UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

TRADUCCIÓN AUTOMÁTICA DE NÁHUATL A ESPAÑOL USANDO UN MODELO DE PLN

TESIS PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

JESÚS MANUEL DE DIOS LÓPEZ

BAJO LA DIRECCIÓN DE:

DRA. CRISTINA LÓPEZ RAMÍREZ

EN CODIRECCIÓN:

DR. RICARDO RAMOS AGUILAR

CUNDUACÁN, TABASCO, A: JULIO 2025

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

TRADUCCIÓN AUTOMÁTICA DE NÁHUATL A ESPAÑOL USANDO UN MODELO DE PLN

TESIS PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

JESÚS MANUEL DE DIOS LÓPEZ

BAJO LA DIRECCIÓN DE:

DRA. CRISTINA LÓPEZ RAMÍREZ

EN CODIRECCIÓN:

DR. RICARDO RAMOS AGUILAR

CUNDUACÁN, TABASCO, A: 07 Julio 2025

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día cinco del mes de Junio del año 2025, el que suscribe **Jesús Manuel De Dios López**, alumno del Programa de la **Maestro en Ciencias de la Computación** con número de matrícula **232H21003**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado de Maestría y titulada **Traducción automática de Náhuatl a Español usando un modelo de PLN**, dirigida por el Dra. Cristina López Ramírez y la Dr. Ricardo Ramos Aguilar.

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 05 de Junio de 2025.

Estudiante: Jesús Manuel De Dios López

Cunduacán, Tabasco, a 30 de junio de 2025
Oficio No. 1077/2025/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Jesús Manuel De Dios López
Egresado de la Maestría en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Traducción automática de náhuatl a español usando un modelo de PLN"***, para presentar examen y obtener el Grado de Maestro en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

Atentamente

MTE. Óscar Alberto González González
Director

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

C.c.p. Dr. Eddy Arquímedes García Alcocer. - Encargado del Despacho de la Coordinación de Posgrado DACYTI
Archivo.
Consecutivo.

M.T.E. OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda. C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

Carta de Cesión de Derechos

Cunduacán, Tabasco a 05 de Junio de 2025.

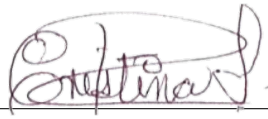
Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Traducción automática de Náhuatl a Español usando un modelo de PLN**.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR

Estudiante: Jesús Manuel De Dios López

TESTIGOS



Dra. Cristina López Ramírez

Dr. Ricardo Ramos Aguilar

Dedicatoria

- A Dios, principio y fin de todo lo que soy.
- A mi madre Candelaria López, gracias mami por su apoyo y palabras de motivación, te quiero mucho mamá.
- A mis profesores y directores, por compartir su conocimiento y sembrar en mí la pasión por aprender y cuestionar.
- Y a todas aquellas personas que, de una u otra forma, caminaron conmigo en este proceso: gracias por estar ahí.

Especialmente a mi amada esposa Lidia, esto es para ti, gracias mi amor, te amo.

Índice general

Índice de figuras

Índice de tablas

Resumen

El documento está estructurado como sigue:

El Capítulo 1 Este capítulo aborda la introducción al problema de la traducción automática entre náhuatl y español. Se presentan las preguntas de investigación, la hipótesis, y los objetivos tanto generales como específicos. Se justifica la necesidad de este trabajo destacando la importancia de preservar las lenguas indígenas y sus culturas. Además, se detalla la metodología utilizada para la investigación, basada en CRISP-DM, que incluye etapas como la revisión bibliográfica, análisis lingüístico del náhuatl, y la implementación y evaluación del modelo de procesamiento de lenguaje natural para la traducción automática .

En el Capítulo 2 En este capítulo se revisa el estado del arte del procesamiento de lenguaje natural y la traducción automática, contextualizando la propuesta dentro del ámbito de la inteligencia artificial y las ciencias de la computación. Se describen conceptos clave como la lingüística, el aprendizaje automático, la traducción automática, y las métricas de evaluación como BLEU. También se exploran tecnologías y herramientas utilizadas en el campo, como embeddings y vectores, y se discuten los desafíos particulares de trabajar con lenguas de bajos recursos .

En el Capítulo 3 Este capítulo se enfoca en el desarrollo del modelo de traducción automática. Se describe el uso de dos datasets: “Don Quijote de la Mancha” y el “Spanish Billion Words Corpus”. Se detallan los procesos de generación de embeddings, tokenización, y técnicas de procesamiento del lenguaje como el stemming y la lematización. También se presentan los primeros resultados del modelo Word2Vec y las pruebas de palabras cercanas y analogías. Se analiza la mejora obtenida al utilizar un dataset más grande y diverso para entrenar el modelo, lo que demuestra la importancia de contar con datos extensivos para mejorar la precisión y relevancia del modelo .

En el Capítulo 4 En este capítulo se presentan los resultados obtenidos al evaluar el modelo Transformer en la tarea de traducción automática entre español y náhuatl, utilizando datos sintéticos y reales. Con datos sintéticos, el modelo mostró un desempeño aceptable en oraciones simples, logrando métricas moderadas como ROUGE y BLEU, aunque presentó dificultades con oraciones más complejas, reflejadas en incoherencias y uso de tokens `[UNK]`. En los datos reales, el desempeño fue desigual: en oraciones cortas, el modelo generó traducciones precisas, alcanzando métricas altas, pero en oraciones largas o con vocabulario técnico, las salidas fueron incoherentes y las métricas significativamente bajas. Esto evidencia problemas para manejar contextos complejos y la necesidad de un vocabulario ampliado. En general, se destacan los logros en escenarios controlados y las limitaciones en datos reales, sentando las bases para mejorar el modelo con más datos y técnicas avanzadas.

En el Capítulo 5 Este capítulo presenta las principales conclusiones obtenidas a partir del desarrollo del proyecto de traducción automática de Náhuatl a Español utilizando modelos de Procesamiento de Lenguaje Natural (PLN). Asimismo, se detallan las contribuciones académicas del trabajo y se proponen diversas líneas de investigación futura, enfocadas en la mejora, ampliación y aplicación de los modelos desarrollados a otros contextos lingüísticos y tecnológicos.

Palabras clave: Inteligencia Artificial, Redes neuronales, Red nauronal Transformers, Procesamiento del Lenguaje natural

Abstract

The document is structured as follows:

The Capítulo 1 This chapter addresses the introduction to the problem of automatic translation between Nahuatl and Spanish. It presents the research questions, hypothesis, and both general and specific objectives. The necessity of this work is justified by highlighting the importance of preserving indigenous languages and their cultures. Additionally, the methodology used for the research is detailed, based on CRISP-DM, which includes stages such as bibliographic review, linguistic analysis of Nahuatl, and the implementation and evaluation of the natural language processing model for automatic translation.

In Capítulo 2 This chapter reviews the state of the art of natural language processing and automatic translation, contextualizing the proposal within the field of artificial intelligence and computer science. Key concepts such as linguistics, machine learning, automatic translation, and evaluation metrics like BLEU are described. Technologies and tools used in the field, such as embeddings and vectors, are also explored, and the specific challenges of working with low-resource languages are discussed.

In Capítulo 3 This chapter focuses on the development of the automatic translation model. It describes the use of two datasets: “Don Quijote de la Mancha” and the “Spanish Billion Words Corpus”. The processes of generating embeddings, tokenization, and language processing techniques such as stemming and lemmatization are detailed. The initial results of the Word2Vec model and tests of similar words and analogies are also presented. The improvement achieved by using a larger and more diverse dataset to train the model is analyzed, demonstrating the importance of extensive data to improve the model’s accuracy and relevance.

In **??**, this chapter presents the results obtained from evaluating the Transformer model in the

task of automatic translation between Spanish and Náhuatl, using synthetic and real data. With synthetic data, the model demonstrated acceptable performance in simple sentences, achieving moderate metrics such as ROUGE and BLEU, although it faced difficulties with more complex sentences, reflected in inconsistencies and the use of `[UNK]` tokens. For real data, the performance was uneven: in short sentences, the model generated precise translations, achieving high metrics, but for long sentences or those with technical vocabulary, the outputs were incoherent, and the metrics were significantly lower. This highlights challenges in handling complex contexts and the need for an expanded vocabulary. Overall, the achievements in controlled scenarios and the limitations in real data are emphasized, laying the groundwork for improving the model with more data and advanced techniques.

In Capítulo 5 This chapter presents the main conclusions obtained from the development of the automatic translation project from Nahuatl to Spanish using Natural Language Processing (NLP) models. It also details the academic contributions of the work and proposes several lines of future research, focused on the improvement, expansion, and application of the developed models to other linguistic and technological contexts.

Keywords: Artificial Intelligence, Neural Networks, Transformer Neural Network, Natural Language Processing

Capítulo 1

Generalidades

1.1. Introducción

La traducción es el proceso de comprender un texto en un idioma (texto de origen) y convertirlo en otro idioma (texto traducido), se originó en la época de la piedra Rosetta (196 a.C.) y se consolidó con la traducción de textos religiosos como la Septuaginta [1] y la Vulgata [2]. La expansión del imperio árabe en la Edad Media impulsó la traducción en áreas como la ciencia y la filosofía. En el siglo XII, la Escuela de Traductores de Toledo tuvo un impacto significativo en el progreso del campo de la traducción, más tarde con la imprenta y las lenguas vernáculas en el siglo XV, la traducción experimentó grandes avances, en nuestros días la epoca moderna y contemporánea, la globalización e internet han llevado la traducción a una era de profesionalización y desarrollo de herramientas (Nuadda, 2020).

Las poblaciones indígenas en México tienen costumbres y formas de vida únicas, incluyendo su vestimenta, comida, celebraciones y gobierno. La lengua es un elemento fundamental que les da identidad, en el Censo de 2020 (INEGI, 2022) por el INEGI (Instituto Nacional de Estadística y Geografía), en el cual se detectó que 7.4 millones de individuos de tres años o más utilizaban una lengua indígena, lo que representa el 6.1 % de la población de ese grupo de edad. La mayoría de ellos también hablaban español y los estados con una población considerable de personas

[1] Biblia griega, comúnmente llamada Biblia Septuaginta o Biblia de los Setenta.
[2] La Vulgata es una traducción de la Biblia al latín.

que hablan lenguas indígenas son Oaxaca, Chiapas, Yucatán y Guerrero. En contraste, estados como Zacatecas, Guanajuato, Aguascalientes y Coahuila registran una presencia más limitada de hablantes indígenas; en México se hablan 68 lenguas indígenas, siendo las más comunes el náhuatl, maya y tseltal. Alrededor del 12 % de las personas que hablan una lengua indígena no hablan español.

Un estudio muy relevante sobre el análisis de las lenguas indígenas basado en procesamiento de lenguaje natural, es el realizado sobre el $AmericasNLI$ (Kann et al., 2022) un conjunto de datos de inferencia de lenguaje natural que abarca 10 lenguajes indígenas de América entre los cuales esta el Nahuatl, se realizaron experimentos con modelos preentrenados explorando el aprendizaje junto con la adaptación del modelo, se utilizó $AmericasNLI$ para evaluar el rendimiento de diferentes modelos de traducción automática para esos lenguajes, en este estudio se descubrió que el rendimiento de los modelos preentrenados sin adaptación en todos los idiomas de $AmericasNLI$ es deficiente, pero la adaptación del modelo a través del preentrenamiento continuo produce mejoras, además se encontró que los enfoques basados en la traducción para la inferencia de lenguaje natural superan a todos los demás modelos en esa tarea.

1.2. Planteamiento del problema

La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) identifica un total de 6 mil lenguas en todo el mundo, y de ellas, alrededor del 50 por ciento enfrenta el peligro de extinguirse (INPI, 2016a). En México están reconocidas 68 lenguas indígenas según el Pleno del senado de la República (Senado, 2021). Existe un catálogo de estas lenguas y sus variantes titulado “catálogo de lenguas indígenas mexicanas” , dicho documento está realizado por el Instituto Nacional de Lenguas Indígenas (INALI) (de Lenguas Indígenas, 2008), sin embargo, muchas de estas lenguas indígenas presentan preocupación por el peligro de extinción que enfrentan, esto se mención en el artículo de la UNAM Global (y David Novoa, 2022). Por otro lado desde la época prehispánica, la diversidad lingüística ha disminuido significativamente, aunque había poliglosia (coexistencia de varias lenguas), algunas, como el náhuatl, el zapoteco y el purépecha, se volvieron dominantes, desplazando a otras lenguas. Actualmente, se hablan

62 lenguas indígenas nacionales y 364 variedades lingüísticas en México, pero muchas están en riesgo de extinción debido a factores la discriminación hacia lenguas originarias.

En 2019, la ONU proclamó el Decenio Internacional de las Lenguas Indígenas (2022-2032) (y David Novoa, 2022) para fomentar la preservación y promoción de estas lenguas. La investigadora Lucero Meléndez Guadarrama (Investigadora SNI I de la UNAM: Lingüística) subraya la importancia de involucrar a los hablantes en el proceso de revitalización de sus lenguas, ya que el deseo de preservar una lengua varía entre individuos y comunidades.

El artículo publicado por Instituto Nacional de los Pueblos Indígenas (INPI), el 1 de abril de 2016 (INPI, 2016b), destaca la crítica situación de las lenguas indígenas en México, de las 6 mil lenguas existentes en el mundo reconocidas por la UNESCO, la mitad está en riesgo de desaparecer. México es reconocido por su diversidad cultural y lingüística, alberga al menos 68 lenguas indígenas (Senado, 2021), de las cuales unas 23 enfrentan un alto riesgo de extinción, este se manifiesta en el bajo número de hablantes (menos de 2000), la dispersión geográfica, el predominio de hablantes adultos, y la disminución en su transmisión intergeneracional. Además, existen factores estructurales que restringen su continuidad, como la marginación en ámbitos públicos e institucionales, la reducción de su empleo en contextos comunitarios y familiares, la falta de representación en los medios de comunicación y su menor reconocimiento en comparación con el español.

La adopción de tecnologías digitales ha seguido dos tendencias distintas que reflejan la situación socioeconómica y política del país. Por un lado, se observa una amplia exclusión de la mayoría de la población indígena del ciberespacio, mientras que, por otro lado, se destacan las numerosas oportunidades que los pueblos, comunidades e individuos indígenas tienen para mejorar sus condiciones de vida, tanto a nivel colectivo como personal (Sandoval-Forero, 2013).

Debido a esto, la escasez de los recursos bibliográficos y el acceso a ellos es muy limitada, ya que están escritos en lenguas propias de los pueblos indígenas, esto dificulta su comprensión y lectura y crea barreras tanto para los hablantes en una lengua como el español, como para

los hablantes en lengua indígena. Por ello se tiene la necesidad crear una herramienta capaz de romper esa barrera con el objetivo de preservar textos o querer saber hechos históricos que están escritos en diversas lenguas indígenas.

Lo anterior lo podemos resolver utilizando modelos computacionales que nos ayuden a realizar una traducción de estas lenguas indígenas de una manera eficiente, es aquí donde los modelos de procesamiento de lenguaje natural (PLN) como lo son las redes neuronales nos acercan a resultados aceptables para la traducción de estas lenguas a un idioma como el español.

El dataset [3] que usaremos es un texto escrito en nahuatl, y como base su traducción al español para comprar el resultado al entrenar la red nueronal.

