

Machine Translation for Northeastern Puebla Sierra Nahuatl: Evaluating Preprocessing Scenarios and Model Convergence

Javier Tolentino-Ramírez

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla,
Mexico

tr226570041@alm.buap.mx

Abstract. This study investigates the impact of text preprocessing and orthographic representation on the neural machine translation of the Northeastern Puebla Sierra Nahuatl variant, an agglutinating language. Using the SLR92 corpus extracted from ELAN files, we evaluate three preprocessing scenarios targeting the vowel-length colon marker (":"): Keeping, Removing, and Substituting. Fine-tuning the open-source NLLB-200-3.3B model across all three scenarios reveals consistent training dynamics across three phases: an explosive early learning phase (Epochs 1–5), a high-accuracy convergence phase reaching peak validation BLEU scores of up to 26.64 (Keeping), 25.42 (Substituting), and 25.26 (Removing), and a stable performance plateau. Final evaluation on a locked test set of 3,865 sentences demonstrates solid structural consistency, with Substituting achieving the top BLEU score of 18.59 (WER 67.38%) and Keeping yielding the lowest Word Error Rate at 66.85% (BLEU 18.44).

Keywords: Neural machine translation, Nahuatl language, morphological segmentation, granularity density.

Traducción automática para el náhuatl de la Sierra Nororiental de Puebla: evaluación de escenarios de preprocesamiento y convergencia del modelo

Resumen. Este estudio investiga el impacto del preprocesamiento de texto y la representación ortográfica en la traducción automática neuronal de la variante del náhuatl de la Sierra Norte de Puebla, una lengua aglutinante. Utilizando el corpus SLR92 extraído de archivos ELAN, evaluamos tres escenarios de preprocesamiento centrados en el marcador de dos puntos (":") que indica la longitud vocálica: mantenerlo, eliminarlo y sustituirlo. El ajuste fino del modelo de código abierto NLLB-200-3.3B en los tres escenarios revela dinámicas de entrenamiento

consistentes a lo largo de tres fases: una fase inicial de aprendizaje explosivo (épocas 1–5), una fase de convergencia de alta precisión —alcanzando puntuaciones máximas de BLEU en validación de hasta 26,64 (mantener), 25,42 (sustituir) y 25,26 (eliminar)— y una meseta de rendimiento estable. La evaluación final en un conjunto de prueba fijo de 3.865 oraciones demuestra una sólida consistencia estructural; la opción de sustitución logra la puntuación BLEU más alta (18.59; WER 67.38%), mientras que la opción de mantener el marcador arroja la tasa de error por palabra (WER) más baja (66.85%; BLEU 18.44).

Palabras clave: Traducción automática neuronal, lengua náhuatl, segmentación morfológica, densidad de granularidad.

1 Introduction

The analysis of indigenous languages is a fundamental pillar for the development of machine translation systems that not only preserve these languages but also guarantee universal access to information. This is critical in areas such as the justice systems of indigenous communities, where accurate linguistic understanding is indispensable to guarantee due process and prevent unjust sentences resulting from misinterpretation.

In Mexico, 364 linguistic variants are spoken, grouped into 68 linguistic groups and 11 language families. In the case of the Nahuatl language, there are at least 14 variants [5].

The original representation of the Nahuatl language is considered a logosyllabic hieroglyphic system [3]. Its use was highly functional for recording names and administrative data; although it was not used extensively for complex sentences like the Mayan script, it possessed the capacity to do so.

Nahuatl is an agglutinating language, which is relevant to understanding its syntax and the construction of lexical meanings. In agglutinating languages, words are formed by joining multiple morphemes to a root, each of which generally maintains a distinct and recognizable meaning or grammatical function [4].

To identify the baseline characteristics of a language, like Nahuatl, it is necessary to perform a preliminary statistical analysis first. This initial step gives important insights into its vocabulary structure, word boundaries, and text density. Understanding these baseline details ensures that subsequent machine learning models can be tailored and fine-tuned with the best configurations to adapt precisely to the morphological complexities of the language.

