Fecha de recepción: 06 de marzo de 2026, fecha de publicación en línea: junio de 2026.

Implementación de Modelo de Lenguaje Multilingüe para Traducción Automática Bidireccional entre Español y Náhuatl

Samuel Pérez Zistecatl, José Crispín Hernández Hernández, Eduardo Sánchez Lucero, Yesenia Nohemí González Meneses , y Edmundo Bonilla Huerta
Tecnológico Nacional de México – Instituto Tecnológico de Apizaco. San Andrés Ahuashuatepec, Municipio de Tzompantepec, Tlaxcala, C.P. 90491, México.

Autor de correspondencia: Samuel Pérez Zistecatl (correo electrónico: m24370014@apizaco.tecnm.mx).

Abstract- In this work, we present the design, construction, and evaluation of a large-scale language model (LLM) conceived for bidirectional translation between Spanish and Nahuatl, developed with technical rigor and with careful respect for the cultural weight inherent to the Indigenous language. A bilingual corpus comprising 15,000 sentence pairs were compiled; SentencePiece tokenization was adopted, and fine-tuning was carried out on Transformer-based architectures (refinement using mBART50-Sequential Encoder and LoRA adaptation). A functional web prototype was deployed as proof of concept. Automated metrics (BLEU, chrF++/ROUGE), along with expert evaluation conducted by native speakers and linguists, demonstrate the feasibility of the system and its capacity to preserve linguistic features of Nahuatl that are often lost in generic translation systems. This endeavor therefore constitutes both a technical and sociocultural contribution aimed at fostering digital inclusion and supporting the preservation of Indigenous languages.

Keywords: cultural preservation, digital inclusion, large language model (LLM), linguistic revitalization, natural language processing (NLP), Spanish–Nahuatl machine translation, Transformers.

I. INTRODUCCIÓN

La pluralidad lingüística de la república mexicana es un patrimonio vasto y venerable, hoy más amenazada por la merma de hablantes jóvenes y por la escasa incorporación de las lenguas indígenas en los ámbitos tecnológicos contemporáneos [1]. El náhuatl en México, padece un retroceso en su transmisión intergeneracional. En territorios como Tlaxcala, la hegemonía del español en la escuela, la administración y los medios ha coadyuvado al desplazamiento del uso cotidiano del náhuatl, restringiendo los recursos digitales disponibles y agrandando la brecha tecnológico–cultural frente a las mayorías [2]. Un LLM (Large Language Model) es un modelo de inteligencia artificial basado en redes neuronales profundas, generalmente del tipo *transformer*, entrenado con grandes volúmenes de texto para aprender patrones del lenguaje natural El desarrollo de herramientas de procesamiento del lenguaje natural (PLN) se erige como estrategia pertinente para la preservación, normalización y revitalización del náhuatl en ámbitos digitales [3]. Este artículo presenta la metodología, implementación y validación de un modelo de lenguaje de gran tamaño para la traducción bidireccional español-náhuatl como contribución para el aprendizaje y conservación de las lenguas indígenas. El modelo fue desarrollado con precisión técnica y sensibilidad cultural como muestra de respeto a la lengua.

II. MARCO TEÓRICO Y TRABAJOS RELACIONADOS

2.1. Arquitecturas Basadas en Transformers y Mecanismos de Atención

La base de la traducción automática moderna es la arquitectura Transformer [13]. A diferencia de sus predecesores recurrentes (RNN- Recurrent Neural Network y LSTM-Long Short Term Memory), el Transformer elimina la secuencialidad obligatoria, permitiendo el procesamiento paralelo mediante el Mecanismo de Auto-atención de Múltiples Cabezales (Multi-Head Self-Attention).

Este mecanismo permite que, para cada unidad lingüística o token (unidad de texto), el modelo calcule un puntaje de relevancia respecto a todos los demás elementos de la secuencia, independientemente de su distancia lineal. En el caso del náhuatl, esta propiedad es fundamental: permite conectar prefijos de sujeto situados al inicio de una palabra

aglutinada con raíces verbales o sufijos de tiempo que aparecen al final, resolviendo dependencias morfológicas complejas que en modelos analíticos se perderían.

2.2. Modelos de Lenguaje Multilingües

Los modelos como NLLB (No Language Left Behind) y mBART50 (Sequential Encoder) representan una evolución del Transformer estándar hacia la traducción multilingüe masiva. Estos modelos son pre-entrenados mediante objetivos de eliminación de ruido (denoising) [9], donde el modelo debe reconstruir fragmentos de texto alterados.