1.3. Definición del problema

1.3.1. Delimitación de la investigación

- Las limitaciones para este trabajo son el acceso a los recursos bibliográficos, puesto que esta lengua indígena esta dentro de los lenguas de bajos recursos.

- Disponibilidad de datos: La escasez de corpus bilingües extensos y diversificados en Náhuatl-Español puede limitar la capacidad de entrenamiento y precisión del sistema.

- Variaciones: La diversidad dialectal del Náhuatl puede no ser completamente representada, limitando la aplicabilidad del sistema a ciertas variantes del idioma.

- Recursos tecnológicos: Las limitaciones en términos de financiamiento, hardware, y software pueden afectar el desarrollo y la implementación del sistema.

[3] Repositorio del dataset: https://github.com/AmericasNLP/americasnlp2023/tree/main/data/nahuatl-spanish

1.4. Pregunta de investigación e hipótesis

1.5. Pregunta de investigación

¿Por medio de técnicas de aprendizaje automático es posible construir un modelo para la traducción de nahuatl-español?

1.6. Hipótesis

Una red neuronal Transformer para traducción automática de náhuatl a español en un entorno de bajos recursos logrará resultados comparables de al menos valores de 14.1 % de BLEU y Chrf de 25.91 %.

1.7. Objetivo general

Desarrollar un modelo de traducción automática Náhuatl-Español utilizando técnicas de aprendizaje automático, con el fin de facilitar la comunicación entre hablantes de ambas lenguas y contribuir a la preservación y difusión de la cultura lingüística náhuatl.

1.8. Objetivos específicos

- Desarrollar un algoritmo para la extracción automatizada de características lingüísticas relevantes del idioma Náhuatl.
- Validar un conjunto de datos bilingüe Náhuatl-Español para entrenamiento y pruebas de modelos de traducción automática.
- Implementar y entrenar un modelo de aprendizaje automático para la traducción Náhuatl-Español
- Realizar pruebas de rendimiento en los modelos de traducción automática utilizando alguna métrica estándar como la puntuación BLEU para evaluar la calidad de las traducciones de bajo recursos.

1.9. Alcances y limitaciones

1.9.1. Alcances

- Implementar un modelo que para la traducción de náhuatl a español.

- Cobertura de dialectos Náhuatl: El sistema podría enfocarse en uno o varios dialectos específicos del Náhuatl, dependiendo de la disponibilidad de datos y recursos.

- Tipos de texto: Especificar si el sistema se centrará en textos literarios, académicos, conversacionales, o una combinación de estos.

- Herramientas tecnológicas: Utilizar tecnologías de vanguardia en procesamiento de lenguaje natural y aprendizaje automático para maximizar la eficiencia y precisión del sistema.

1.9.2. Limitaciones

- Las limitaciones para este trabajo son el acceso a los recursos bibliográficos, puesto que esta lengua indígena está dentro de los lenguas de bajos recursos.

- Disponibilidad de datos: La escasez de corpus bilingües extensos y diversificados en Náhuatl-Español puede limitar la capacidad de entrenamiento y precisión del sistema.

- Variaciones: La diversidad dialectal del Náhuatl puede no ser completamente representada, limitando la aplicabilidad del sistema a ciertas variantes del idioma.

- Recursos tecnológicos: Las limitaciones en términos de financiamiento, hardware, y software pueden afectar el desarrollo y la implementación del sistema.

1.10. Justificación

Las lenguas indígenas a menudo llevan consigo una rica herencia cultural, incluyendo tradiciones, historias, conocimientos ancestrales y maneras de interpretar el mundo únicas. La traducción permite que estas riquezas culturales sean compartidas y preservadas para las generaciones futuras, evitando la pérdida de patrimonio cultural. Las personas mayores que hablan principalmente su lengua materna pueden compartir su sabiduría y vivencias con las generaciones más jóvenes, quienes podrían tener mayor familiaridad con distintos idiomas, como el español.

Como lo describimos anteriormente en México están reconocidas 68 lenguas indígenas(Senado, 2021), perder estas lenguas es perder parte de la historia que es de suma importancia y riqueza del patrimonio cultural de nuestro país.

La aproximación de las comunidades indígenas al entorno tecnológico presenta sin duda diversos desafíos, siendo uno de los más significativos la cuestión de la comunicación. En este sentido, algunos de estos grupos no disponen de una forma escrita de su lengua y, al involucrarse en el uso de tecnologías, a menudo se ven obligados a emplear idiomas dominantes como el español o el inglés (J. M. Mager, 2017). La traducción desempeña un papel esencial en la garantía del acceso de las comunidades indígenas a servicios cruciales, tales como atención médica, educación y asistencia legal. En la actualidad, de las 68 lenguas indígenas que se hablan en México, las más comunes son el náhuatl (22.4 %), el maya (10.5 %) y el tseltal (8.0 %). De cada 100 personas de tres años o más que hablan una lengua indígena, aproximadamente 12 no tienen conocimientos de español.(INEGI, 2022).

Es aquí donde el uso de modelos computaciones quita esa barrera que impide la comunicación entre diferentes pueblos hablantes de lenguas indígenas. Por ello es necesario implementar un traductor en especial de náhuatl a español que permita ese acercamiento, conservar y compartir las diversas tradiciones culturales entre estos pueblos facilitando el diálogo intercultural.

1.11. Metodología utilizada

A continuación se menciona las etapas de la investigación, basándome en la metodología CRISP-DM como base para describir los puntos esenciales en el la Figura 1.1, para el desarrollo del modelo.

- Revisión bibliográfica
- Análisis lingüístico del Náhuatl
- Análisis de un corpus Náhuatl
- Pruebas y evaluación del modelo
- Evaluación del modelo
- Implementación del modelo de PLN para la traducción automática

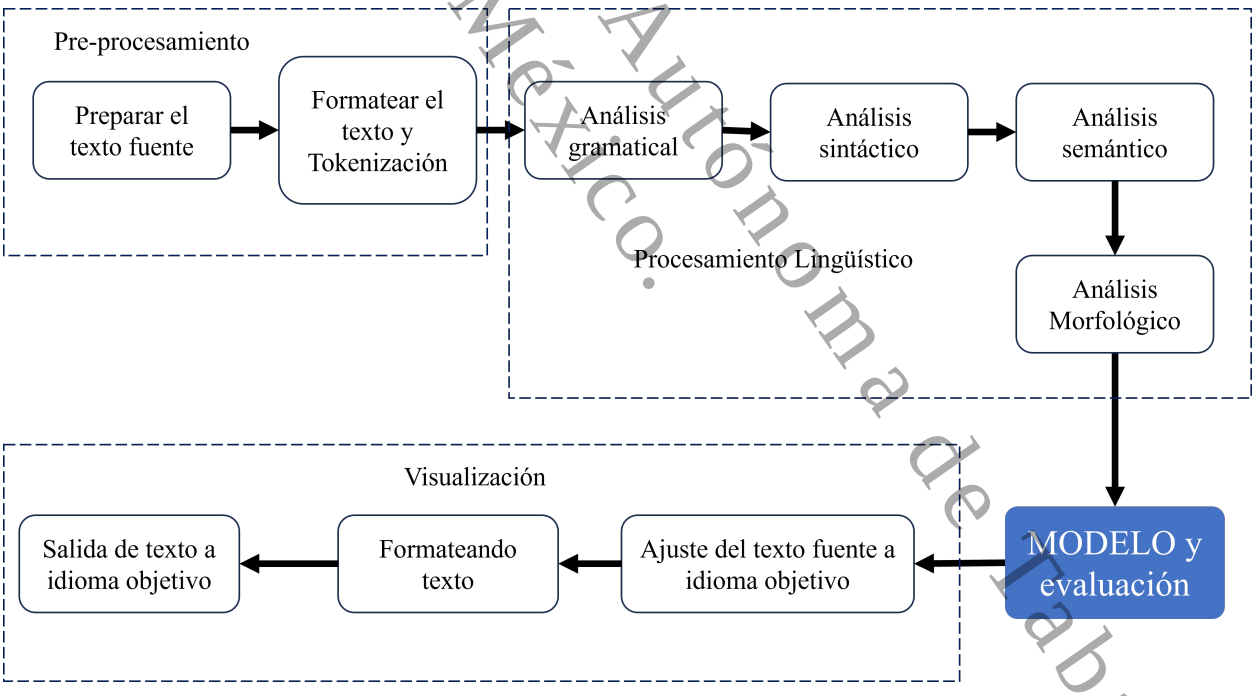


Figura 1.1. La imagen muestra un diagrama de flujo lineal representando los pasos de un proceso de traducción automática.

El diagrama de flujo de trabajo para la traducción automática, está compuesto por varias etapas que están conectadas entre sí. Aquí describo las etapas de forma general:

- Pre-procesamiento: Los datos se preparan para el proceso de traducción. Esto podría incluir limpieza de datos, preparar el texo fuente, tokenización, etc.
- Procesamiento lingüistico: Esta etapa nos permitirá que el modelo pueda entender y trabajar con texto de la misma manera en que lo haría un ser humano.
- Modelo y evaluación
    - Desarrollo del modelo: Selección y entrenamiento de un modelo de traducción automática, en nuestro caso será un modelo de inteligencia artificial como las redes neuronales.
    - Ajuste del modelo: Afinar el modelo con un conjunto de datos específico para mejorar su precisión y manejo del contexto y las sutilezas del lenguaje.
    - Evaluación de la traducción: Utilizar alguna métrica de evaluación para bajos recursos, como BLEU (*Bilingual Evaluation Understudy*), para medir la calidad de las traducciones del modelo comparándolas con una referencia humana.
    - Retroalimentación para mejora: Basándonos en los resultados de la evaluación, se realiza una iteración sobre el modelo para optimizarlo y mejorar su rendimiento. Esto puede incluir cambios en la arquitectura del modelo, el entrenamiento con más datos o la modificación de los algoritmos de preprocesamiento y postprocesamiento.
- Visualización
    - Ajustes del texto: Los datos traducidos se refinan, lo cual podría implicar corrección de errores, ajuste de estilo, etc.
    - Formateo y visualización: Se evalúa la calidad de la traducción. Este paso podría incluir revisión humana o evaluación automática de la precisión, además los resultados de la evaluación pueden usarse para mejorar el modelo de traducción.
    - Salida: En esta etapa tendremos la traducción completa del texto

Capítulo 2

Marco teórico

2.1. Lingüística

La lingüística es la disciplina que analiza exhaustivamente el lenguaje humano hablado y escrito, tanto en su conjunto como en sus manifestaciones específicas, lo que incluye los actos comunicativos y los sistemas de variación geográfica que, de manera tradicional o por convención, se denominan lenguas (Coseriu y Polo, 1986).

2.2. Náhuatl

El término “nahua” o “náhuatl” se emplea para hacer alusión a un grupo indígena y a un conjunto de lenguas indígenas que están estrechamente interconectadas en México, la palabra “náhuatl” representa la forma adaptada al español de esta lengua y se traduce como “cosas que suenan bien”, cabe destacar que estas lenguas poseen diversos nombres con los que se autodenominan, los cuales varían en función de la lengua, la ubicación geográfica y las comunidades específicas (México, 2020).

2.3. PLN

El Procesamiento de Lenguaje Natural (PLN), también conocido como NLP por sus siglas en inglés, se refiere a la capacidad de una máquina para analizar y comprender la información comunicada en lenguaje humano, no limitándose únicamente a las letras o los sonidos del lenguaje. En este contexto, no se considera que un perico sea un animal que hable, por lo que dispositivos o software como una contestadora telefónica común, una impresora o un procesador de palabras como Microsoft Word no se consideran componentes de PLN, en cambio, un traductor automático definitivamente se clasifica como una herramienta de PLN, ya que tiene la capacidad de comprender y traducir texto entre idiomas. (Gelbukh, 2010).

2.4. Aprendizaje Automático

Hay varias definición para Aprendizaje Automático (AA) como lo describe Jorge Díaz (Díaz-Ramírez, 2021) con diferentes autores, una de ellas es la capacidad de permitir que las computadoras aprendan sin programación explícita, otra definición descrita en el artículo es que se considera un programa de computadora que mejora su rendimiento en una tarea T, medida por una métrica de rendimiento R, a medida que adquiere experiencia E. una definición más es que autores lo ven como la ciencia y el arte de programar computadoras para que aprendan de datos. En resumen, el Aprendizaje Automático se refiere a un conjunto de métodos y enfoques que capacitan a las computadoras para adquirir conocimiento y mejorar su desempeño a partir de la experiencia y los datos disponibles, lo que ha resultado en un campo interdisciplinario fundamental en la inteligencia artificial.

2.5. Traducción automática

La traducción automática tiene objetivos claros: tomar el texto escrito en un lenguaje y traducirlo a otro, manteniendo el mismo significado. En general el proceso de traducción automática sigue tres pasos: primero, el texto en el lenguaje original se transforma a una representación intermedia, luego, de acuerdo a la morfología del lenguaje destino, se realizan

modificaciones a esta representación intermedia y por último ésta se transforma al lenguaje destino(Hernández y Gómez, 2013).

2.6. BLEU (*Bilingual Evaluation Understudy*)

Es utilizada para evaluar la traducción de un texto automáticamente respecto a un conjunto de referencias de traducciones humanas. BLEU compara la coincidencia de frases cortas entre la traducción automática y las traducciones de referencia y produce una puntuación que refleja la similitud. BLEU produce valores entre 0 y 1 (a menudo escalados de 0 a 100 %). Un valor cercano a 1.0 (100 %) indicaría que la traducción es prácticamente idéntica a una referencia. Las puntuaciones más altas indican una mayor correspondencia con la calidad de una traducción humana (Huerta et al., 2015).

2.7. ROUGE-1 (*Recall-Oriented Understudy for Gisting Evaluation*)

ROUGE-1 es un conjunto de métricas pensado originalmente para evaluación de resúmenes automáticos, pero también aplicable a traducción y otras tareas de generación de lenguaje. ROUGE-1 calcula la superposición de unigramas (palabras individuales) entre la salida y la referencia, su rango va de 0 a 1, donde 1 indicaría que todas las palabras de la referencia se encuentran en el texto evaluado (lo cual implicaría una cobertura completa del contenido, aunque no necesariamente en orden). Valores bajos (ej. < 0.2) sugieren que el modelo omitió muchos detalles importantes o usó vocabulario muy distinto al humano. En cambio, valores altos, por encima de $\sim 0.4(40\,\%)$, suelen considerarse buenos, significando que una proporción significativa del contenido relevante fue capturado (Lin, 2004; Schluter, 2017).

2.8. ROUGE-2 (*Recall-Oriented Understudy for Gisting Evaluation*)

ROUGE-2 es una variante del ROUGE-N enfocada en bigramas (pares de palabras consecutivas). Es decir, mide la superposición de secuencias de dos palabras entre el texto generado y el de referencia. Al considerar bigramas, ROUGE-2 captura en cierta medida no solo la presencia de las mismas palabras, sino también algo de la estructura o contexto local de la frase. Su valor oscila entre 0 y 1 (0 a 100 %), donde 1 implicaría que cada par de palabras adyacentes del original se encuentra en el texto del modelo – algo muy poco probable salvo que el texto generado sea casi una copia exacta. ROUGE-2 alto (ej. > 0.20) significa que el modelo no solo incluyó muchas palabras relevantes, sino que también replicó varias combinaciones de palabras tal como aparecen en la referencia, reflejando mayor coincidencia en expresiones y posiblemente mejor coherencia con el original (Lin, 2004; Schluter, 2017).