To effectively exploit these statistical insights, state-of-the-art machine translation architectures capable of handling low-resource and morphologically rich languages are required. In this context, open-source multilingual pretrained systems provide a robust foundation for transfer learning. Among these, the No Language Left Behind (NLLB) model stands out as a powerful multilingual Transformer framework designed to scale translation across massively parallel language directions.

2 NLLB Model

No Language Left Behind (NLLB) is an open-source system designed to translate more than 200 languages based on three pillars [7]:

- Massive Data: It combines advanced web mining (using the LASER3 encoder) with FLORES-200, a high-quality, human-translated dataset used to evaluate thousands of translation directions.
- Mixture of Experts (MoE) Architecture: It utilizes a model that activates only the necessary parts of the neural network for each specific task. This allows handling hundreds of languages efficiently without the information of one interfering negatively with another.
- Training and Safety: It applies progressive learning strategies and toxicity filters to ensure that translations are not only accurate but also safe.

The NLLB-200 model includes the Northern Puebla Nahuatl variant within its training, although it does not specify whether it is the eastern or northeastern version. Retraining (or fine-tuning) the model with the corpus of the northeastern variant will allow observing the model's capabilities in translating this variant into Spanish.

The model used in the present research was NLLB-200 Dense 3.3B, which is a multilingual encoder-decoder Transformer based on the M2M-100 architecture. It comprises 24 encoder and 24 decoder layers, each with 16 attention heads, a hidden dimension of 2,048, and feed-forward networks of dimension 8,192. The model uses a 256,206-token SentencePiece vocabulary with language identifiers and supports sequences of up to 1,024 positional embeddings (trained on inputs of up to 512 tokens). Regularisation is achieved with 0.1 dropout on both hidden states and attention weights, while ReLU is used as the activation function. Unlike the NLLB-200 MoE model, its computational cost scales directly with its parameter count because every layer is executed for each inference.

3 Results in Other Translation Tasks

There are several case studies on the application of machine translation for indigenous languages of Mexico, a summary of these studies is listed in Table 1.

4 Methodology

In this study, the methodology is structured into three main stages. First, the dataset was extracted and cleaned from the raw EUDICO Linguistic Annotator (ELAN) files. Second, the corpus is analyzed using four Natural Language Processing (NLP) techniques: Tokenization, Bag of Words (BoW), Bigrams, and Granularity Density under three scenarios. Third, the scenarios are compared after a separated fine-tuning of the multilingual NLLB model for each one.

Table 1. Comparative analysis of machine translation models for Mexican indigenous languages.

Ref.	Direction	Model	Corpus Size	Configuration	BLEU
[10]	Mazatec → Spanish	mBART50	~9,000 sentences	Epochs: 3, batch	10.45
	Spanish → Mazatec			size: 4, beams: 5,	21.20
	Mixtec → Spanish		~13,000 sentences	learning rate:	15.66
	Spanish → Mixtec			5e-2	16.03
[12]	Spanish → Ayuuk	Transformer	~6,000 sentences	Layers: 6, embedding: 256, batch: 128, epochs: 250	5.56
	Ayuuk → Spanish				3.52
[2]	Spanish → Nahuatl	Transformer	~35,000 sentences	Epochs: 25 to 40, Optimizer: AdamW, Learning rate: 9.12e-04	66.45
[1]	Spanish → Mixtec	Transformer	~1,000 sentences	Layers: 6, heads: 4, embedding: 256, batch size: 128, epochs: 250	15.97
[6]	Purépecha → Spanish	Transformer	~10,000 sentences	Layers: 64–256, epochs: 50–100	15.85

4.1 Data Extraction and Cleaning from ELAN Files

The primary linguistic data consists of annotated ELAN files under the identifier SLR92 [9]. An automated extraction pipeline using `pympi` was developed to align Nahuatl tiers with their corresponding Spanish translations based on temporal overlap.

To ensure data purity, a multi-stage regular expression cleaning process was implemented:

- Text encapsulated within metadata tags (`<sci>`, `<i>`, `<b>`) was identified as non-linguistic noise and removed entirely.
- All characters were normalized using the Unicode Normalization Form C (NFC) to ensure consistent encoding of Nahuatl-specific diacritics and the Spanish *ñ*.

