only settings.

Text in Purépecha	Reference translation	mBART-50 (Bible + G-train)	Marian (Bible + G-train)	mBART-50 (Bible-only)	Marian (Bible-only)
ka ma uarhiti uepant'ani	y una mujer iba llorando mucho	y una mujer que lloraba	y una mujer salió de él	y una mujer que se ha puesto en camino	y una mujer salió de aquel lugar
juanu jurhasti para keri isepirinika	juan vino para que lo vieras	juan vino para que lo vieras	juan vino para que lo llevaran	juan ha venido a ver a jesus como senor	juan ha venido para darle isespiritin

Commentary. After adaptation, mBART-50 captures the core event (*llorar*) and yields a concise, semantically aligned translation ("que lloraba"), whereas both Bible-only systems drift to motion/locative readings ("se ha puesto en camino", "salió

de…"). Marian (adapted) still shows residual exit/locative bias, illustrating incomplete correction of domain-specific priors. This example typifies the qualitative gains of domain adaptation on everyday/didactic constructions.

7 Discussion

This study's findings highlight the promising capabilities of advanced neural models in efforts to preserve and revitalize indigenous languages. By offering a comprehensive translation tool, this research not only contributes to language preservation but also enhances cultural exchange and understanding. The integration of artificial intelligence in this context prompts important ethical considerations, particularly regarding the representation and distribution of cultural knowledge.

In our experiments, the contrast between the Bible-only (zero-shot) and Bible+grammar (adapted) settings quantifies this potential: adaptation yields large absolute gains on G-test performance (e.g., mBART-50 from BLEU 1.9992 to 28.8776; Marian from 0.2272 to 21.2699) and qualitatively reduces repetition while improving coverage of metalinguistic terms and everyday constructions. These substantial differences confirm that even a modest, domain-focused grammar corpus can significantly reduce domain mismatch and improve generalization to unseen constructions, especially those tied to Purépecha morphology and pedagogical explanations.

Subsequent research should prioritize expanding the linguistic corpus and exploring more sophisticated linguistic phenomena, including idiomatic expressions, metaphorical language, and cultural references. Additionally, investigating the application of cross-lingual transfer learning techniques in conjunction with Large Language Models (LLMs) could further improve the model's ability to generalize from limited data. Concretely, we recommend parameter-efficient adaptation (e.g., adapters/LoRA), back-translation and round-trip data augmentation, and orthography-aware tokenization/normalization for Purépecha to enhance subword coverage. Future evaluations should complement BLEU/ROUGE with chrF and reference-based learned metrics (e.g., COMET/BLEURT), plus human judgments focused on adequacy, fluency, and cultural appropriateness. In particular, given the significant variation observed between zero-shot and adapted systems, future work should systematically evaluate how different corpus compositions (e.g., narrative vs. didactic) influence the model's handling of Purépecha's agglutinative morphology.
Moreover, extending the qualitative analysis to include more detailed error categorization (e.g., repetition, mistranslation of metalinguistic markers, or omission of grammatical morphemes) would help identify which phenomena benefit most from adaptation.

Furthermore, the study emphasizes the critical role of community engagement in developing language technologies. Collaborating with native speakers and cultural authorities can ensure that translations are not only accurate but also culturally relevant, reducing the likelihood of misrepresentation or miscommunication. We advocate for participatory evaluation sessions (co-design of guidelines, in-context error review) and transparent data governance: clear consent, culturally appropriate licensing, and mechanisms for data removal or correction when requested by the community. Establishing reviewer panels of Purépecha speakers to audit system outputs—especially in domains involving identity, ceremony, or place names—can further reduce harm and improve trust.

Limitations and scope. Our grammar-book corpus is modest in size and didactic in style; while it exposes domain shift effectively, results may not fully reflect conversational or domain-specific registers (e.g., health, education). Automatic metrics may underrepresent morphological adequacy and discourse coherence; future releases should broaden domains and include multi-reference/human-rated subsets. Reproducibility is strengthened by fixing seeds/hyperparameters and reporting tokenization/SacreBLEU settings; nonetheless, small-data variance remains a challenge and should be addressed with multiple runs and confidence intervals.
Given the magnitude of the observed gains, additional controlled experiments isolating the contribution of each corpus type would clarify how much of the improvement arises from exposure to grammatical paradigms versus increased lexical variety.
Finally, the current work highlights the need to study the stability of results across random seeds and training runs, as this remains a known issue in low-resource settings.