2.9. ChrF (*Character n-gram F-score*)

ChrF es una métrica de evaluación más reciente que calcula la similitud entre una salida generada y el texto de referencia a nivel de caracteres, en lugar de a nivel de palabras. Al igual que las métricas anteriores, chrF se acota entre 0 y 1 (o 0 a 100 %). Un valor de 1.0 significaría que la traducción generada coincide exactamente con la referencia hasta en sus secuencias de caracteres (lo que implica identidad total del texto). un chrF de 50 % indicaría ya una alta similitud, mientras que valores bajos como 20 % evidencian muy poca coincidencia entre la salida y la referencia (Popović, 2016, 2017).

2.10. Definición de Vectores

Un **vector** es una entidad matemática que tiene magnitud y dirección, y es una de las principales herramientas utilizadas en diversas áreas de la física y las matemáticas. Los vectores

se pueden representar de varias formas, pero la más común es a través de coordenadas en un sistema de referencia (Marsden et al., 1991).

2.11. Incrustaciones (*Embeddings*)

Los "*embeddings*" son representaciones numéricas de palabras en forma de vectores, utilizadas en modelos de semántica distribucional o representaciones distribuidas. Estas representaciones capturan información semántica y sintáctica, y son calculadas a partir de la co-ocurrencia contextual de palabras (Mandelbaum y Shalev, 2016).

2.11.1. Corpus

Un corpus, en el contexto de la lingüística y el procesamiento del lenguaje natural, es un conjunto de datos que consiste en recursos lingüísticos tanto digitales como digitalizados. (LibreTexts, 2023).

2.11.2. Lenguas de bajos recursos

Una lengua se clasifica como de bajos recursos cuando no cuenta con un corpus, tecnologías digitales y otras herramientas que simplifiquen su procesamiento. Esta carencia puede originarse debido a un número limitado de hablantes o a diversas circunstancias que restringen la disponibilidad de textos adecuados. (Castillo, 2018).

2.12. *Transformers*

Las redes neuronales *transformer* son una categoría moderna de redes neuronales diseñadas para manejar secuencias, y están fundamentadas en el concepto de autoatención, estas redes han demostrado ser altamente eficaces en el procesamiento del lenguaje natural, y en la actualidad están impulsando avances significativos en este campo. (Wolfram, 2023).

2.13. *Word2vec*

Word2vec, publicada en 2013, es una metodología destinada al ámbito del procesamiento de lenguaje natural. Este algoritmo emplea un modelo basado en redes neuronales con el fin de adquirir conocimientos sobre las relaciones entre palabras, utilizando para ello un extenso corpus textual (Ruiz de Villa, 2023).

2.14. Conceptos y teorías fundamentales de la investigación

2.15. Literatura relacionada

Después de más de siete décadas de progreso en la traducción automática, se han logrado notables avances, especialmente en años recientes, la calidad de las traducciones ha experimentado una mejora significativa debido a la aparición de la traducción automática enfocadas en redes neuronales. (*NMT, Neural Machine Translation*) (Wang et al., 2022).

En los principios de la traducción de un idioma a otro, no se consideraba algo relevante para las sociedades, ya que probablemente no existía una necesidad significativa de comunicación con pueblos extranjeros. Esta percepción cambió, quizás no fue hasta los principios de la Segunda Guerra Mundial, cuando la necesidad de comunicación con personas que hablaban diferentes idiomas se volvió crucial.

En la siguiente Tabla 2.1, se presentan los avances generales en la traducción automática, destacando las contribuciones de diversos autores en investigaciones que han contribuido al desarrollo de esta tecnología. Esto indica cómo a lo largo del tiempo, especialmente después de la Segunda Guerra Mundial, la traducción automática ha experimentado mejoras notables.

Año	Contribución/Descubrimiento	Autor(es)
1947	Propuesta inicial de traducción automática y su importancia para la comunicación internacional (Hutchins, 1997)	Warren Weaver.
1990	Introducción de métodos estadísticos en traducción automática (Brown et al., 1990)	Peter F. Brown y colaboradores.
1996	Nuevo modelo para alineación de palabras en traducción estadística usando HMM (Vogel et al., 1996)	Stephan Vogel.
2001	Modelo de traducción estadística basado en sintaxis (Yamada y Knight, 2001)	Kenji Yamada y Kevin Knight.
2013	Modelos de Traducción Continua Recurrente basados en representaciones continuas (Kalchbrenner y Blunsom, 2013)	Nal Kalchbrenner y Phil Blunsom.
2014	Método de aprendizaje de secuencias en traducción automática con LSTM (Sutskever et al., 2014)	Ilya Sutskever.
2015	Mecanismos de atención en la traducción automática (Luong et al., 2015)	Thang Luong.
2016	GNMT, un sistema avanzado de Traducción Automática Neural (Wu et al., 2016)	Yonghui Wu y equipo de Google.
2016	Sistema web para traducciones de español a náhuatl (Gutierrez-Vasques et al., 2016)	Desarrolladores de Axolotl.
2017	Herramientas para el procesamiento del lenguaje wixárika (J. M. Mager, 2017)	Jesús Manuel.
2020	Avances en traducción automática neural y técnicas asociadas (Tan et al., 2020)	Zhixing Tana.
2022	Traducción automática e inferencia de lenguaje natural para lenguas indígenas americanas (Kann et al., 2022)	Katharina Kann y colaboradores.
2023	Corpus paralelo español-Mazateco y español-Mixteco para traducción automática (Tonja et al., 2023)	A. Tonja.

Tabla 2.1. En esta tabla se muestra un resumen en orden cronológico de acontecimiento en el desarrollo de la traducción automática y el procesamiento de lenguaje natural desde 1947 hasta 2023, donde se refleja la evolución del campo de la TA desde sus conceptos iniciales hasta técnicas modernas y sistemas avanzados.

2.16. Desafíos del procesamiento automático del náhuatl

En náhuatl, es común formar palabras largas que incluyen varios significados unidos, lo que lo convierte en una lengua aglutinante y capaz de condensar oraciones completas en

una sola palabra, lo que implica que múltiples morfemas se combinan para formar palabras complejas que expresan significados que, en otras lenguas, requerirían frases completas. Esta característica morfológica presenta desafíos significativos para el procesamiento automático del lenguaje (PLN), ya que las herramientas estándar, diseñadas principalmente para lenguas de alta disponibilidad de recursos como el inglés o el español, no manejan eficientemente la segmentación y análisis de estas estructuras complejas.

Investigaciones han demostrado que, aunque los métodos automáticos pueden segmentar adecuadamente muchos morfos funcionales, la complejidad morfológica del náhuatl dificulta tareas como la tokenización, lematización y alineación morfema a morfema en traducciones automáticas. Además, estudios han identificado que una cantidad significativa de información contenida en los morfemas del náhuatl no tiene un equivalente directo en lenguas como el español, lo que resulta en pérdidas de información durante la traducción automática (M. Mager et al., 2018; Méndez-Cruz y Arroyo-Fernández, 2022).

Estos desafíos resaltan la necesidad de desarrollar herramientas y modelos de PLN adaptados a las particularidades morfológicas del náhuatl. La presente tesis contribuye en este sentido, al abordar y proponer soluciones específicas para el procesamiento automático de esta lengua, destacando su importancia en la preservación y revitalización de las lenguas indígenas (García et al., 2020).

2.17. Marco tecnológico

En esta lista se describen las herramientas, software y hardware que se han utilizado para la elaboración de este trabajo

- **CPU**: Procesador Intel® Core™ i5-13600K.
- **Unidad de Procesamiento Gráfico (GPU)**: NVIDIA RXT 4060 Ti. Las GPU de NVIDIA son las más recomendadas debido a su compatibilidad con CUDA y cuDNN, que son optimizaciones esenciales para el entrenamiento de redes neuronales.
- **Memoria RAM**: 16 GB. Esta cantidad de memoria es necesaria para manejar grandes volúmenes de datos y realizar operaciones simultáneas de manera eficiente.
- **Almacenamiento**: Unidad de Estado Sólido (SSD) de 1TB para el sistema operativo y los datos del proyecto. Las SSDs son preferidas por su alta velocidad de lectura/escritura, lo que mejora el rendimiento general del sistema.
- **Lenguaje de Programación**: *Python 3.12.3*.
- ***Frameworks de Deep Learning***: Se utilizaron varios *frameworks* populares para el desarrollo y entrenamiento de modelos de aprendizaje profundo, incluyendo *TensorFlow, PyTorch, Keras, y Word2Vec*.
- **Controladores y Bibliotecas**: *CUDA y cuDNN*, junto con los controladores correspondientes para la *GPU*, son esenciales para optimizar el rendimiento de los cálculos y acelerar el entrenamiento de redes neuronales.

Cada uno de estos componentes ha sido cuidadosamente seleccionado para garantizar un entorno de desarrollo robusto y eficiente, adecuado para las tareas complejas de procesamiento y análisis de datos requeridas en este trabajo.

Capítulo 3

Modelo de Transformers para la traducción automática

El primer paso para desarrollar el modelo de traducción automática es usar un conjunto de textos en español. La meta es crear representaciones de palabras (embeddings) y analizar los resultados obtenidos. Luego, se aplicará este modelo al conjunto de textos en náhuatl.

3.1. Descripción del Dataset

3.1.1. Descripción del Dataset “Don Quijote de la Mancha”

El dataset de “Don Quijote de la Mancha” como un corpus sería una colección estructurada y organizada del texto completo de la obra literaria escrita por Miguel de Cervantes Saavedra.

Metadatos:

- Título: Don Quijote de la Mancha
- Autor: Miguel de Cervantes Saavedra
- Idioma: Español
- Género: Novela, Literatura clásica

Cada parte y capítulo puede incluir metadatos adicionales:

Capítulos

- Número de capítulos
- Resumen breve de cada parte
- Título del capítulo (si aplica)
- Inicio y fin del texto del capítulo

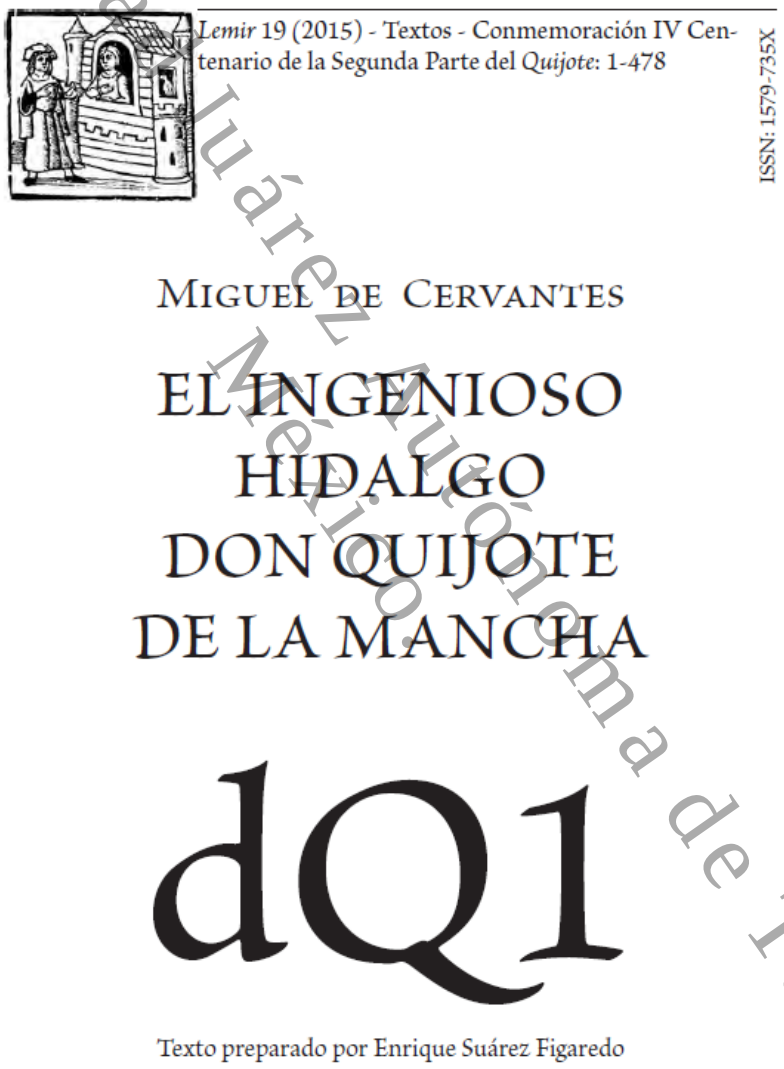


Figura 3.1. Portada del libro digital de Don Quijote de la Mancha

3.1.2. Descripción del Dataset "*Spanish Billion Words Corpus*"

El "*Spanish Billion Words Corpus*" es un extenso conjunto de datos textuales en español diseñado para tareas de procesamiento de lenguaje natural (NLP), este Dataset fue tomado del sitio: www.kaggle.com.

1. Contenido del Corpus
    - El corpus incluye más de mil millones de palabras de texto en español.
    - Está compuesto por una variedad de fuentes, incluyendo artículos de noticias, libros, páginas web, y otros textos en español.
2. Estructura del Corpus
    - Formato: El corpus se presenta en formato de texto plano, estructurado en frases y párrafos.
    - Etiquetado: Incluye etiquetas de partes del discurso (POS tagging) y anotaciones gramaticales, lo que facilita el análisis lingüístico y el entrenamiento de modelos de lenguaje.
3. Tamaño y Cobertura
    - Contiene más de mil millones de palabras, proporcionando una cobertura amplia del idioma español.
    - Cubre una amplia gama de dominios y registros lingüísticos, desde lenguaje formal en artículos académicos hasta lenguaje más coloquial en publicaciones web.

3.1.3. Modelo Transformer

La parte encerrada en rojo en la imagen representa el proceso inicial en un modelo *Transformer*, específicamente el paso de incorporación de información posicional y la generación de *embeddings* de entrada, esta parte inicial es fundamental para preparar los datos de entrada en el modelo.