4.2 Corpus Analysis and Metrics

After the pre-processing and text normalization pipeline, the cleaned corpus of the Northeastern Puebla Sierra Nahuatl variant was analyzed. To identify the baseline characteristics of Nahuatl writing, the following techniques were applied:

- Tokenization: This process consists of dividing the content into basic components called tokens (such as words, numbers, and punctuation marks) based on their structure. For this study the colon symbol (":") was explicitly preserved for subsequent comparative evaluation.
- Bag of Words (BoW): This is a classical and compact text representation that aggregates the set of terms appearing within a corpus. Despite being a significant simplification, this method is not free from the complexities of natural language due to the ambiguity of terms out of context, which manifests through synonymy and polysemy.
- Bigrams: The *One Sense per Collocation* principle establishes that the meaning of a word remains constant each time it appears within the same collocation or bigram. This property is fundamental for Word Sense Disambiguation (WSD) tools, which utilize classifiers based on decision lists that implicitly assume this rule.
- Granularity Density: Granularity density was analyzed to understand how character-level markers change the vocabulary structure. Specifically, this analysis measures how the text density, unique vocabulary size, and lexical ambiguity change across different scenarios.

In Nahuatl, a colon (:) typically denotes vowel length, an essential feature that can completely alter a word's meaning. To determine how representing this feature impacts machine translation, it was conducted a statistical comparison across three dataset scenarios: keeping the original colons, removing them entirely, and substituting them with whitespace.

To guarantee the statistical validity of the evaluation, the master dataset $\mathcal{D}$ was partitioned into three disjoint sets using a seeded pseudo-random distribution:

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}}, \quad (1)$$

where:

- $\mathcal{D}_{\text{train}}$ (70%): Used for weight optimization via backpropagation.
- $\mathcal{D}_{\text{val}}$ (15%): Used for evaluating the accuracy of translation during training.
- $\mathcal{D}_{\text{test}}$ (15%): A "locked" set used exclusively for the final reporting of evaluation metrics, ensuring no data leakage during the training stage.

The Bilingual Evaluation Understudy (BLEU) metric compares tokenized n -grams of generated text against reference human translations. The final output is scaled to a percentage between 0 and 100, where 0 indicates low quality or no overlap with the reference, and close to 100 represents an almost perfect match with the human translation. This value is obtained multiplying a Brevity Penalty (BP) by the geometric mean of n -gram precisions [8]:

$$\text{BLEU} = 100 \cdot \text{BP} \cdot \exp \left(\sum_{n=1}^4 0.25 \cdot \log p_n \right). \quad (2)$$

The Brevity Penalty (BP) penalizes translations that are shorter than the reference texts to prevent high precision through truncation:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{(1-r/c)} & \text{if } c \leq r, \end{cases} \quad (3)$$

where:

- $c = \sum_{i=1}^M |c_i|$: Total length of generated translations across M sentences after standardized tokenization.
- $r = \sum_{i=1}^M r_i$: Effective reference length across M sentences, where r_i is the reference length closest to candidate length $|c_i|$.

The modified n -gram precision p_n calculates the proportion of matched n -grams across the entire dataset. A clip is applied to avoid overcounting:

$$p_n = \frac{\sum_{i=1}^M \sum_{g \in C_i} \text{count_clip}(g)}{\sum_{i=1}^M \sum_{g \in C_i} \text{count}(g)}, \quad (4)$$

where:

- $\text{count_clip}(g)$: Count of n -gram g in candidate sentence C_i , capped at the maximum frequency of g in any single matching reference sentence.
- $N = 4$: Evaluates unigrams up to 4-grams ($n \in \{1, 2, 3, 4\}$).
- $w_n = 0.25$: Uniform weight assigned to each n -gram order ($w_1 = w_2 = w_3 = w_4 = \frac{1}{4}$).
- Smoothing: Exponential decay smoothing is applied when $p_n = 0$ to avoid taking $\log(0)$.

Word Error Rate (WER) is also a widely used measure for evaluating the accuracy of speech recognition and machine translation systems. Its values generally range from 0 to 1, where a score of 0 represents a perfect match between the reference and the predicted text, while a score of 1 (or higher) indicates that the two texts are entirely dissimilar, sharing little or no overlap, it is calculated as:

$$\text{WER} = \frac{S + D + I}{N}, \quad (5)$$

where, S , D , and I represent substitutions, deletions, and insertions, while N is the total number of words in the reference.