8 Conclusion

The findings of this study highlight significant advancements in the automatic translation of the Purépecha language into Spanish, demonstrating the effectiveness of fine-tuning Transformer-based neural networks. This approach addresses the unique challenges posed by the complex morphology and syntactic structures of Purépecha, a language that has historically been under-resourced in terms of digital tools and linguistic data. In particular, targeted domain adaptation with a small grammar-book corpus proved decisive to bridge the gap between biblical narrative training data and didactic/colloquial usage.

The results indicate that the fine-tuning process has notably improved translation quality, suggesting that tailored models can better capture the linguistic nuances of indigenous languages. This improvement is crucial for fostering effective communication between speakers of different languages and for ensuring that the translations are culturally relevant and accurate. Quantitatively, on the grammar test set (G-test), mBART-50 improved from BLEU 1.9992 (zero-shot) to 28.877 after adaptation, while Marian rose from 0.2272 to 21.2699; ROUGE-L similarly increased (e.g., mBART-50 to 0.579), reflecting gains in adequacy and fluency. Qualitatively, adaptation reduced repetition/degenerate outputs and improved coverage of metalinguistic terminology and everyday constructions.

Moreover, the research underscores the potential of advanced machine learning techniques and Large Language Models (LLMs) in supporting linguistic diversity and cultural preservation. By providing tools for automatic translation, this work not only aids in bridging communication gaps but also contributes to the documentation and revitalization of endangered languages. This is particularly important in a global context where many indigenous languages are at risk of disappearing.

Future evaluations should complement

BLEU/ROUGE with chrF and learned metrics (e.g., COMET/BLEURT) and, crucially, incorporate speaker-informed human judgments of adequacy, fluency, and cultural appropriateness.

The study also opens up several opportunities for further research and development. Future efforts should focus on expanding and diversifying the Purépecha-Spanish corpus to encompass a broader range of linguistic phenomena, including idiomatic expressions and complex sentence structures. Additionally, enhancing the model's architecture to better handle rare linguistic features will be essential for improving its performance. Promising directions include parameter-efficient adaptation (adapters/LoRA), back-translation and round-trip augmentation, and orthography-aware tokenization/normalization to improve subword coverage for Purépecha. We also recommend multiple-run reporting with fixed seeds and confidence intervals to mitigate small-data variance and strengthen reproducibility.

Engaging with native speakers and cultural experts is vital for ensuring the cultural relevance of translations. This collaboration can help prevent misinterpretations and preserve the cultural context of the language, making the translation tools not only technically sound but also culturally sensitive. We advocate participatory workflows (co-designed evaluation guidelines, in-context error review) and transparent data governance—clear consent, appropriate licensing, and mechanisms for correction/removal upon community request. Establishing reviewer panels of Purépecha speakers to audit outputs in sensitive domains (e.g., identity, ceremony, place names) can further reduce harm and improve trust.

In summary, this research represents a meaningful step forward in the application of AI to language preservation, particularly for under-resourced languages like Purépecha. The advancements made through this study provide a foundation for ongoing efforts to document, preserve, and revitalize endangered languages, ultimately contributing to a more inclusive and linguistically diverse digital landscape. Our contribution includes a documented clean split of the grammar-book corpus (G-train/G-test) and evidence that even modest in-domain additions can yield large, practical gains under domain shift; nevertheless, the didactic nature and size of the grammar set limit generalization to conversational and specialized domains, motivating broader multi-domain corpora and community-driven evaluation in future work.

9 Acknowledgments

This work was supported by SECIHTI. We thank Jason Angel for helpful feedback during corpus preparation and experimental planning.

Use of AI tools. We used AI-assisted writing tools (e.g., **ChatGPT**, OpenAI) **exclusively** to improve clarity, grammar, and style. All scientific content, analysis, results, and conclusions are the authors' own; no AI system is listed as an author, and no AI tool was used to generate data, design experiments, or interpret results. All text produced with AI assistance was **reviewed and approved** by the authors.