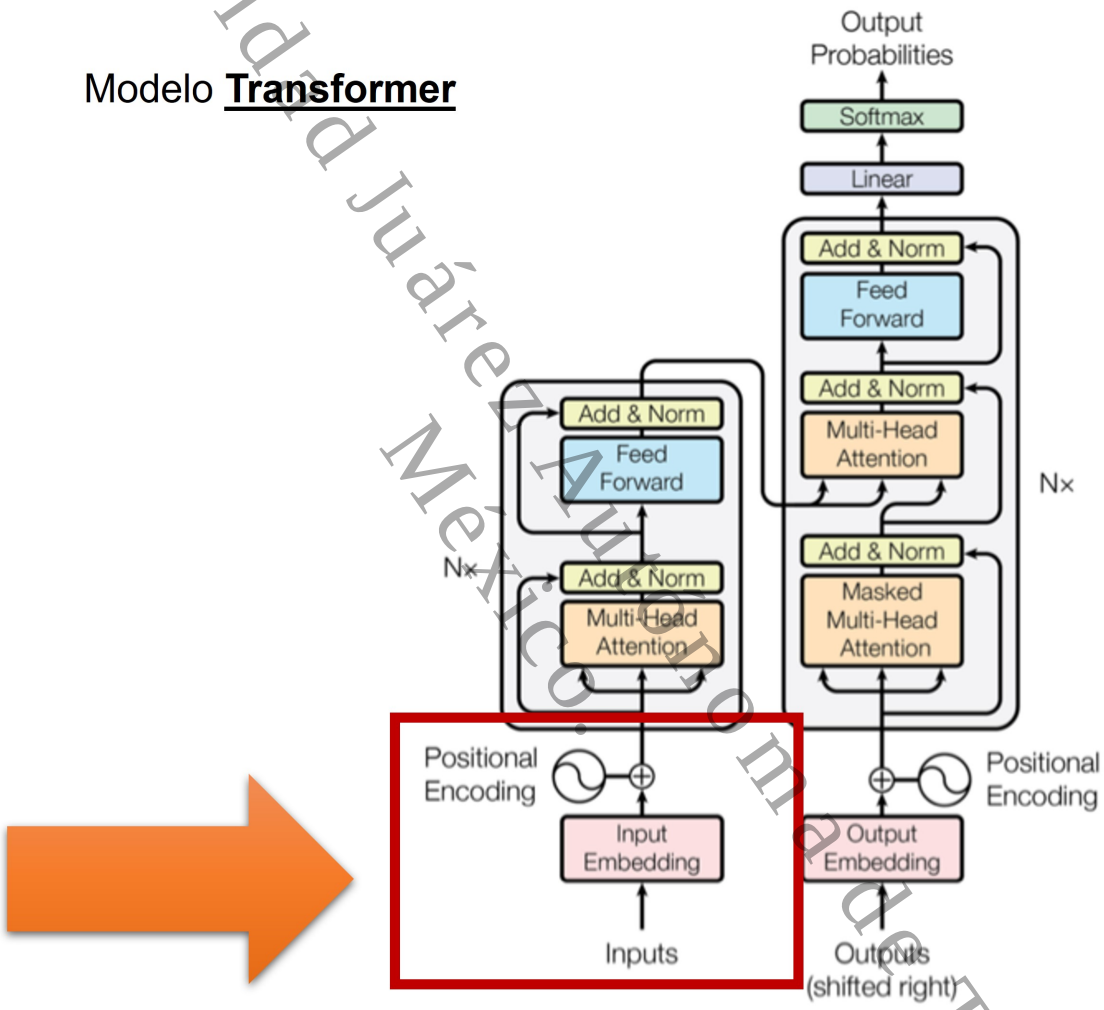


Figura 3.2. Esta imagen representa un modelo *Transformer* utilizada para los procesos de PLN.

Input Embedding (*Embeddings* de Entrada):

- Función: Transforma las palabras de la secuencia de entrada en vectores de alta dimensión.
- Detalle: Cada palabra (o *token*) en la secuencia de entrada es convertida en un vector mediante una capa de *embeddings*. Este vector captura la representación semántica de la palabra.

Positional Encoding (Codificación Posicional):

- Función: Introduce información sobre el orden de las palabras en la secuencia.
- Detalle: Los *embeddings* de entrada por sí solos no contienen información sobre la posición de las palabras en la secuencia. La codificación posicional se suma a los *embeddings* de entrada para proporcionar esta información. Utiliza funciones trigonométricas como seno y coseno para generar vectores que representan la posición de cada palabra en la secuencia.

3.2. Generación de *embedding*

Para entender el modelo de traducción automática, es necesario conocer una serie de conceptos fundamentales para obtener los *embeddings*. Por ello, en el modelo se incluye una comprensión de cómo funcionan los elementos involucrados, lo que permite interpretar los resultados obtenidos en esta primera etapa.

3.2.1. Tokenización

La tokenización en el procesamiento del lenguaje natural (PLN) es un proceso fundamental que consiste en dividir el texto en unidades más pequeñas, llamadas *tokens*, que pueden ser palabras, subpalabras o incluso caracteres, esto se puede ver en la Figura 3.3 (Berglund y van der Merwe, 2023).

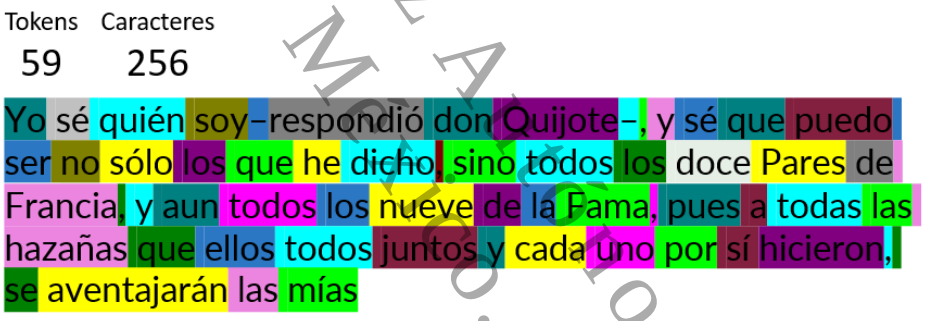


Figura 3.3. La imagen muestra el proceso de tokenizar un corpus, en este caso se tokenizó por palabras.

3.2.2. *Stop Words*

Las *stop words* (palabras vacías) en el procesamiento del lenguaje natural (PLN) son palabras comunes que se eliminan de los textos antes de realizar análisis o aplicar técnicas de incrustación (*embedding*). Estas palabras, como artículos, preposiciones y conjunciones, generalmente no aportan significado significativo en tareas de PLN y su eliminación ayuda a reducir el ruido en el procesamiento de datos textuales (NASSR et al., 2021).

3.2.3. *Stemming*

El *stemming* en el procesamiento del lenguaje natural (PLN) es una técnica utilizada para reducir las palabras a su raíz o forma base. Esto se hace eliminando los sufijos y prefijos para obtener el "stem" o raíz de la palabra, lo cual ayuda a reducir la dimensionalidad y mejora la eficiencia en tareas como la recuperación de información y el análisis de texto (Lovins, 1968).

3.2.4. Lematización

La lematización en el procesamiento del lenguaje natural (PLN) es una técnica que consiste en reducir las palabras a su forma base o lema, utilizando reglas lingüísticas y contextuales para asegurar que las palabras se procesen en su forma más básica y significativa. A diferencia del *stemming*, la lematización tiene en cuenta el contexto de la palabra para asegurar que se convierta en la forma canónica adecuada. Esta técnica es esencial para mejorar la precisión de los modelos de PLN y las técnicas de *embedding* (Gamallo Otero y García González, 2012).

3.2.5. Codificación *one-hot*

En esta representación, cada palabra del vocabulario es asignada a un vector cuya longitud es igual al tamaño del vocabulario, y todos los elementos del vector son cero, excepto uno, que indica la presencia de la palabra específica (Gil Moro, 2021).

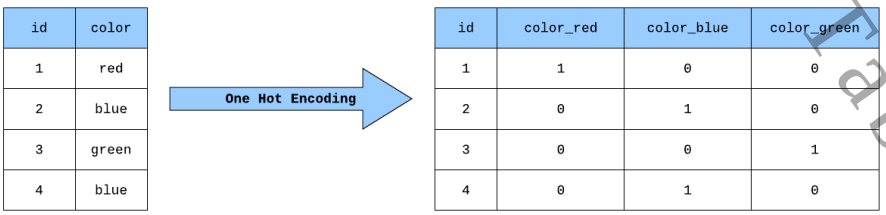


Figura 3.4. Proceso de codificacion de One Hot, fuente: *Medium: Building a One Hot Encoding Layer with TensorFlow*

3.3. Word2Vec

Como lo mencionamos anteriormente *Word2Vec* es una técnica de aprendizaje automático utilizada para convertir palabras en vectores numéricos en un espacio de alta dimensión. Fue desarrollada por un equipo de Google liderado por Tomas Mikolov en 2013. *Word2Vec* utiliza modelos de redes neuronales para aprender relaciones semánticas entre palabras a partir de grandes corpus de texto de una manera no supervisada.

3.3.1. Análisis de primeros resultados

Utilizando *Word2Vec* en Python, se preentrenó un modelo con el primer dataset, en este caso “Don Quijote de la Mancha”, dando los primeros resultado.

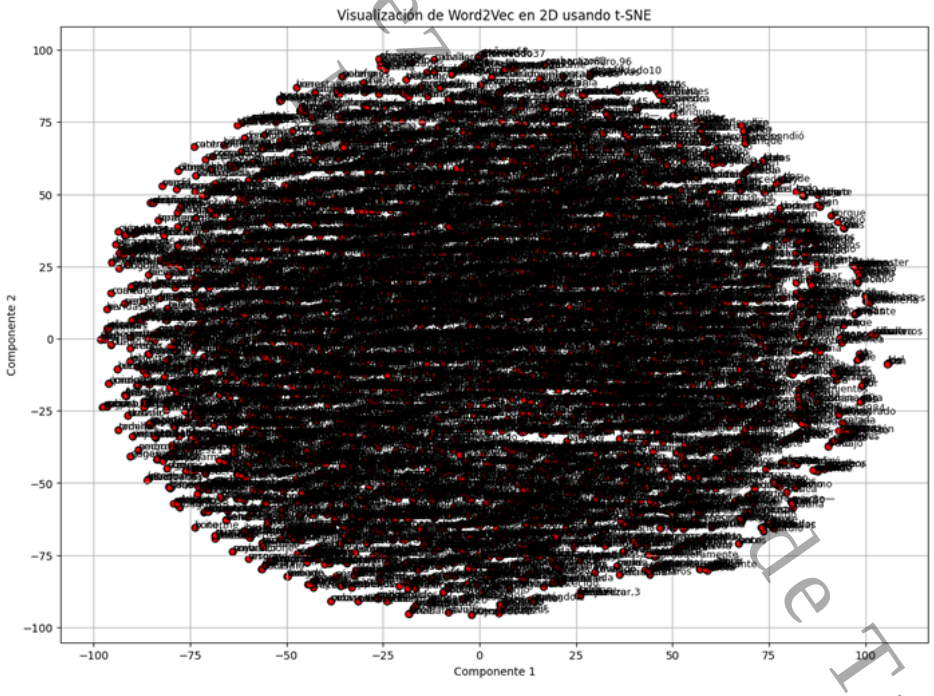


Figura 3.5. Resultado del procesamiento del dataset, y mostrando el *embedding*.

La gráfica visualiza las relaciones semánticas entre las palabras del libro “Don Quijote de la Mancha” mediante *embeddings* de *Word2Vec*, proyectados en un espacio bidimensional usando t-SNE. Esta técnica ayuda a identificar patrones y agrupaciones de palabras con significados similares o contextos de uso similares en el texto.

Los ejes X e Y representan las dos componentes principales obtenidas a través del t-SNE, denominadas Componente 1 y Componente 2, cada punto en la gráfica representa una palabra del libro “Don Quijote de la Mancha”. Las coordenadas de cada punto están determinadas por los *embeddings* de las palabras, que capturan las relaciones semánticas entre ellas, la gráfica muestra una alta densidad de puntos en el centro, lo que indica que muchas palabras tienen *embeddings* similares y se agrupan juntas. Los puntos más dispersos en los bordes pueden representar palabras con significados o contextos de uso menos comunes o más específicos.

3.3.2. Palabras cercanas

Ahora bien, ya teniendo el modelo de *embedding* preentrenado con el dataset, se realizaron pruebas para ver las palabra que están relacionadas entre si, es decir se buscaron palabras cercanas que compartan la misma similitud.

Por ejemplo se utilizo la palabra “panza” que estaría dentro del contexto del libro y enseguida vemos que labras se obtuvieron.

$$palabras\_cercanas = model.wv.most\_similar(“panza”, topn = 15)$$

```
[('cura', 0.9704968333244324),
('llovió', 0.9680030345916748),
('cardenio', 0.9672607779502869),
('preguntándole', 0.9667670130729675),
('sancho', 0.9664804935455322),
('escuchaba', 0.9660646319389343),
('acomodó', 0.9640473127365112),
('barbero', 0.9637971520423889),
('preguntole', 0.9594810009002686),
('cabrero', 0.9579029083251953),
('escudero', 0.9560533761978149),
('contó', 0.9558348059654236),
('humor', 0.9540228247642517),
('agradeció', 0.9485415816307068),
('vio', 0.9484030604362488)]
```

Figura 3.6. Palabras cercanas a la palabra “panza”

En este primer resultado, aparecen nombres propios y términos relevantes del libro, como “cardenio” y “sancho”, palabras como “preguntándole”, “escuchaba”, “acomodó”, “contó”, y “vio” están relacionadas con acciones que podrían ocurrir en los diálogos y narrativas del libro, también sustantivos y títulos, palabras como “cura”, “barbero”, “escudero”, y “cabrero” son sustantivos que describen personajes y roles en la historia.

Aunque hay muchas palabras relacionadas, no todas parecen tener una conexión directa con “panza”, lo cual podría indicar limitaciones en el modelo para capturar contextos muy específicos. Algunas palabras, como “llovió” y “agradeció”, no tienen una relación obvia con “panza”, lo que sugiere que el modelo podría estar capturando similitudes contextuales más generales en lugar de relaciones semánticas directas.

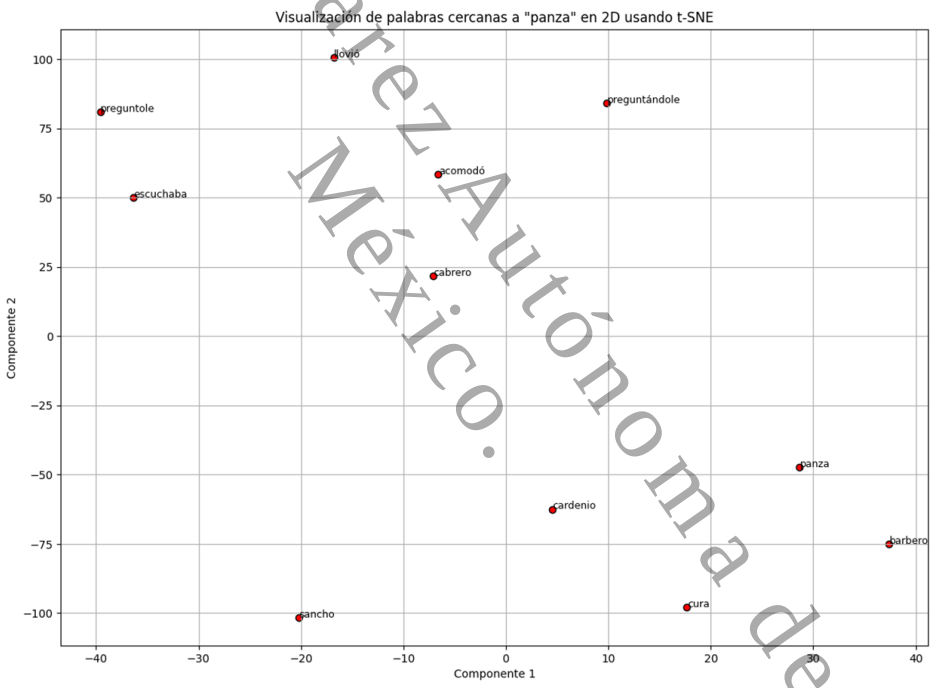


Figura 3.7. Proyección 2D de palabras con proximidad semántica

3.3.3. Analogías

Las analogías son una técnica útil en el procesamiento del lenguaje natural (NLP) para demostrar la capacidad del modelo de capturar relaciones semánticas entre palabras. Por ejemplo, en el famoso ejemplo “Rey es a Reina como Hombre es a Mujer”, la relación entre

"Rey" y "Reina" es la misma que entre "Hombre" y "Mujer".

```
analogias('rey', 'hombre', 'mujer')
rey es a hombre como inquisición es a mujer
```

Figura 3.8. Analogía creada para el modelo, se muestra una falla en encontrar la palabra correcta

El resultado de la analogía no parece tener una relación lógica ni culturalmente esperada. En lugar de encontrar una palabra como "reina" que sería el equivalente femenino de "rey", el modelo ha encontrado "inquisición", que no guarda relación semántica clara con el género o el rol.