4.3 Model Configuration

The model was configured using the default parameters of NLLB-200-3.3B except for the number of epochs, the evaluation strategy, generation parameters, memory optimization of the available hardware, Learning Rate (LR) and Weight Decay (WD).

The parameters LR and WD were adjusted following the recommendations for adafactor optimizer[11], a summary of model configuration is shown in Table 2.

Table 2. Training hyperparameters and configurations for Seq2SeqTrainingArguments.

Hyperparameter	Value	Technical Description
Epoch Control		
num_train_epochs	25	Total number of complete passes over the training corpus.
eval_strategy	"epoch"	Performs model evaluation at the end of each epoch.
Generation Parameters (Metrics)		
predict_with_generate	True	Forces autoregressive decoding to evaluate with BLEU/WER.
Memory Optimization (A100 GPU)		
bf16	True	Enables 16-bit brain floating point precision.
optim	"adafactor"	Optimizer designed for massive models with lower memory footprint.
dataloader_num_workers	4	Parallel threads for data loading and pre-processing.
Anti-Overfitting Strategies		
learning_rate	1e-5 (0.00001)	Initial learning rate for the optimizer.
weight_decay	1e-3 (0.001)	L2 regularization penalty applied to weights.

5 Results

Bag of Words technique was conducted to get the 20 top words for each scenario and 20 top bigrams as well. It was observed that the "Keeping" and "Removing" scenarios share the exact same word boundaries and token distributions because stripping out the vowel-length marker (":") is a simple character-level string replacement that reduces orthographic complexity. In the "Keeping" scenario, the vocabulary is dominated by words like *wa:n* and *yo:n* peaking at around 14,500 occurrences, followed closely by *se:* and the letter *n*, while the top bigram pairs are *yehwa yo:n* (~1,800 occurrences), *wa:n yo:n*, and *a:mo a:mo*.

The "Removing" scenario preserves this entire distribution shape and order, simply normalizing the tokens to *wan* and *yon* without changing any absolute frequency counts. On the other hand, the "Substituting" scenario introduces a radical change in tokenization by implementing morphological segmentation. This aggressive mechanism causes the single-letter token *n* to explode to over 72,000 occurrences, becoming a massive statistical outlier because it acts as a highly frequent relational particle, determiner, or clitic detached from its host words, while other fragments like *wa*, *a*, *yo*, and *ke* take over the top spots. Consequently, the bigram distribution directly reflects this split, causing the

original single words to reappear as the top bigrams *wa n* ($\sim 15,000$) and *yo n* ($\sim 14,000$) matching their original unigram frequencies.

From Figure 1 it is observed that the transition from "Keep" to "Remove" represents an orthographic normalization layer where stripping the vowel-length marker (":") collapses variants like *wa:n* and *wan* into a single token, successfully reducing unique vocabulary sparsity while keeping the total token count exactly identical at 347,109.

Conversely, moving from "Remove" to "Substitute" introduces aggressive morphological segmentation by breaking compound words or attached particles into independent units, this structural change causes the unique vocabulary to plummet to its lowest point (15,679) while the total token count explodes to over half a million, ultimately reducing the dictionary size the model needs to learn at the cost of significantly increasing the overall sentence sequence lengths across the corpus.

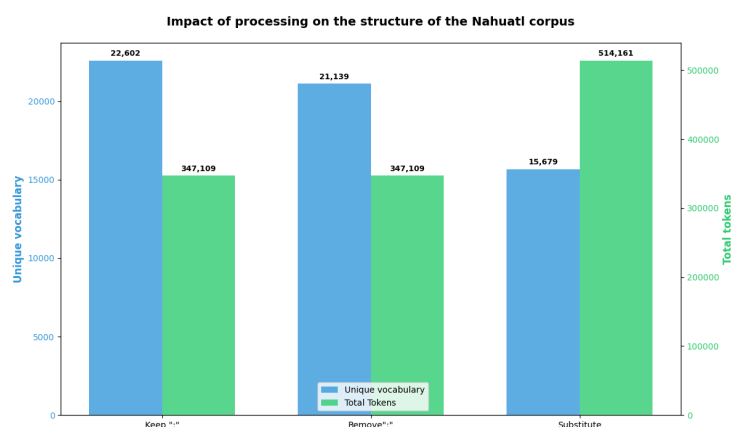


Fig. 1. Impact of each scenery in total number of tokens.