El modelo proyecta las palabras a vectores de alta dimensión (embeddings). Al ser multilingüe, el modelo aprende que el concepto "agua" en español y "atl" en náhuatl deben ocupar regiones cercanas en el espacio vectorial, facilitando la transferencia de conocimiento entre idiomas con abundantes recursos y aquellos de bajo recurso [6], [8].

Otro aspecto importante en los modelos de lenguaje multilingüe es el sesgo lingüístico digital (SLD). Es imperativo considerar que los LLM masivos suelen heredar sesgos de los datos de entrenamiento. Muñoz-Basols et al. (2024) [12] advierten que, sin un ajuste fino adecuado, el modelo puede "españolizar" la sintaxis de lenguas indígenas, perdiendo la esencia polisintética del náhuatl.

2.3. Lingüística Computacional del Náhuatl y Desafíos Técnicos

El náhuatl presenta una complejidad morfológica superior a las lenguas indoeuropeas debido a su naturaleza aglutinante. Una palabra puede contener información de sujeto, objeto, dirección y tiempo.

Teniendo presente lo anterior, es recommendable utilizar la tokenización por sub-palabras. Para manejar esta complejidad, se emplea SentencePiece (Sequence to Sequence Tokenizer) que divide el texto en unidades sub-léxicas. Esto previene el problema del "vocabulario fuera de diccionario" (OOV), permitiendo que el modelo entienda palabras que nunca vio en el entrenamiento, siempre que reconozca sus componentes morfológicos.

Otro de los retos es la asimetría de datos, la cual consiste en establecer una categorización em los datos obtenidos del tokenizador. La categorización del náhuatl como lengua de bajo recurso (low resource) no es cultural, sino estadística. La disparidad entre el volumen de datos digitales para el español y el náhuatl exige técnicas de optimización específicas, como el ajuste de hiperparámetros sugerido por Araabi y Monz (2020) [10].

2.4. Fundamentación de las Métricas de Evaluación

Para validar la efectividad del modelo, se emplean métricas que cuantifican la fidelidad de la traducción [7]:

BLEU (Bilingual Evaluation Understudy): Evalúa la precisión basada en la coincidencia de n-gramas. Un puntaje BLEU alto valida la coherencia sintáctica global.

chrF++(Character n-gram F-score): Esta métrica es especialmente relevante para esta investigación, ya que mide n-gramas a nivel de caracteres. Al ser el náhuatl una lengua que construye palabras "pegando" piezas, chrF++ captura si el modelo está ensamblando correctamente los morfemas, incluso si hay pequeñas variaciones ortográficas.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Se utiliza para medir la recuperación (recall), asegurando que el modelo no omita conceptos clave presentes en la oración fuente.

III. IMPLEMENTACION DE UN LLM

La materialización de un LLM no es tarea facil, si bien varias arquitecturas contemporáneas admiten entradas de larga duración, sobreviene con frecuencia el fenómeno denominado “loss-in-the-middle” —la pérdida de información situada en la parte mediana del contexto—. En su percepción [4] menciona que tal déficit proviene de una supervisión explícita insuficiente durante el entrenamiento de contextos extensos, ausencia que impide reconocer cualquier posición de los mismos.

Para este trabajo, se construyó un corpus bilingüe de 7 500 pares de oraciones, se entrenó un tokenizador SentencePiece y se ajustaron con precisión modelos basados en Transformer — particularmente mBART50 con adaptación LoRA—; además, se liberó un prototipo web operativo como prueba de concepto, demostrando así la factibilidad del modelo. Se realizaron evaluaciones automáticas (BLEU, chrF++ y ROUGE) y retroalimentación humana de hablantes nativos junto con especialistas, ambos validaron la viabilidad del sistema y su capacidad para conservar matices del náhuatl que suelen perderse en traductores genéricos; por tanto, el modelo representa una mejora significativa respecto a los sistemas de traducción existentes. En esencia, el modelo constituye una innovación técnica con una misión sociocultural centrada en la accesibilidad digital y la conservación de lenguas indígenas, y por ende posee el potencial de incidir positivamente en la preservación y promoción del náhuatl

IV. METODOLOGIA

El desarrollo del sistema de traducción automática bidireccional se fundamenta en un pipeline de ingeniería de lenguaje [6], [10]. La arquitectura seleccionada es el modelo NLLB-1.3B (No Language Left Behind), cuya capacidad de representación multilingüe masiva ofrece una base sólida para el aprendizaje por transferencia (transfer learning) en lenguas indígenas americanas [11]. Se desarrollaron las siguientes actividades:

4.1 Ingeniería de Datos y Curación del Corpus

La calidad del entrenamiento supervisado depende críticamente de la integridad del corpus paralelo, por ello se consolidó un conjunto de datos mixto de 15,000 pares de sentencias

bilingües, siguiendo un proceso de tres etapas para asegurar la documentación adecuada de la lengua [2]:

- Recolección Multifuente: Se integraron datos provenientes de repositorios educativos del INALI [14], contenidos pedagógicos de la SEP [15] e información de recursos bilingües de JW.org [16]. Esta diversidad de fuentes busca reducir el sesgo lingüístico digital (SLD) que suelen presentar los modelos de gran escala [12].
- Normalización Ortográfica: Dada la ausencia de una norma única para el náhuatl, se unificaron las variantes gráficas hacia la convención del náhuatl del centro. Se aplicaron técnicas de limpieza para eliminar ruidos ortográficos y caracteres especiales que incrementan la entropía del vocabulario y dificultan la convergencia del modelo [1].
- Tokenización Basada en Sub-palabras: Se implementó el algoritmo SentencePiece para descomponer las estructuras polisintéticas del náhuatl en unidades sub-léxicas recurrentes [13]. Este enfoque previene el problema de palabras fuera de vocabulario (out-of-vocabulary) y permite que el modelo capture la compleja morfología aglutinante de la lengua [9], [10].

4.2 Arquitectura del Sistema y Mecanismos de Atención

Como motor de traducción se empleó la arquitectura Transformer [13], la cual utiliza un codificador para procesar la secuencia fuente y un decodificador para generar la secuencia destino mediante mecanismos de auto-atención (self-attention), los cuales se mencionan a continuación:

- Atención Cruzada (Cross-Attention): Se aplicó este mecanismo ya permite que el decodificador consulte las representaciones contextuales del codificador para mantener la concordancia entre los prefijos de sujeto en náhuatl y los pronombres en español, factor determinante en lenguas con alta densidad morfológica [4], [8].
- Eficiencia en Bajos Recursos: Se aplicaron técnicas de optimización y ajuste de hiperparámetros para adaptar el modelo a las restricciones de memoria de una infraestructura académica, asegurando la estabilidad del gradiente durante el proceso de retropropagación (backpropagation) [10].

4.3 Protocolo de Entrenamiento: Se aplicó aprendizaje por currículo (Curriculum Learning)

Para maximizar la convergencia ante la escasez de datos, el entrenamiento se estructuró en tres fases progresivas [4]:

- Fase 1 (Fine-Tuning Unidireccional): Se realizó un entrenamiento inicial enfocado en la dirección Náhuatl↔Español con una tasa de aprendizaje de $3x10^{-5}$ para estabilizar la comprensión semántica del modelo sobre la morfología indígena.
- Fase 2 (Generación Sintética y Back-translation): Se utilizó el modelo intermedio para traducir frases monolingües de español hacia un náhuatl sintético, ya que estos pares adicionales actúan como un mecanismo de regularización que mejora la fluidez del decodificador y la capacidad de traducción del modelo [6], [9].
- Fase 3 (Optimización Bidireccional): Se ejecutó un entrenamiento simultáneo en ambas direcciones (ESP↔NAH) utilizando el optimizador AdamW y un esquema de incremento progresivo (warmup) para refinar los pesos finales del modelo [13].

4.4 Marco de Evaluación Híbrido

La validación de los resultados se realizó bajo un esquema dual que garantiza la precisión técnica y la fidelidad lingüística [7]:

- Evaluación Automática: Se emplearon las métricas BLEU y chrF++ para cuantificar la similitud léxica, y ROUGE para la recuperación de información frente a referencias humanas [5], [7].
- Evaluación Cualitativa: Se contempló un análisis de interacción directa para examinar la adecuación contextual y la naturalidad de la traducción, considerando que el rendimiento de los modelos actuales aún presenta discrepancias entre las métricas automáticas y los juicios expertos [5], [11]

V. RESULTADOS

5.1 Evaluación Cuantitativa y Convergencia del Modelo

El desempeño se evaluó tras completar las 10 épocas totales de entrenamiento definidas en el esquema de currículo. Los resultados, segmentados por métricas de precisión, recuperación y estructura de caracteres, demuestran la eficacia de la arquitectura NLLB-1.3B adaptada.

TABLA I. Resultados de Evaluación Automática del Modelo Final.