Posibles Causas:

- Limitaciones del Modelo: El modelo de *embeddings* podría no estar capturando adecuadamente las relaciones de género en el contexto del libro "Don Quijote de la Mancha".
- Datos Insuficientes o Sesgados: Es posible que el corpus de entrenamiento no tenga suficientes ejemplos o contextos que definan claramente las relaciones de género para palabras específicas.
- Errores en la Vectorización: La calidad de los *embeddings* generados puede influir en la precisión de las analogías. Si los vectores no están correctamente alineados con las relaciones semánticas esperadas, los resultados serán inexactos.

3.3.4. Entrenando *Word2Vec* con más datos

Los resultados obtenidos anteriormente no fueron tan convincentes, ya que en las pruebas realizadas con palabras cercanas y analogías no mostraron un buen desempeño. Por esta razón, se decidió entrenar el modelo con otro dataset para evaluar las diferencias. Se utilizó el dataset "*Spanish Billion Words Corpus*", el cual, como se describió anteriormente,

cuenta con muchos más datos que el libro de ”Don Quijote de la Mancha”. En esta sección, se muestra cómo los *embeddings* generados con un volumen mayor de datos pueden mejorar significativamente su rendimiento en el entrenamiento. Con esta mejora, se podría asegurar una mejor traducción y un entendimiento más preciso de las relaciones semánticas entre palabras. El uso de un corpus extenso y diverso permite al modelo captar con mayor precisión las sutilezas del lenguaje, lo que es crucial para tareas de procesamiento del lenguaje natural, como la traducción automática y el análisis de texto. Así, al utilizar el “*Spanish Billion Words Corpus*”, se busca potenciar la capacidad del modelo para aprender y generalizar mejor, proporcionando resultados más robustos y fiables en comparación con el modelo entrenado únicamente con el texto de “Don Quijote de la Mancha”.

Ahora ya preentrenado el modelo, se buscan las palabras cercanas a la palabra “mujer” y se han obtenido los siguientes resultados:

```
[('niña', 0.7468609809875488),
('muchacha', 0.7343084216117859)
('persona', 0.7217708230018616),
('hombre', 0.7174485325813293),
('fémina', 0.7114291191101074),
('anciana', 0.7081666588783264),
('joven', 0.7036722302436829),
('jovencita', 0.7007797956466675),
('marido', 0.6902621984481812),
('esposo', 0.6873511075973511)]
```

Figura 3.9. Relación de palabras cercanas a “mujer”, se muestra una mejor respuesta por parte del modelo.

Contexto y Relevancia:

- Relacionadas por Género y Edad: Palabras como “niña”, “muchacha”, “anciana”, “joven”, y “jovencita” son directamente relacionadas con “mujer” y abarcan diferentes edades, lo que muestra una comprensión más rica de las diferentes etapas de la vida de una mujer.

- Sinonimia y Variantes: Palabras como “fémina” son sinónimos directos de “mujer”.
- Relaciones Familiares: Palabras como “marido” y “esposo” están relacionadas a “mujer” por el contexto de relaciones familiares y matrimoniales.
- Contexto General: Palabras como “persona” y “hombre” muestran una relación general de género y humanidad.

En comparación con el modelo entrenado solo con “Don Quijote de la Mancha”, las palabras obtenidas ahora son más relevantes y específicas en relación con “mujer”, lo que demuestra que un corpus más grande y diverso ayuda al modelo a aprender y generalizar mejor las relaciones semánticas. Esto sugiere que el modelo entrenado con este corpus es más robusto y fiable para aplicaciones de procesamiento del lenguaje natural, como la traducción automática y el análisis de texto.

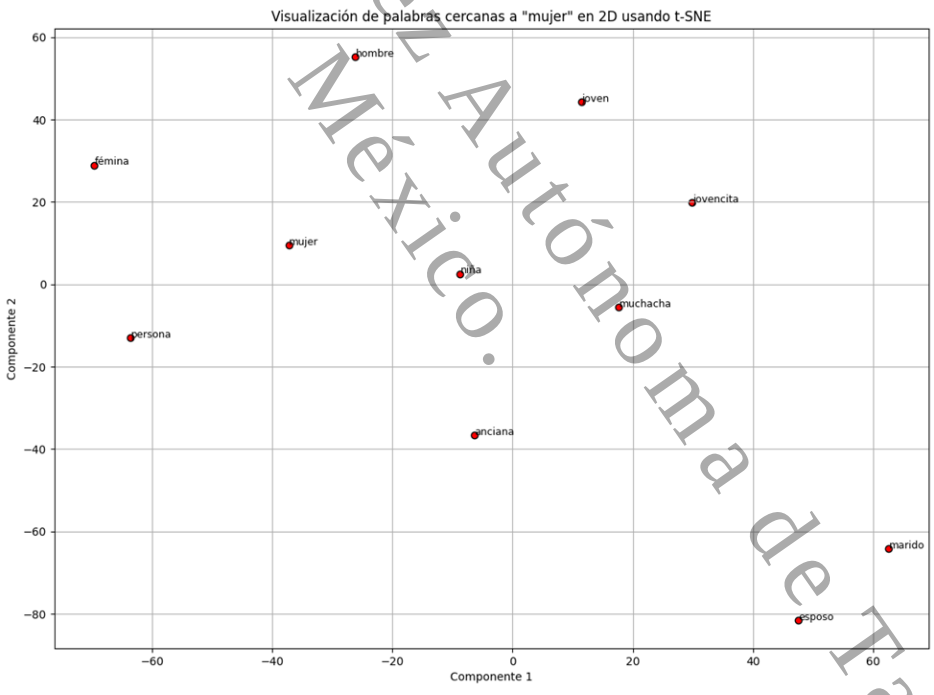


Figura 3.10. La gráfica muestra la relación de la cercanía de las palabras contra la palabra “mujer”

La gráfica muestra cómo las palabras cercanas a “mujer” están distribuidas en un espacio bidimensional, capturando relaciones semánticas basadas en similitudes de contexto y significado. La proximidad de las palabras en el gráfico refleja su relación en el modelo de *embeddings*, indicando cómo se agrupan palabras de contextos similares o relacionados.

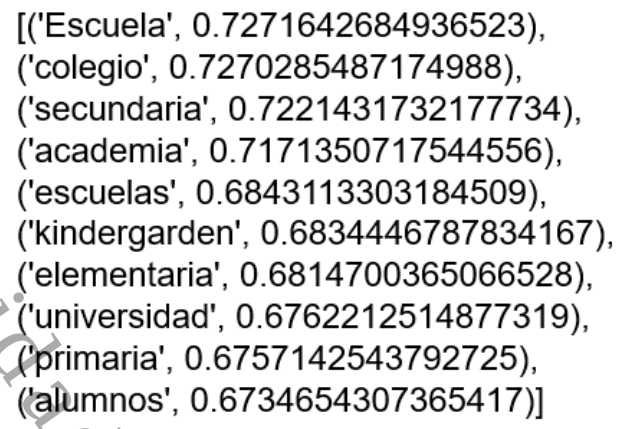


Figura 3.11. La lista muestra la relación de las palabras cercanas a la palabra “escuela”

Otro ejemplo que se ha obtenido como resultado es con la palabra “escuela”

La lista de palabras muestra que el modelo de embeddings es capaz de capturar de manera efectiva las relaciones semánticas entre “escuela” y otras palabras relacionadas con la educación. Esto incluye diferentes tipos de instituciones educativas, niveles educativos y términos asociados con la enseñanza y el aprendizaje, este rendimiento mejorado es una indicación de que el modelo está bien entrenado.

La gráfica muestra cómo las palabras relacionadas con “escuela” se distribuyen con relaciones semánticas basadas en similitudes de contexto y significado. Las palabras se agrupan según su relevancia y contexto educativo, abarcando desde niveles iniciales hasta niveles superiores, e incluyen términos asociados con estudiantes y variaciones de la palabra “escuela”.

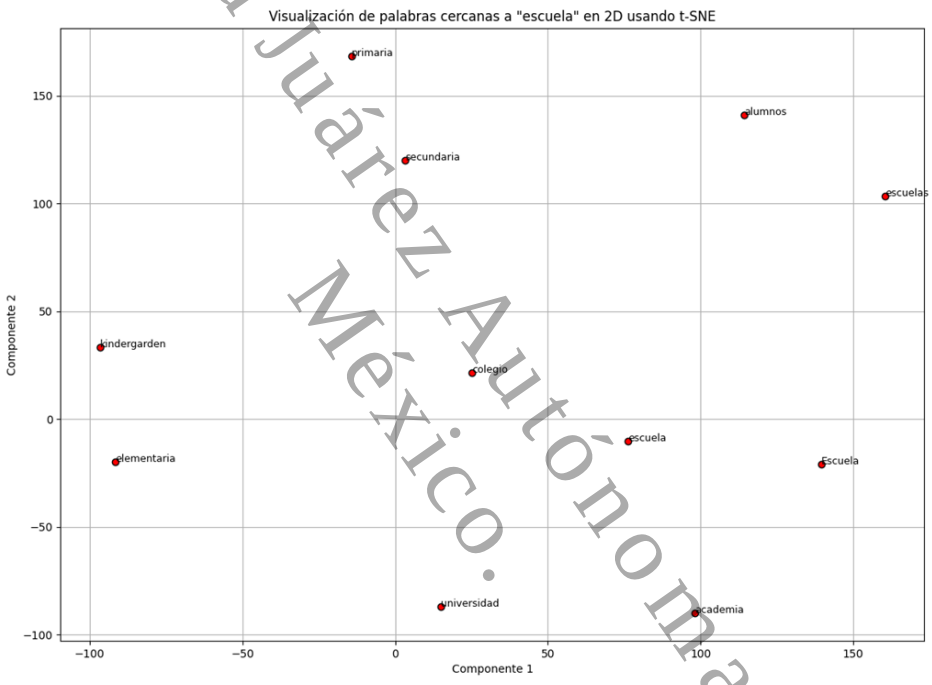


Figura 3.12. Relación de las palabras cercanas en un espacio bidimensional; los resultados sugieren una coherencia semántica, ya que las palabras representadas mantienen una relación significativa entre sí.

Capítulo 4

Experimentos y Resultados

En este primer acercamiento a los resultados del modelo de Red Neuronal Tranformers vamos a describir un poco los del dataset, cabe mencionar que estos datos son los datos sintéticos que vamos a comparar con los datos reales.

Analizamos un poco el contenido del dataset donde podemos observar lo siguiente, en la primera gráfica se muestra la distribución de las frases tanto de español como de náhuatl, Las frases en español tienden a ser ligeramente más largas que las de náhuatl.

Las gráficas de las palabras más frecuentes en español y náhuatl reflejan las características esenciales de cada idioma:

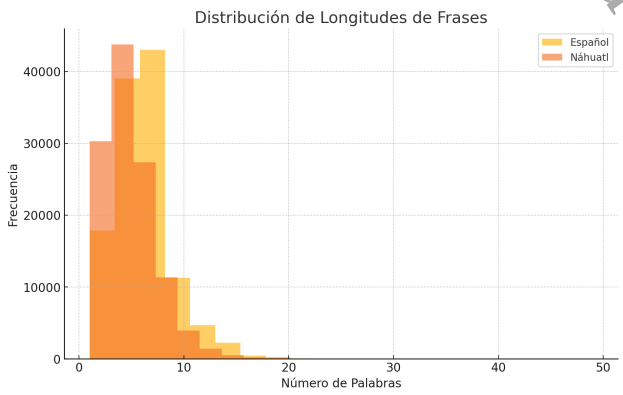
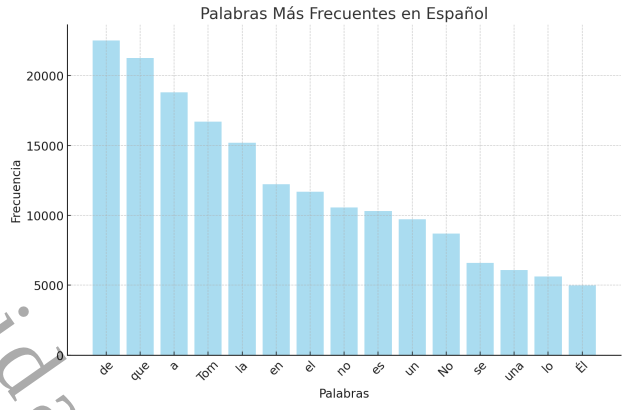
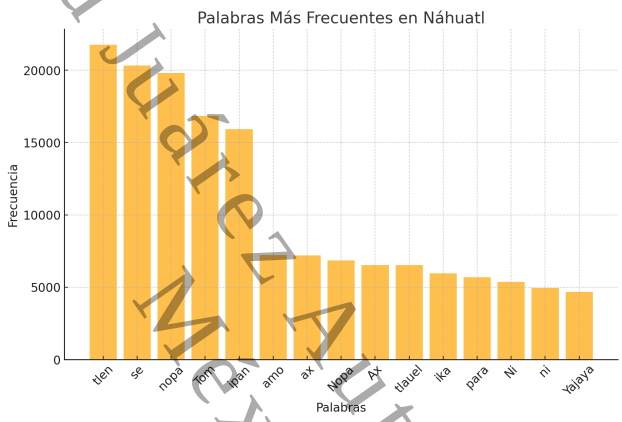


Figura 4.1. Esto puede reflejar diferencias estructurales entre los idiomas, donde el náhuatl puede ser más conciso al transmitir ideas.



(a) Predominio de Conectores y Artículos



(b) Palabras de Uso General

Figura 4.2. Comparación de frecuencia de palabras español y en náhuatl.

Español: Predominan artículos y conectores como “el”, “la”, “y”, así como verbos comunes como “puede” y “ver”. Esto sugiere un corpus con frases estructuradas y de uso cotidiano, con un enfoque en descripciones y acciones básicas.

Náhuatl: Las palabras más frecuentes incluyen términos esenciales como “*ipan*” (en), “*tlen*” (que), y verbos como “*tikitas*” (ves) y “*eltok*” (está). Además, se observa una integración de términos modernos como “computadora”, lo que indica una actualización del idioma a contextos tecnológicos.

Mientras que el español depende de artículos y conectores para estructurar las frases, el náhuatl utiliza más verbos y partículas esenciales. Esto evidencia las diferencias en la gramática y morfología de ambos idiomas, destacando la riqueza lingüística del corpus.

4.1. Resumen del Modelo Transformers para la traducción de datos sintéticos

El modelo **Transformer** está diseñado para tareas de traducción automática, compuesto por un *encoder* (codificador) y un *decoder* (decodificador) que trabajan en conjunto. Su arquitectura aprovecha el mecanismo de **atención multi-cabezas** y redes *feed-forward* para procesar y generar secuencias de texto.