The information given in Figure 2 confirms the invariance of removing colon for context analysis, both scenarios lie at around 5 to 7 tokens per sentence. Conversely, the "Substituting" scenario is significantly flattened and shifted to the right, where the peak density broadens to around 8 to 12 tokens per sentence while the right tail stretches well past 30 to 40 tokens as a direct structural consequence of the morphological splitting. For modern NLP frameworks like Transformers, this Substituting scenario significantly increases the maximum and average sequence lengths and lowers the information density per token compared to the other scenarios.

In order to evaluate the performance impact of these structural differences, all three preprocessing configurations (Keeping, Removing, and Substituting) were implemented and evaluated during NLLB model training. Applying the NLLB

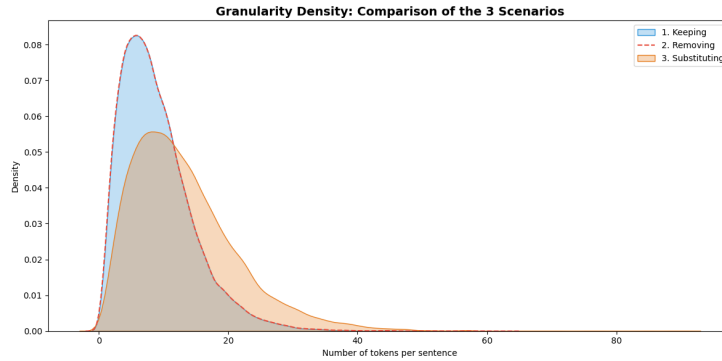


Fig. 2. Granularity density of tokens per sentence for each scenery.

model across all three scenarios yields the BLEU and WER evaluation scores shown in Figure 3 and Figure 4, respectively.

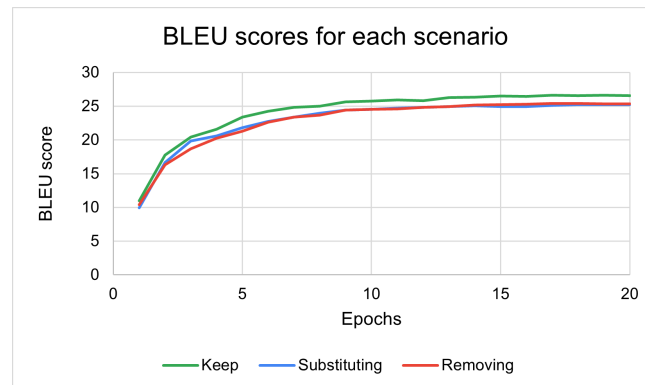


Fig. 3. BLEU scores for each epoch in validation set.

By analyzing the behavior of BLEU scores across epochs for all three scenarios, three distinct phases are clearly identified:

- Explosive Learning Phase (Epochs 1 to 5): Models across all three scenarios rapidly assimilate foundational linguistic structures. BLEU scores increase sharply from initial values of 10.99 (Keeping), 9.96 (Removing), and 10.37 (Substituting) at Epoch 1 to 23.37, 21.83, and 21.30, respectively, by Epoch 5.
- Convergence Phase (Epochs 6 to 15): Performance stabilizes into a high-accuracy region. The "Keeping" scenario converge to BLEU scores near to 26 at Epoch 15, while the "Removing" and "Substituting" scenarios converge to BLEU scores near to 25.

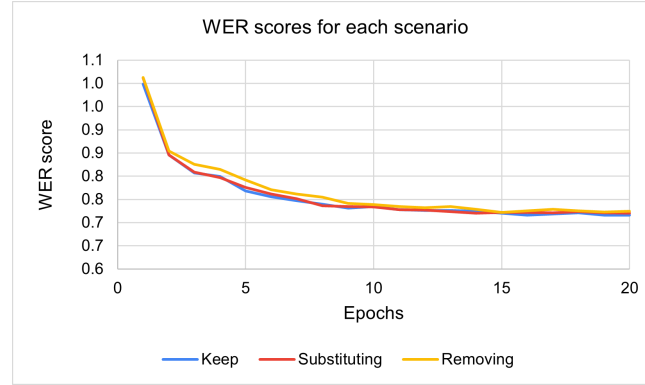


Fig. 4. WER scores for each epoch in validation set.