Dirección de Traducción	BLEU	chrF++	ROUGE-L
Náhuatl ↔Español	31.25	58.40	0.54
Español ↔Náhuatl	26.15	49.12	0.42
Promedio General	**28.70**	**53.76**	**0.48**

Se observa que la métrica chrF++ es significativamente más alta que el BLEU. Esto indica que, aunque el modelo a veces no acierta la palabra exacta (especialmente en náhuatl), sí es capaz de generar los componentes morfológicos correctos, lo cual es un indicador de éxito para una lengua aglutinante [7].

5.2 Discusión sobre la Polisíntesis y el Error de Atención

Tras un análisis cualitativo de las muestras, se identificaron patrones de error específicos:

Incorporación Sustantiva: El modelo identifica correctamente la raíz verbal, pero en ocasiones falla al integrar el objeto directo dentro del verbo, optando por una estructura más analítica similar al español. Esto confirma el sesgo lingüístico digital (SLD) advertido por Muñoz-Basols et al. (2024) [12].

- Ambigüedad Dialectal: En frases con alta varianza regional, el modelo tiende a la normalización, eligiendo la forma gramatical más frecuente en el corpus del INALI o la SEP [14], [15].
- Mecanismo de Atención Cruzada: Se verificó que los cabezales de atención logran alinear correctamente los prefijos de sujeto (ej. ni-, ti-) con los pronombres correspondientes en español, lo que valida la robustez de la arquitectura Transformer en este dominio [13].

5.3 Discusión sobre la Polisíntesis y el Error de Atención

Tras un análisis cualitativo de las muestras, se identificaron patrones de error específicos:
Incorporación Sustantiva: El modelo identifica correctamente la raíz verbal, pero en ocasiones falla al integrar el objeto directo dentro del verbo, optando por una estructura más analítica similar al español. Esto confirma el sesgo lingüístico digital (SLD) advertido por Muñoz-Basols et al. (2024) [12].

- Ambigüedad Dialectal: En frases con alta varianza regional, el modelo tiende a la normalización, eligiendo la forma gramatical más frecuente en el corpus del INALI o la SEP [14], [15].

5.4 Comparativa con el Estado del Arte

Comparado con modelos base como mBART50 (sin ajuste fino específico para náhuatl central) o GPT-4 (en tareas zero-shot), la propuesta actual muestra una ventaja competitiva en la preservación de la estructura gramatical indígena [5], [11]. Mientras que los modelos comerciales suelen alucinar términos o usar estructuras simplificadas, el modelo ajustado respeta la densidad morfológica requerida para una interpretación lingüística confiable.

VI. CONCLUSIONES

A partir de los resultados experimentales, se infiere que los modelos de lenguaje de gran tamaño (LLM), cuando se ajustan finamente con corpus bilingües seleccionados y metodologías de *fine-tuning* —particularmente sobre la arquitectura NLLB generan traducciones para el náhuatl más precisas y culturalmente sensibles.

- Estrategia Viable: La sinergia de un procesamiento de datos riguroso (limpieza y aumento) junto con la validación humana constituye una estrategia efectiva para la accesibilidad digital.
- Escalabilidad: Esta metodología es escalable y puede emularse para otras lenguas indígenas de México, posicionando a la inteligencia artificial como una herramienta para el renacimiento cultural.

Refinamiento: Se concluye que la cantidad de datos sin refinamiento genera ruido, mientras que el refinamiento metodológico amplifica el valor del sistema.

REFERENCIAS

[1] G. A. Gomashie y R. Terborg, "Nahuatl, selected vitality indicators and scales of vitality in an Indigenous language community in Mexico," Open Linguistics, vol. 7, n.º 1, pp. 166–180, 2021.

[2] J. Shi, J. D. Amith, X. Chang, S. Dalmia, B. Yan, S. Watanabe, "Highland Puebla Nahuatl-Spanish Speech Translation Corpus for Endangered Language Documentation", capítulo de conferencia/libro, 2021–2022.

[3] C. A. Aguilar Santiago, H. A. García Zúñiga, "Language technologies applied to processing indigenous languages in Mexico: an overview", Lingüística y Literatura, n.º 84, 2022.

[4] S. An, Z. Ma, Z. Lin, N. Zheng, J. G. Lou, W. Chen, "Let your LLM make the most of the context," Advances in Neural Information Processing Systems, vol. 37, pp. 62160–62188, 2024.

[5] A. Hendy et al., "What is the performance of GPT models on machine translation? An exhaustive assessment," arXiv:2302.09210, 2023.