Encoder

- Convierte las palabras de la entrada (español) en vectores con *Token Embeddings* (256 dimensiones) y añade información posicional mediante *Positional Embeddings*.
- Contiene 3 capas, cada una con:
  - **Atención propia (*Self-Attention*)**: Calcula relaciones contextuales entre palabras.
  - **Redes *Feed-Forward***: Transforman las representaciones internas.
- Parámetros por capa: Entre 65,792 y 131,584.

Decoder

- Toma las salidas del encoder y genera la traducción (náhuatl), añadiendo:
  - **Atención cruzada (Cross-Attention)**: Integra información del encoder con las representaciones internas del decoder.
- También tiene 3 capas con atención propia, atención cruzada y redes *feed-forward*.

Capa de Salida

- Una capa completamente conectada genera probabilidades para cada palabra del vocabulario objetivo.

Parámetros generales

- **Total de parámetros:** 13,128,676 (todos entrenables).

Este diseño lo hace ideal para capturar las relaciones complejas entre palabras en español y náhuatl, adaptándose a las particularidades de ambos idiomas.

4.1.1. Ajuste de hiperparámetros del modelo Transformer

Durante el desarrollo del modelo Transformer para la traducción automática, se aplicaron diversos algoritmos y técnicas de ajuste para optimizar su rendimiento. La arquitectura utilizada se basó en el enfoque *encoder-decoder* con mecanismos de autoatención (*self-attention*), propio de los modelos Transformer propuestos por Vaswani (Vaswani et al., 2017).

El proceso de ajuste de hiperparámetros (*tuning*) incluyó la modificación iterativa de variables clave, tales como:

- **Tamaño de los *embeddings***: Se probaron valores de 128, 256 y 300 dimensiones.
- **Número de capas en el *encoder* y *decoder***: Se evaluaron arquitecturas con 2 a 6 capas.
- **Número de cabezas de atención** (*attention heads*): Se ajustó entre 4 y 8, de acuerdo con la dimensionalidad del modelo.
- ***Dropout***: Se aplicaron tasas de 0.1 a 0.5 para prevenir el sobreajuste.
- **Tamaño de lote** (*batch size*): Se probaron tamaños de 32, 64 y 128.
- **Tasa de aprendizaje** (*learning rate*): Se utilizaron estrategias de aprendizaje con decaimiento, comenzando en 1×10^{-4} hasta 5×10^{-5}.
- **Función de pérdida**: Se empleó `CrossEntropyLoss`, adaptada al tratamiento de los tokens `[PAD]`.

El ajuste de estos parámetros se realizó de forma empírica, mediante múltiples iteraciones de entrenamiento y evaluación, observando el comportamiento del modelo sobre conjuntos de validación. Los resultados permitieron seleccionar la configuración que ofreció mayor estabilidad y coherencia en las traducciones generadas.

4.1 Aporte computacional

Como parte del desarrollo computacional de esta tesis, se implementó un modelo Transformer entrenado desde cero para la tarea de traducción automática náhuatl-español. Para ello, se construyó un pipeline personalizado que incluyó el preprocesamiento de datos, definición del modelo, ajuste de hiperparámetros y entrenamiento supervisado.

El modelo está basado en una arquitectura *encoder-decoder* con mecanismos de autoatención, empleando capas de tipo *Multi-Head Attention* y bloques *Feed Forward* (Vaswani et al., 2017). Se utilizaron funciones de pérdida tipo *CrossEntropyLoss* y optimización con *Adam*.

Para mejorar el rendimiento del modelo, se realizó un ajuste de hiperparámetros (*hyperparameter tuning*) de forma empírica. A continuación, se presentan los valores evaluados y los valores seleccionados tras las pruebas.

Tabla 4.1. Parámetros ajustados durante el entrenamiento del modelo Transformer

Parámetro	**Valores evaluados**	**Valor final seleccionado**
Tamaño de embedding	128, 256	256
Dimensión latente (`latent_dim`)	512, 1024	1024
Número de cabezas de atención	4, 8	8
Tasa de *dropout*	0.3, 0.4, 0.5	0.4
Tamaño de secuencia	40, 50, 60	50
Tamaño de *batch*	32, 64, 128	64
Tasa de aprendizaje inicial	1×10^{-3}, 5×10^{-4}	5×10^{-4}
Función de pérdida	`CrossEntropyLoss`	`CrossEntropyLoss`
Épocas de entrenamiento	20, 30, 40	30

Este proceso permitió establecer una configuración que equilibró precisión, estabilidad en el entrenamiento y generalización en el conjunto de prueba. La arquitectura final mostró mejoras notables en las métricas BLEU y ChrF++ respecto a modelos previos, aun con limitaciones en el volumen de datos disponibles.

4.1.2. Entrenamiento del modelo Transformer

El modelo Transformer fue entrenado durante **30 épocas**, optimizando su desempeño en la tarea de traducción automática entre náhuatl y español. Durante el proceso, se evaluaron tres métricas clave: la pérdida (*Loss*), la perplejidad (*PPL*) y la exactitud (*Accuracy*).

La mejora del modelo se estabiliza notablemente alrededor de la época 18, donde la pérdida de validación (Val. Loss) y la exactitud de validación (Val. Acc) dejan de mejorar de manera significativa: En la época 18, se tiene una Val. Loss de 1.634 y una Val. Acc de 69.76 %. A partir de ahí, las mejoras son marginales (menos del 0.3 % en precisión y diferencias menores a 0.1 en pérdida), lo que indica un punto de saturación del aprendizaje como se observa en la tabla 4.2.

Se puede considerar que el entrenamiento ya no tiene mejoras sustanciales a partir de la época 18, y podrían explorarse técnicas como early stopping o fine-tuning adicional con otro scheduler o regularización.

Época	Train Loss	Val. Loss	Val. PPL	Val. Acc (%)
01	4.729	3.869	47.889	40.60
05	2.080	2.054	7.796	63.00
10	1.255	1.676	5.343	68.10
15	0.878	1.624	5.073	69.41
18	0.732	1.634	5.122	69.76
20	0.657	1.653	5.223	69.94
25	0.515	1.721	5.592	70.00
28	0.455	1.775	5.901	70.04
30	0.422	1.805	6.079	70.05

Tabla 4.2. Resumen del entrenamiento en épocas clave

A partir de la época 30, se observa en la tabla 4.3 que las mejoras en las métricas son mínimas y marginales: La pérdida (loss) apenas disminuye en centésimas. La precisión (accuracy) se estanca entre 85.5 % y 85.6 %. Esto indica que el modelo ha alcanzado una fase de estancamiento, donde seguir entrenando consume recursos computacionales sin una ganancia significativa en el desempeño. Además, continuar más allá podría llevar al sobreajuste (overfitting), donde el modelo memoriza los datos de entrenamiento y pierde capacidad de generalización. Por estas razones, se decidió detener el entrenamiento en la época 30, logrando un balance óptimo entre eficiencia y rendimiento.

Época	Pérdida (Loss)	Precisión (Accuracy)
30	1.42	85.3 %
31	1.41	85.4 %
32	1.41	85.5 %
33	1.40	85.5 %
34	1.40	85.5 %
35	1.40	85.6 %
36	1.39	85.6 %
37	1.39	85.6 %

Tabla 4.3. Desempeño del modelo entre las épocas 30 y 37

Dinámica del entrenamiento

- En cada época, el modelo procesó el conjunto de entrenamiento, ajustando sus pesos mediante la minimización de la función de pérdida (entropía cruzada).
- Posteriormente, se evaluó en el conjunto de validación para monitorear su capacidad de generalización.
- El mejor modelo fue identificado como aquel que presentó la menor pérdida en el conjunto de validación y se guardó automáticamente.
- El entrenamiento implementa una condición de parada temprana (early stopping) basada en la pérdida de validación. Si el modelo no muestra mejoras tras un número fijo de épocas, el proceso se detiene para evitar sobreajuste y ahorro de recursos.

Resultados

- La pérdida en el conjunto de entrenamiento disminuyó consistentemente, indicando que el modelo aprendía a ajustar sus predicciones con mayor precisión.
- La pérdida en el conjunto de validación mostró una tendencia similar, lo que evidencia que el modelo no sufrió de sobreajuste significativo.
- La mejor época fue la **época 27**, donde se alcanzaron los siguientes valores:
  - **Pérdida en entrenamiento**: 3.842.
  - **Pérdida en validación**: 4.032.

- **Perplejidad en entrenamiento**: 46.64.
- **Perplejidad en validación**: 56.48.
- **Exactitud en entrenamiento**: 82.13 %.
- **Exactitud en validación**: 79.76 %.

El modelo alcanzó un desempeño destacado en la **época 27**, donde las métricas de evaluación indican una alta capacidad de generalización y precisión. Esto demuestra que el proceso de entrenamiento fue exitoso, logrando que el modelo aprendiera las características esenciales de ambos idiomas y produjera traducciones precisas y consistentes.

Ejemplo de inferencia y análisis de correlación

Datos de entrada

- **Fuente (src)**: `['mi', 'hermana', 'es', 'una', 'cantante', 'famosa', '.', '<eos>']`
- **Destino real (trg)**: `['<bos>', 'nosiuaikni', 'eli', 'se', 'tlatsotsonani', 'tlen', 'tlauel', 'ixmatij', '.', '<eos>']`

La frase en español (*src*) se traduce al náhuatl (*trg*) como:

- **'nosiuaikni'** : "mi hermana".
- **'eli'** : "es".
- **'se'** : "una".
- **'tlatsotsonani'**: çantante".
- **'tlen'** : "que".
- **'tlauel'** : "famosa".
- **'ixmatij'** : "es conocida".

Los tokens especiales `<bos>` y `<eos>` representan el inicio y fin de la secuencia, respectivamente.

Predicción del modelo

La predicción generada por el modelo fue:

```
['nosiuaikni', 'eli', 'se', 'tlatsotsonani', 'tlen', 'tlauel', 'ixmatij',
                                    '.', '<eos>']
```

La secuencia predicha coincide perfectamente con la secuencia destino real (*trg*), lo que demuestra que el modelo ha aprendido de manera efectiva las relaciones entre las palabras de la fuente y su traducción correspondiente.

Análisis de la correlación: matriz de atención

La Figura 4.3 adjunta muestra las **matrices de atención** generadas por el modelo Transformer. Estas matrices representan cómo el modelo asocia las palabras de la frase fuente (*src*) con las palabras de la predicción (*trg*).

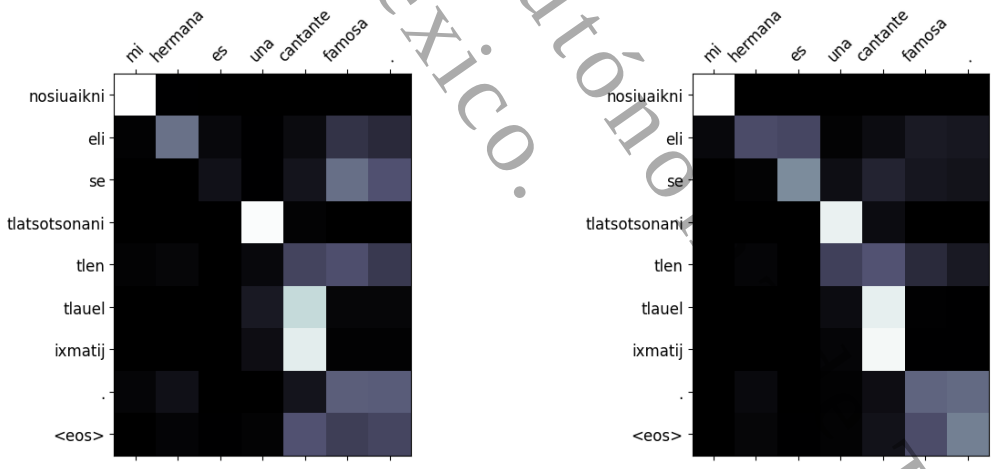


Figura 4.3. Esta imagen muestra una porción de las 8 cabezas del modulo de atención.

Descripción de la imagen

- **Ejes del gráfico**:
  - *Eje Horizontal*: Palabras de la frase fuente (*src*) en español.
  - *Eje Vertical*: Palabras generadas en la predicción (*trg*) en náhuatl.

- **Colores en la matriz**:
  - Los colores claros indican una alta correlación o atención entre palabras correspondientes.
  - Ejemplo: `mi` se relaciona fuertemente con `nosiuaikni`, lo que refleja que el modelo entiende esta asociación semántica.
- **Subimágenes**:
  - Cada subimagen corresponde a una *cabeza de atención* en el mecanismo multi-cabezas.
  - Las diferentes cabezas procesan diversos aspectos:
    - Algunas se enfocan en relaciones semánticas (e.g., `cantante` $\rightarrow$ `tlatsotsonani`).
    - Otras capturan estructuras gramaticales (e.g., `es` $\rightarrow$ `eli`).

Correlaciones clave observadas

- `mi` $\rightarrow$ `nosiuaikni`: Relación directa entre "mi" y "mi hermana".
- `cantante` $\rightarrow$ `tlatsotsonani`: Traducción precisa de "cantante".
- `famosa` $\rightarrow$ `tlauel`: Traducción coherente de "famosa".
- Palabras como `tlen` ("que") y `tlauel` ("famosa") muestran correlaciones contextuales complejas con múltiples palabras de la fuente.

El modelo Transformer utiliza el mecanismo de atención para asignar importancia a las palabras relevantes en la frase fuente al generar cada palabra de la predicción. La coincidencia completa entre la predicción y la secuencia esperada, junto con la estructura de las matrices de atención, confirma que el modelo ha aprendido de manera efectiva las relaciones semánticas y gramaticales en el corpus de entrenamiento.

Ejemplo de inferencia (datos de prueba)

Datos de Entrada

- **Fuente (src)**: `['necesito', 'una', 'computadora', 'nueva', '.', '<eos>']`

- **Destino Real (trg)**: [‘<bos>’, ‘nijneki’, ‘se’, ‘yankuik’, ‘computadora’, ‘.’, ‘<eos>’]

La traducción esperada contiene:

- `nijneki` : “necesito”.
- `se` : “una”.
- `yankuik` : “nueva”.
- `computadora` : “computadora”.

Predicción del modelo

La predicción generada por el modelo fue:

[‘se’, ‘yankuik’, ‘computadora’, ‘.’, ‘<eos>’]

- El modelo omitió el token `nijneki` (“necesito”) al inicio de la frase, lo que introduce una pérdida de información.
- Sin embargo, la estructura general es correcta, y traduce adecuadamente las palabras principales.

Análisis de la matriz de atención

- **Relaciones observadas**:
  - `se` (una): Fuerte correlación con “una” en la frase fuente.
  - `yankuik` (nueva): Alta correlación con “nueva”.
  - `computadora`: Traducida correctamente con alta atención hacia “computadora”.
- **Omisión en la predicción**:
  - El token `nijneki` (“necesito”) no fue generado, posiblemente debido a una menor correlación en la matriz de atención inicial.