- Plateau Phase (Epochs 16 to 20): Rather than experiencing severe performance degradation, BLEU scores plateau across all setups, maintaining stable scores above 25.2 for Removing and Substituting, and above 26.4 for "Keeping", demonstrating consistent generalization without sharp overfitting through 20 epochs.

The WER metric measures the word error rate by accounting for substitutions, deletions, and insertions required to align predicted sequences with the reference text. Across all three scenarios, initial error rates at Epoch 1 are high due to complete undertraining—starting at 99.82% (0.9982) for "Keeping", 101.20% (1.0120) for "Removing", and 101.30% (1.0130) for "Substituting". Following this, a rapid downward trajectory emerges as WER drops significantly by Epoch 2 (84.58% for Keeping, 84.62% for Removing, and 85.43% for Substituting).

Error rates continue to decline in parallel with BLEU growth before reaching their respective minimums: 71.60% at Epoch 20 for Keeping, 72.01% at Epoch 20 for "Removing", and 72.24% at Epoch 15 for Substituting, after which scores stabilize within tight plateaus.

The final evaluation on the locked test set, comprising 3,865 sentences, demonstrated competitive performance across all three tokenization strategies. The "Substituting" scenario achieved the highest test BLEU score of 18.59 (with a WER of 67.38%), closely followed by "Removing" (BLEU 18.49, WER 66.89%) and "Keeping" (BLEU 18.44, WER 66.85%). Compared to validation training curves—where validation BLEU peaked between 25.26 and 26.64 and WER stabilized near 71.60%–72.24%—the test evaluation reveals a slight generalization gap in translation precision (BLEU) on completely unseen data, alongside improved word-level accuracy (WER drops by approximately 5 percentage points across all scenarios on the test set).

During training, "Keeping" reached the highest validation BLEU score (26.64 vs. 25.42 for "Substituting"), this shows that holding whole words works well when the validation set shares the exact same words as the training data. However,

on the final unseen test set, "Substituting" achieved a slightly better BLEU score, this confirms that breaking words into smaller pieces helps the model handle new, unknown words (out-of-vocabulary) much better when training data is limited.

Even though breaking up words made sentences much longer—increasing total tokens by over 44% and doubling average sentence lengths in Substituting—the final test scores across all three setups were very similar (BLEU 18.44–18.59; WER 66.85%–67.38%). This shows that the large pretrained NLLB model is strong enough to handle big changes in text formatting without losing translation quality, and that simple text cleanup mitigates sparsity without making sentences too long or complicated for the fine tuning process.

6 Conclusions

In this study, we investigated the impact of character-level text preprocessing and orthographic representation on the neural machine translation of the Northeastern Puebla Sierra Nahuatl variant, a low-resource agglutinating language. By conducting a comprehensive statistical evaluation across three distinct tokenization setups (Keeping, Removing, and Substituting), we demonstrated how character-level markers directly alter vocabulary granularity and corpus density. While the "Keeping" and "Removing" configurations preserved identical word boundaries, the "Substituting" scenario implemented an aggressive morphological segmentation. This structural alteration collapsed the unique vocabulary size while causing the total token count to increase, thereby altering sequence lengths and context distribution during transformer training.

Fine-tuning the open-source NLLB-200 model across all three configurations revealed consistent training dynamics characterized by rapid early learning (Epochs 1–5) followed by convergence in later epochs (Epochs 15–20), where validation BLEU scores reached 26.64 for Keeping, 25.42 for Substituting, and 25.26 for Removing.