[6] B. Zhang, P. Williams, I. Titov, R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," arXiv:2004.11867, 2020.

[7] M. Freitag et al., "Experts, errors, and context: A large-scale study of human evaluation for machine translation," Trans. Assoc. Comput. Linguist., vol. 9, pp.1460–1474, 2021.

[8] A. Chowdhery et al., "Scaling language modeling with pathways (PaLM)," J. Mach. Learn. Res., vol. 24, no. 240.

[9] Y. Tang et al., "Multilingual translation from denoising pre-training", Proceedings of Findings ACL-IJCNLP, August 2021, pp. 3450–3466.

[10] A. Araabi, C. Monz, "Optimizing Transformer for low-resource neural machine translation", arXiv:2011.02266, 2020.

[11] M. Lucas et al., "The Performance of Artificial Intelligence in the Use of Indigenous American Languages", Inter-American Development Bank, May 2025.

[12] J. Muñoz-Basols, M. del M. Palomares, F. Moreno Fernández, "Digital linguistic bias (SLD) in artificial intelligence: consequences for large-scale language models in Spanish," Language and Society, vol. 23, n.º 2, pp. 623–648, 2024.

[13] A. Vaswani et al., "Attention Is All You Need," NeurIPS.

[14] INALI, Recursos lingüísticos y educativos en lenguas nacionales indígenas.

[15] SEP, Secretaría de Educación Pública, Contenidos educativos oficiales en lenguas nacionales indígenas.

[16] JW.org, Contenidos lingüísticos y educativos en español y náhuatl.

Samuel Pérez Zistecatl recibió el título de Ingeniería en Sistemas Computacionales por el Instituto Politécnico de Tlaxcala, región Poniente, Campus Hueyotlipan, actualmente cursa la Maestría en Sistemas Computacionales en el Tecnológico Nacional de México, Campus Apizaco. Su trabajo de investigación se centra en el desarrollo de modelos de lenguaje de gran tamaño (LLM) parala traducción automático-computacional entre español y náhuatl, integrando técnicas de minería de datos, procesamiento de lenguaje natural y arquitecturas neuronales avanzadas.

José Crispín Hernández-Hernández Obtuvo el título de Licenciado en Informática por el Instituto Tecnológico de Apizaco, Apizaco, Tlaxcala, México, en 1993. El grado de Maestro en Ciencias, en Ciencias Computacionales por el Instituto Tecnológico de Apizaco, Apizaco, Tlaxcala, México, en 1998. Y el grado de Doctorado en Ciencias de la Computación por la Université d'Angers. Campus Belle Beille, Francia en 2008. Sus principales áreas de investigación incluyen: metaheurísticas, optimización, bioinformática, biomedicina, lógica difusa, visión por computadora. aprendizaje máquina, minería de datos

Eduardo Sánchez Lucero recibió la licenciatura en Informática por el Instituto Tecnológico de Puebla, México, en 2002 y la maestría en ciencias en ciencias computacionales por el Instituto Tecnológico de Apizaco, México, en 2004. Actualmente está cursando el doctorado en ciencias de la ingeniería en el TecNM-Apizaco. Sus intereses de investigación incluyen el procesamiento digital de señales, reconocimiento de patrones, redes neuronales artificiales, modelos de lenguaje de gran tamaño, procesamiento del lenguaje natural y estudio de la lengua náhuatl.

Yesenia Nohemí González Meneses Obtuvo en el año 2021 el grado de Doctor en Ingeniería del Lenguaje y del Conocimiento por parte de la Universidad Autónoma de Puebla. Actualmente trabaja en la División de Estudios de Posgrado del Tecnológico Nacional de México Campus Apizaco. Sus áreas de interés incluyen reconocimiento automático de emociones, aplicaciones de IA, ingeniería de software e interacción humano computadora.

Edmundo Bonilla-Huerta recibió los grados de Licenciatura en Ciencias (B.Sc.) y Maestría en Ciencias (M.Sc.) en ciencias de la computación por parte del Instituto Tecnológico de Apizaco (ITA), Apizaco, Tlaxcala, México, en 1994 y 1996, respectivamente, y el grado de Doctorado (Ph.D.) en ciencias de la computación por la Universidad de Angers, Angers, Francia, en 2008.

Actualmente es Profesor en la División de Estudios de Posgrado e Investigación del ITA. Sus intereses de investigación incluyen la optimización, lógica difusa, aprendizaje automático (machine learning), técnicas para el análisis de datos de microarreglos y bioinformática**.**