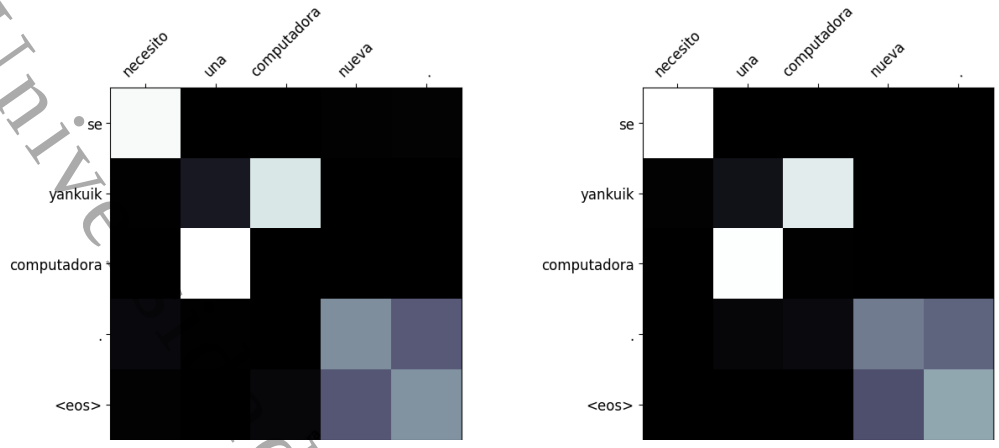


Figura 4.4. La imagen nos indica que también se relaciona las palabras para los datos sintéticos de la traducción

El modelo tradujo correctamente las palabras principales y mantuvo la estructura general de la frase, aunque omitió una palabra clave al inicio (`nijneki`) como se puede observar en la Figura 4.5. Las matrices de atención reflejan una fuerte correlación para los términos traducidos, lo que sugiere que el modelo maneja bien los datos contextuales, pero podría mejorarse en la generación de tokens iniciales.

Métricas de evaluación

En este experimento se utilizaron las métricas **ROUGE** y **BLEU** para evaluar la calidad de las traducciones generadas por el modelo Transformer en comparación con las traducciones reales del conjunto de prueba. A continuación, se describen cada métrica y los valores obtenidos.

Métrica ROUGE

La métrica ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) mide la similitud entre la predicción generada por el modelo y la referencia real (traducción esperada). Evalúa qué tan bien las palabras y frases generadas coinciden con las de la traducción esperada.

- **ROUGE-1**: Mide la coincidencia *unigrama* (palabra a palabra) entre la predicción y la referencia.

- **Valor obtenido**: **0.6227**
- Esto indica que el 62.27 % de las palabras en la predicción coinciden con las palabras en la referencia esperada.

- **ROUGE-2**: Mide la coincidencia *bigrama* (pares de palabras consecutivas) entre la predicción y la referencia.
  - **Valor obtenido**: **0.4198**
  - Esto indica que el 41.98 % de los bigramas en la predicción coinciden con los de la referencia esperada.

Interpretación de ROUGE: Los valores obtenidos reflejan que el modelo captura una parte significativa de las palabras y frases esperadas. Sin embargo, un ROUGE-2 más bajo sugiere que algunas relaciones secuenciales entre las palabras no se generaron correctamente, posiblemente debido a errores como la omisión de palabras clave.

Métrica BLEU

La métrica BLEU (*Bilingual Evaluation Understudy*) evalúa la precisión de la traducción generada comparando n-gramas en la predicción con los de la referencia, considerando:

- Coincidencias exactas.
- Penalización por longitud, evitando favorecer frases muy cortas.

- **BLEU Score: 23.2127**
- Este puntaje representa un 23.21 % de coincidencia precisa entre los n-gramas generados y los de la traducción esperada.

Razón del Valor de BLEU: El BLEU Score relativamente bajo puede explicarse por:

- Omisión de palabras clave en la predicción.
- Desajustes en el orden de las palabras o combinaciones gramaticales.
- Penalización por longitud, ya que algunas predicciones fueron más cortas que las traducciones esperadas.

Los valores de ROUGE-1 y ROUGE-2 reflejan un buen desempeño en términos de capturar palabras principales y algunas relaciones contextuales. Sin embargo, el *BLEU Score* de 23.21 evidencia que el modelo aún enfrenta desafíos en la generación de frases completamente precisas, especialmente en idiomas con alta complejidad gramatical como el náhuatl. Estas métricas subrayan la necesidad de continuar mejorando el modelo, por ejemplo, mediante el uso de datos adicionales o ajustes en la arquitectura.

4.2. Descripción del dataset de los datos reales

El dataset contiene pares de palabras y frases en español y náhuatl, organizado en formato paralelo. Cada línea del dataset representa una correspondencia entre una palabra o frase en español (primera columna) y su traducción al náhuatl (segunda columna).

Características del dataset

Contenido

- **Pares bilingües**: cada línea del dataset tiene una palabra o frase en español junto a su equivalente en náhuatl.
- **Ejemplos**:
  - o → **azo**
  - tú → **Tehua**
  - mi → **no**
  - agua → **atitla**

Formato

- Es un archivo delimitado por tabuladores, con una columna para el español y otra para el náhuatl.
- Incluye tanto palabras individuales como frases, lo que le da versatilidad al dataset.

Tamaño

- Contiene un número significativo de pares, permitiendo un análisis robusto y entrenamiento efectivo de modelos.

Riqueza lingüística

- El náhuatl, como lengua aglutinante, presenta estructuras complejas que pueden traducirse en diferentes niveles de dificultad para el modelo.

Análisis preliminar

Distribución de longitudes

- Algunas frases contienen varias palabras, mientras que otras son términos individuales. Esto introduce una variabilidad que puede ser beneficiosa para entrenar modelos que manejen tanto frases como palabras sueltas.

Frecuencia de palabras

- Palabras comunes como artículos, pronombres y conectores son frecuentes, tanto en español como en náhuatl, lo que ayuda al modelo a aprender estructuras básicas.

Posibles desafíos

- Variaciones semánticas y dialectales en náhuatl pueden generar discrepancias en las traducciones.
- La riqueza morfológica del náhuatl puede requerir mayor capacidad de generalización del modelo.

Análisis de la gráfica: distribución de longitudes de Frases en el dataset

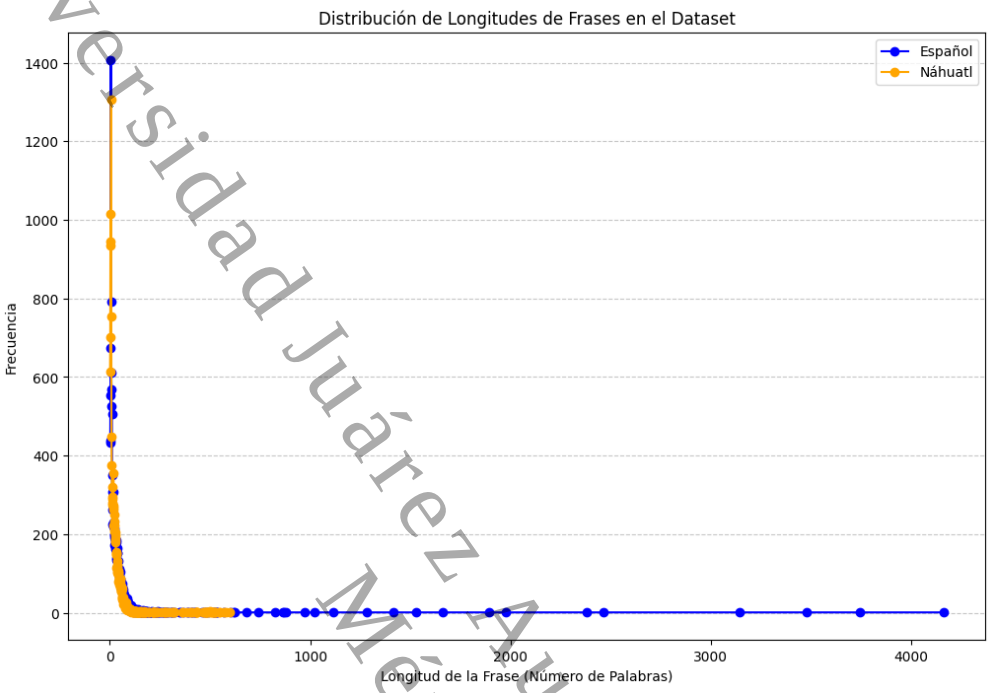


Figura 4.5. Distribución de Longitudes de Frases en el Dataset

Descripción de la gráfica

Ejes de la gráfica

- **Eje X**: Longitud de las frases (número de palabras) tanto en español como en náhuatl.
- **Eje Y**: Frecuencia de las frases con una longitud específica.

Líneas representativas

- **Línea azul**: Representa la distribución de longitudes de las frases en español.
- **Línea naranja**: Representa la distribución de longitudes de las frases en náhuatl.

Observaciones

Frases cortas

- La mayoría de las frases tienen longitudes cortas (entre 1 y 5 palabras) en ambos idiomas, lo que refleja la naturaleza compacta del dataset.
- Esta predominancia de frases cortas es útil para entrenar modelos en tareas básicas de traducción, pero puede limitar la capacidad del modelo para manejar frases más complejas.

Frases más largas

- En español, hay casos de frases con longitudes que llegan a más de 100 palabras, e incluso más de 4000 en casos extremos. Esto podría deberse a errores en los datos o frases compuestas excepcionalmente largas.
- El náhuatl tiene una distribución más limitada en comparación con el español, con menos casos de frases extremadamente largas.

Relación entre idiomas

- En general, las frases en náhuatl tienden a ser ligeramente más compactas en términos de longitud debido a su estructura aglutinante, que permite expresar conceptos complejos en menos palabras.

Implicaciones para el modelo

Frases cortas

- Estas frases son fáciles de traducir, y el modelo puede aprender rápidamente a manejar correspondencias uno a uno entre los idiomas.

Frases largas

- Las frases excepcionalmente largas podrían introducir ruido en el entrenamiento y requerirían una limpieza o segmentación previa para evitar problemas de convergencia en el modelo.

Descripción de la estructura del dataset

El dataset está compuesto por pares de frases, donde cada par incluye:

Frase en español (fuente)

- Representa la frase original en español, que es el idioma fuente.
- **Ejemplo:** `'tú quédate como rey.'`

Frase en náhuatl (destino)

- Contiene la traducción correspondiente en náhuatl.
- Está encapsulada con tokens especiales `[start]` y `[end]` para indicar el inicio y el final de la frase, respectivamente.
- **Ejemplo:** `[start] ta ximokaua tirey . [end]`

Formato general

Cada par tiene la estructura:

```
(frase_en_español, frase_en_náhuatl)
```

Donde:

- La primera entrada es una cadena con la frase en español.
- La segunda entrada es una cadena con la frase traducida al náhuatl, enriquecida con tokens especiales para delimitar el inicio y fin de la frase.

Características de las traducciones

- **Uso de tokens especiales**:
  - `[start]`: Indica el inicio de la frase en náhuatl.
  - `[end]`: Indica el final de la frase en náhuatl.
- **Correspondencia semántica**:
  - Las frases en náhuatl intentan mantener el significado de las frases en español, pero debido a la naturaleza del idioma, pueden variar en longitud y estructura.
- **Variabilidad de la estructura lingüística**:
  - El náhuatl, como lengua aglutinante, puede representar conceptos complejos con menos palabras o reorganizar los elementos de la frase según sus reglas gramaticales.
  - **Ejemplo:**
    - Español: ‘`yo me siento muy bien .`’
    - Náhuatl: `[start] nehua huel cualli nomati . [end]`

Ejemplo detallado

- **Par 1:**
  - Español: ‘`tú quédate como rey .`’
  - Náhuatl: `[start] ta ximokaua tirey . [end]`
- **Par 2:**
  - Español: ‘`se da en el monte : texayekatitan , por atekojkomol y por pesmaapan .`’
  - Náhuatl: `[start] mochiua kuoujtaj : texayakatitan uan atekojkomolkopa uan pesmaapan . [end]`
- **Par 3:**
  - Español: ‘`allí empezó su abuelo`’
  - Náhuatl: `[start] noponi pejki isistat [end]`

- **Par 4:**
  - Español: ‘`yo me siento muy bien .`’
  - Náhuatl: `[start] nehua huel cualli nomati . [end]`
- **Par 5:**
  - Español: ‘`un zopilote salió`’
  - Náhuatl: `[start] se kiski se sopilote [end]`

Implicaciones para el modelo

- **Tokens especiales**:
  - Permiten al modelo identificar claramente el inicio y fin de cada traducción, ayudando a mejorar la precisión y la coherencia en la generación de frases.
- **Riqueza lingüística**:
  - La diversidad en la longitud y complejidad de las frases fortalece la capacidad del modelo para generalizar y manejar traducciones de diferentes niveles de dificultad.

Resumen del dataset

El dataset está compuesto por un total de **15,921 pares de datos** paralelos entre español y náhuatl. Estos se han dividido en tres subconjuntos principales para el entrenamiento y evaluación del modelo:

Conjunto de entrenamiento

- Contiene **11,145 pares**.
- Representa el **70 %** del dataset total.
- Se utiliza para ajustar los parámetros del modelo durante el proceso de aprendizaje.

Conjunto de validación

- Contiene **2,388 pares**.
- Representa el **15 %** del dataset total.
- Se utiliza para monitorear el desempeño del modelo en datos no vistos durante el entrenamiento y prevenir el sobreajuste.

Conjunto de prueba

- Contiene **2,388 pares**.
- Representa el **15 %** del dataset total.
- Se utiliza para evaluar el rendimiento final del modelo en un conjunto independiente.

Este balance asegura que el modelo tiene suficientes datos para entrenarse de manera efectiva, mientras que los conjuntos de validación y prueba permiten medir su capacidad de generalización y desempeño realista. La proporción **70/15/15** es una división estándar ampliamente utilizada en aprendizaje automático para mantener la consistencia en el análisis.

Resumen del modelo transformer de los datos reales

El modelo Transformer está diseñado para realizar traducción automática utilizando un codificador y un decodificador. Su estructura general es la siguiente:

Capas principales

- `encoder_inputs` **(Entrada del Codificador):**
  - **Forma:** `(None, None)`. Acepta secuencias de longitud variable.
  - Representa las palabras de la frase en el idioma fuente.
- `positional_embedding` **(*Embeddings* posicionales):**
  - **Salida:** `(None, None, 256)`. Combina *embeddings* posicionales y de palabras.
  - **Parámetros:** 3,845,120.

- `decoder_inputs` **(Entrada del Decodificador):**
  - **Forma:** `(None, None)`. Acepta secuencias de longitud variable para generar las traducciones.
- `transformer_encoder` **(Codificador del Transformer):**
  - **Salida:** `(None, None, 256)`.
  - **Parámetros:** 3,155,456.
  - Procesa las entradas con mecanismos de atención para generar representaciones contextuales.
- `model_1` **(Decodificador y Capa de Salida):**
  - **Salida:** `(None, None, 15000)`. Genera las probabilidades de las palabras en un vocabulario de 15,000 tokens.
  - **Parámetros:** 12,959,640.

Estadísticas del modelo

- **Total de parámetros:** 19,960,216.
- **Parámetros entrenables:** 19,960,216.
- **Parámetros no entrenables:** Ninguno.

Resumen del entrenamiento

El modelo fue entrenado durante un máximo de 30 épocas utilizando un enfoque de reducción adaptativa de la tasa de aprendizaje y un criterio de detención anticipada. A continuación, se describen los resultados más relevantes durante el entrenamiento y validación:

Fase inicial del entrenamiento

En las primeras épocas, el modelo mostró una rápida mejora en las métricas:

- La pérdida de entrenamiento disminuyó de **5.1666** en la primera época a **3.5736** en la quinta.