The final evaluation on the locked test set of 3,865 sentences demonstrated overall structural consistency across preprocessing methods. The "Substituting" approach achieved the top BLEU score of 18.59 on test data, whereas the "Keeping" configuration yielded the lowest Word Error Rate at 66.85%. Although a modest generalization gap manifests in BLEU when comparing test results to validation peaks, word-level error rates showed improved alignment on test data. Ultimately, these findings underscore that comparative orthographic evaluation combined with advanced state-of-the-art multilingual architectures enhances translation performance in endangered agglutinating languages.

For future work, it is highly recommended to explore more specialized text processing techniques designed for translation tasks of agglutinated languages, like Nahuatl, which could help manage complex word formations more effectively. Additionally, future studies should revise other promising transformer topologies and architectures to see how they perform on indigenous translation tasks. When comparing different frameworks, it is crucial to check carefully if other models separate a complete, isolated test set after training to completely avoid

data leakage and ensure a fair evaluation. Finally, an extensive fine-tuning of hyperparameters (such as experimenting with different learning rates, batch choices, and optimization steps) is strongly recommended to find the absolute best configuration for translating these indigenous variants.

Acknowledgment. The author thankfully acknowledge the computer resources, technical expertise and support provided by the National Supercomputing Laboratory of Southeastern Mexico and the Ministry of Science, Humanities, Technology and Innovation.

References

1. Bautista Morales, R., Martínez Ramírez, Y., Rocha Peña, L.E., Montes Santiago, R.E.: Arquitectura de un traductor automático para el idioma mixteco: un enfoque específico para lenguas indígenas con escasos recursos lingüísticos. *Revista de Investigación en Tecnologías de la Información* 12(28), 71–81 (Nov 2024)
2. Bello García, S.K., Sánchez Lucero, E., Bonilla Huerta, E., Hernández Hernández, J.C., Ramírez Cruz, J.F., Pedroza Méndez, B.E.: Nahuatl neural machine translation using attention based architectures: A comparative analysis for RNNs and transformers as a mobile application service. In: *Advances in Soft Computing*, pp. 120–139. *Lecture Notes in Computer Science*, Springer International Publishing, Cham (2021)
3. Cossich Vielman, M.V.: El sistema de escritura jeroglífico náhuatl: Análisis epigráfico de los onomásticos de cinco documentos del siglo XVI de Tepetlaóztoc. Tesis de maestría, Universidad Nacional Autónoma de México (UNAM), Ciudad de México, México (2014)
4. Davletshin, A.: Descripción funcional de la escritura jeroglífica náhuatl. *Estudios de Cultura Náhuatl* 62, 43–93 (2021)
5. Embriz Osorio, A., Zamora Alarcón, Ó. (eds.): *Lenguas Indígenas Nacionales en Riesgo de Desaparición*. Instituto Nacional de Lenguas Indígenas (INALI), México, D.F. (2012)
6. González Servín, C., Maldonado Sifuentes, C.E., Kolesnicova, O., Sidorov, G.: Evaluating the impact of domain adaptation on transformer-based models for low-resource Purépecha-Spanish translation. *Int. J. Comb. Optim. Probl. Inform.* 17(2), 27–37 (Feb 2026)
7. NLLB Team, Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G.M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K.R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N.F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., Wang, J.: No language left behind: Scaling human-centered machine translation (Jul 2022)
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Association for Computational Linguistics, Morristown, NJ, USA (2001)
9. Shi, J., Amith, J.D., Chang, X., Dalmia, S., Yan, B., Watanabe, S.: Highland puebla nahuatl speech translation corpus for endangered language documentation. In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. pp. 53–63 (2021)

10. Tonja, A.L., Maldonado-sifuentes, C., Mendoza Castillo, D.A., Kolesnikova, O., Castro-Sánchez, N., Sidorov, G., Gelbukh, A.: Parallel corpus for indigenous language translation: Spanish-mazatec and Spanish-Mixtec. In: Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP). Association for Computational Linguistics, Stroudsburg, PA, USA (2023)
11. Wang, Y., Xu, C., Sun, Q., Hu, H., Tao, C., Geng, X., Jiang, D.: PromDA: Prompt-based data augmentation for low-resource NLU tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Stroudsburg, PA, USA (2022)
12. Zacarías Márquez, D., Meza Ruiz, I.V.: Ayuuk-Spanish neural machine translator. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas. Association for Computational Linguistics, Stroudsburg, PA, USA (2021)