References

Abrego-Mendoza, S., Angel, J., Meque, A. G. M., Maldonado-Sifuentes, C., Sidorov, G., & Gelbukh, A. (2023). *Comparison of translation models for low-resource languages*. In *Mexican International Conference on Artificial Intelligence (MICAI 2023)*.

Aycock, S., Stap, D., Wu, D., Monz, C., & Sima'an, K. (2024). *Can LLMs really learn to translate a low-resource language from one grammar book?* arXiv. https://arxiv.org/abs/2409.19151

Bible.com. (n.d.). *Bible.com*. Retrieved February 19, 2024, from https://www.bible.com/

Chamoreau, C. (2009). *Hablemos purépecha*. Universidad Intercultural Indígena de Michoacán.

Chamoreau, C. (2009). *Hablemos purépecha*. Universidad Intercultural Indígena de Michoacán.

González-Servín, C., Maldonado-Sifuentes, C. E., Sidorov, G., Kolesnikova, O., & Nuñez-Prado, C. J. (2024). Neural approaches to translating Purépecha: A comprehensive study on indigenous language preservation using Transformer networks. *Preprint*.

Hernández, P. M. (2002). *En torno a la traducción automática*. *Cervantes, 1*(2), 101–117.

Huarcaya Taquiri, D. (2020). *Traducción automática neuronal para lengua nativa peruana* (Doctoral thesis, Universidad Peruana Unión).

Hugging Face. (2024). *Hugging Face Transformers documentation*. Retrieved February 6, 2024, from https://huggingface.co/docs/transformers/index

Instituto Nacional de Estadística, Geografía e Informática. (1996). *Hablantes de lengua indígena: Perfil sociodemográfico*. INEGI.

Joshi, R., Singla, K., Kamath, A., Kalani, R., Paul, R., Vaidya, U., Chauhan, S. S., Wartikar, N., & Long, E. (2024). *Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus*. arXiv. https://arxiv.org/abs/2410.14815

Liao, Y.-C., Yu, C.-J., Lin, C.-Y., Yun, H.-F., Wang, Y.-H., Li, H.-M., & Fan, Y.-C. (2024). *Learning-from-mistakes prompting for indigenous language translation*. arXiv. https://arxiv.org/abs/2407.13343

Lin, C.-Y. (2004, July). *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics. https://www.aclweb.org/anthology/W04-1013

Mager, M., & Meza, I. (2021). *Retos en construcción de traductores automáticos para lenguas indígenas de México*. *Digital Scholarship in the Humanities, 36*(Supplement_1), i43–i48. https://doi.org/10.1093/llc/fqz093

Merx, R., Mahmudi, A., Langford, K., de Araujo, L. A., & Vylomova, E. (2024). *Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language*. arXiv. https://arxiv.org/abs/2404.04809

Nag, A., Mukherjee, A., Ganguly, N., & Chakrabarti, S. (2024). *Cost performance optimization for processing low-resource language tasks using commercial LLMs*. arXiv. https://arxiv.org/abs/2403.05434

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: A method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). https://doi.org/10.3115/1073083.1073135

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: A method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). https://doi.org/10.3115/1073083.1073135

Parra Escartín, C. (2018). *¿Cómo ha evolucionado la traducción automática en los últimos años? La Linterna del Traductor*.

Tonja, A. L., Kolesnikova, O., Arif, M., Gelbukh, A., & Sidorov, G. (2022). *Improving neural machine translation for low-resource languages using mixed training: The case of Ethiopian languages*. In *MICAI 2022* (pp. 30–40). Springer.

Tonja, A. L., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2023). *Low-resource neural machine translation improvement using source-side monolingual data*. *Applied Sciences, 13*(2), 1201.

Tonja, A. L., Maldonado-Sifuentes, C., Mendoza Castillo, D. A., Kolesnikova, O., Castro-Sánchez, N., Sidorov, G., & Gelbukh, A. (2023). *Parallel corpus for indigenous language translation: Spanish–Mazatec and Spanish–Mixtec*. arXiv. https://arxiv.org/abs/2305.17404

Tonja, A. L., Nigatu, H. H., Kolesnikova, O., Sidorov, G., Gelbukh, A., & Kalita, J. (2023). *Enhancing translation for indigenous languages: Experiments with multilingual models*. arXiv. https://arxiv.org/abs/2305.17406

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008).