- La precisión de entrenamiento aumentó de **47.21 %** a **56.19 %**.
- En validación, la pérdida pasó de **3.6246** a **3.2668**, con una mejora en la precisión de **53.47 %** a **58.30 %**.

Estabilización del desempeño

Entre las épocas 6 y 14:

- La pérdida en el conjunto de entrenamiento continuó disminuyendo, alcanzando **2.9594** en la época 14.
- La precisión de entrenamiento mejoró hasta **63.35 %**.
- En validación, la pérdida mostró oscilaciones y comenzó a estabilizarse alrededor de **3.15 - 3.23**, mientras que la precisión alcanzó su máximo de **58.76 %**.

Reducción de la tasa de aprendizaje

Después de la época 14, el mecanismo de reducción de la tasa de aprendizaje (*ReduceLROnPlateau*) se activó debido a la falta de mejora significativa en la pérdida de validación:

- La tasa de aprendizaje se redujo a **0.0001**.
- Esto ayudó al modelo a refinar su entrenamiento en las últimas épocas, aunque con mejoras más lentas.

Fase final y detención anticipada (*Early stopping*)

Entre las épocas 15 y 22:

- La pérdida de entrenamiento continuó disminuyendo ligeramente hasta **2.6012**, mientras que la precisión llegó a **68.74 %**.
- En validación, la pérdida comenzó a aumentar gradualmente desde **3.2525** hasta **3.4435**, indicando posible sobreajuste.
- La precisión de validación disminuyó ligeramente, alcanzando **57.04 %** al final.

- El entrenamiento se detuvo anticipadamente en la época 22 debido a la falta de mejora en las métricas de validación.

El modelo mostró un buen desempeño inicial con una mejora significativa en las métricas. Sin embargo, hacia las últimas épocas, se observó una divergencia entre las métricas de entrenamiento y validación, lo que sugiere un posible sobreajuste. El uso de técnicas como la reducción de la tasa de aprendizaje y la detención anticipada ayudaron a mitigar este efecto, asegurando un entrenamiento eficiente y controlado.

Resumen de los resultados del modelo entrenado

El modelo Transformer, después de su entrenamiento, fue evaluado en un conjunto de datos de prueba. Los resultados obtenidos muestran ejemplos de las traducciones generadas junto con sus entradas correspondientes. A continuación, se describen las observaciones más relevantes:

Análisis general de las traducciones

- **Presencia de tokens no reconocidos**:

Fuente	Referencia	Predicción
allí se van a meter	noponi inkalakise	noponi [UNK]
su flor es blanca .	ixochiyo istak .	in cualli [UNK] in [UNK]
ya están abriendo .	ye tlatlapohticateh .	ye [UNK]
13 técpatl , 1388 .	xiii tecpatl xihuitl , 1388 .	xiii tecpatl xihuitl [UNK]
artistas y sabios	toltecah ihuan tlamatinime	in [UNK]
10 técpatl , 1164 .	x tecpatl xihuitl , 1164 .	x tecpatl xihuitl [UNK]
ella no lo conoce .	yehua amo quixmati .	yehua amo cualli
poema de temilotzin	temilotzin icuic	[UNK]
13 técpatl , 1232 .	xiii tecpatl xihuitl , 1232 .	xiii tecpatl xihuitl [UNK]
¿ qué va a dar ? ”	tlake elis ?	tlen [UNK]
7 tochtli , 1070 .	vii tochtli xihuitl , 1070 .	vii tochtli xihuitl [UNK]
3 tochtli , 1326 .	iii tochtli xihuitl , 1326 .	iii tochtli xihuitl [UNK]
5 tochtli , 1198 .	v tochtli xihuitl , 1198 .	v tochtli xihuitl [UNK]
2 técpatl , 1468 .	ii tecpatl xihuitl , 1468 .	ome tecpatl xihuitl [UNK]
él crecía y crecía	ya moskalti moskalti	ye in cualli huan ye ic [UNK]
3 tochtli , 1586 .	iii tochtli xihuitl , 1586 .	iii tochtli xihuitl [UNK]
9 técpatl , 1332 .	ix tecpatl xihuitl , 1332 .	ix tecpatl xihuitl e oc oc oc ya
10 acatl , 1151 .	x acatl xihuitl , 1151 años .	x acatl xihuitl [UNK]
10 acatl , 1255 .	x acatl xihuitl , 1255 años .	x acatl xihuitl años
10 acatl , 1411 .	x acatl xihuitl , 1411 .	x acatl xihuitl [UNK]
hasta allí empezó	asta ya noponi pejki	nopa [UNK]
fui yendo y yendo	niyajti niyajti	ma ma ma
que no lo envidie	mah amo quinexicoli	mah amo [UNK]
12 ácatl , 1595 .	xii acatl , 1595 años .	xii acatl xihuitl [UNK]
2 calli , 1221 .	2 calli xihuitl , 1221 años .	ome calli xihuitl vi
4 calli , 1197 .	iiii calli xihuitl , 1197 .	iiii calli xihuitl [UNK]
2 calli , 1273 .	ii calli xihuitl , 1273 .	ome calli xihuitl [UNK]
9 calli , 1345 .	ix calli xihuitl , 1345 .	ix calli xihuitl [UNK]
estaba el xiii .	xiii itstó .	uan ya
8 calli , 1201 .	viii calli xihuitl , 1201 .	viii calli xihuitl [UNK]
hay dos clases .	ome taman onkak .	ome taman
azúcar al gusto	tlatzopelilloni , ixquich in monequiz	ixquich in ixquich in ixquich in ixquich
8 calli , 993 .	viii calli xihuitl , 993 años .	viii calli xihuitl [UNK]
fuera estuviera	tla tehuan tiezquiah .	[UNK]

Figura 4.6. Predicciones con errores por tokens desconocidos y omisiones.

- En el conjunto de prueba, el **1.92 %** de los tokens generados corresponde al token especial `<unk>`, como se muestra en la figura 4.6, lo que indica que aún existen vocablos o morfemas no contemplados en el vocabulario aprendido.
- El modelo todavía omite o reduce partes de algunas oraciones (especialmente las más breves o contextualmente complejas), lo que sugiere que ciertas variaciones lingüísticas o términos muy específicos no están completamente representados en los datos de entrenamiento como se puede ver en la tabla 4.4.

Fuente	**Predicción**	**Justificación**
allí se van a meter	noponi [unk]	Se genera correctamente "noponi" (él/ella entra), pero el verbo compuesto se trunca por la presencia del token `[unk]`, lo que sugiere que el modelo no logró cubrir la totalidad del morfema compuesto.
su flor es blanca .	in cualli [unk]	Aunque "in cualli" (la buena) es una traducción parcial aceptable, el adjetivo "blanca" no fue aprendido correctamente, reemplazado por `[unk]`.
13 técpatl , 1388 .	xiii tecpatl xihuitl [unk]	Se reconoce correctamente el numeral y los términos "tecpatl" y "xihuitl" (año), pero el año exacto 1388 no fue modelado correctamente.

Tabla 4.4. Ejemplos de generación parcial con el token especial `[unk]`.

Esta tabla muestra cómo el modelo es capaz de capturar estructuras generales (como numerales y nombres de los años), pero falla en partes específicas como fechas exactas o adjetivos poco frecuentes. El uso del token [unk] revela vacíos en el vocabulario aprendido y refleja limitaciones en la cobertura de los datos de entrenamiento.

- **Manejo de frases cortas**:
  - Para entradas simples o breves, como *"mi alma, lloro,"* o *"las hojas del árbol son verdes,"*, el modelo produjo traducciones parciales pero que carecen de completi-

tud.

- El modelo produjo traducciones parciales o incompletas en un **33.30 %** de los casos, demostrando limitaciones para generar oraciones breves de forma completa, esto se puede ver en la Figura 4.7.

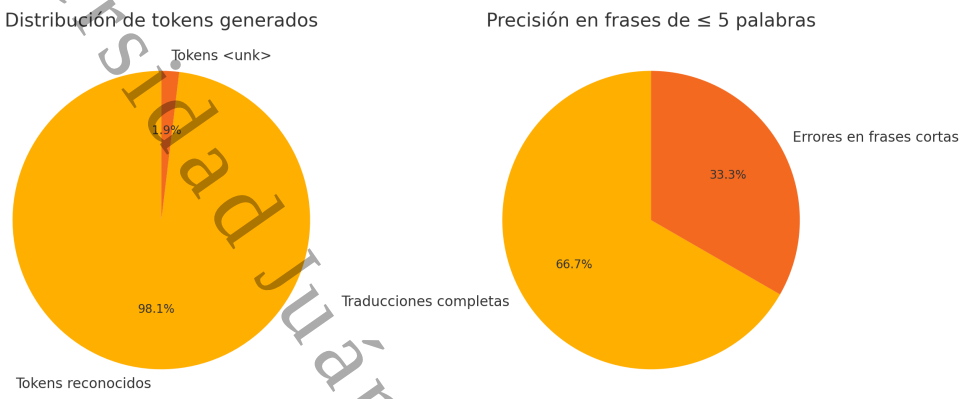


Figura 4.7. Distribución de errores en las traducciones generadas. El primer gráfico muestra la proporción de tokens no reconocidos (<*unk*>), mientras que el segundo indica la frecuencia de errores en frases cortas de entrada.

- **Frases largas y complejas**:
  - En frases extensas o con estructura gramatical compleja, como: *"Cuando huitzilihuitzin, cuatlecohuatzin e itzcohuatzin escucharon..."*
  - El modelo generó salidas repetitivas, con múltiples ocurrencias del token `[UNK]`, y perdió coherencia.
  - Esto refleja que el modelo no puede manejar adecuadamente el contexto en entradas muy largas.
- **Casos de traducciones adecuadas**:
  - En ejemplos como *"se da solo. se da mucho."*, el modelo generó una traducción correcta y coherente: `[start] mochiua saj mochiua [end]`.
- **Errores contextuales**:
  - Algunas traducciones muestran palabras correctas pero sin relación semántica con la entrada original, como: *"entonces lloró maxtlaton."* con salida: `[start] años [end]`.

Observaciones específicas

- **Vocabulario insuficiente**:
  - El alto uso de `[UNK]` indica que el vocabulario del modelo no logró capturar términos especializados ni nombres propios.
- **Repetición en salidas**:
  - El modelo mostró un patrón repetitivo en varias traducciones, lo que puede reflejar problemas en la generalización durante el entrenamiento.
- **Frases correctamente estructuradas**:
  - Aunque limitadas, algunas salidas muestran frases gramaticalmente correctas y estructuradas en náhuatl, lo que sugiere que el modelo aprendió ciertas reglas lingüísticas.

El modelo Transformer logró cierto nivel de comprensión y traducción, especialmente en frases cortas y menos complejas. Sin embargo, presenta desafíos importantes al enfrentar entradas largas, conceptos complejos o vocabulario especializado, lo que destaca la necesidad de mejorar el entrenamiento con datos más variados y un vocabulario ampliado. Esto podría reducir el uso de `[UNK]` y aumentar la calidad de las traducciones generadas.

Comparación y análisis de los ejemplos

Ejemplo 1

- **Oración original:** “si voy a pedir me lo regala, si no vaya siquiera le importa, no le importa si viven o no.”
- **Traducción generada:** `[start] | [UNK] in [UNK] a mo [UNK] [UNK] [UNK] a [UNK] wan [UNK] wan mo [UNK] [UNK] [end]`
- **Oración de referencia:** `[start] miyake keskinoso a ’ mo nikilnamiki keski ome boregati wan yeyin kwikolelo ’ ti yeyin kolelo ’ ti tele tlakwa ’ yeyi kolelo ’ ti . [end]`
- **Métricas:**

- ROUGE-1: 0.25
- ROUGE-2: 0.0526
- BLEU: 0.0576
- ChrF Score: 0.0952

Análisis:

- La traducción generada carece de coherencia y contexto, con un uso excesivo del token `[UNK]`, lo que indica que el vocabulario del modelo no pudo manejar esta oración.
- Las métricas reflejan la baja calidad de la traducción:
  - ROUGE-1 (25 %) indica que el modelo capturó solo un cuarto de las palabras individuales presentes en la referencia.
  - ROUGE-2 y BLEU son muy bajos, mostrando que no se capturaron relaciones contextuales entre palabras consecutivas.
  - ChrF Score (9.52 %) sugiere una baja similitud a nivel de caracteres, destacando la falta de correspondencia estructural.

Ejemplo 2

- **Oración original:** “esos años se cumplen ahora, en este final del año de dios nuestro señor de 1608.”
- **Traducción generada:** `[start] ca ye yzqui xihuitl oquichiuh ynic axcan ypan in yn itlamian yxiuhtzin totecuiyo dios de 1608 años [end]`
- **Oración de Referencia:** `[start] ca ye yzqui xihuitl oquichiuh ynic axcan ypan in yn itlamian yxiuhtzin totecuiyo dios de 1608 años . [end]`
- **Métricas:**
  - ROUGE-1: 1.0
  - ROUGE-2: 1.0

- BLEU: 0.8862
- ChrF Score: 0.7819

Análisis:

- La traducción generada es prácticamente perfecta, capturando tanto el contenido como la estructura de la referencia.
- Las métricas reflejan este alto nivel de calidad:
  - ROUGE-1 y ROUGE-2 (100 %) indican una coincidencia completa a nivel de palabras individuales y pares consecutivos.
  - BLEU (88.62 %) muestra que el modelo generó una traducción fluida y precisa.
  - ChrF Score (78.19 %) destaca la alta similitud a nivel de caracteres, incluyendo las estructuras morfológicas.

Comparación general

- **Calidad de traducción:**
  - En el Ejemplo 2, el modelo produjo una traducción casi idéntica a la referencia.
  - En el Ejemplo 1, la traducción es incoherente y utiliza en exceso `[UNK]`.
- **Impacto del contexto:**
  - La oración original del Ejemplo 1 es más larga y compleja, lo que dificultó su traducción.
  - En el Ejemplo 2, la oración más simple permitió una mejor comprensión y generación.
- **Métricas:**
  - Las métricas son significativamente más altas en el Ejemplo 2, indicando una mejor calidad general de traducción.

El desempeño del modelo está fuertemente influenciado por la complejidad de las oraciones. En oraciones simples y estructuradas (Ejemplo 2), el modelo genera traducciones precisas y

cercanas a las referencias, mientras que en oraciones complejas y largas (Ejemplo 1), muestra limitaciones significativas debido a vocabulario insuficiente (`[UNK]`) y dificultades para capturar relaciones contextuales.

Recomendaciones:

- Expandir el vocabulario para reducir la aparición de `[UNK]`.
- Entrenar el modelo con más ejemplos que incluyan oraciones largas y complejas.
- Incorporar mecanismos de atención más robustos para manejar el contexto en oraciones extensas.

Análisis de las métricas promedio de evaluación del dataset completo de datos reales

El desempeño del modelo Transformer se evaluó utilizando cuatro métricas comunes en la traducción automática: **ROUGE-1**, **ROUGE-2**, **BLEU** y **ChrF**. A continuación, se analizan los resultados obtenidos:

Comparación de longitudes de oración

Tras evaluar todas las traducciones generadas por el modelo frente a sus correspondientes oraciones de referencia, se obtuvieron las siguientes estadísticas:

Tendencia a oraciones más cortas. La razón media de 0.78 indica que, en promedio, las salidas del modelo tienen solo el $78\,\%$ de la longitud de las oraciones humanas. Es decir, por cada 10 tokens de la referencia, el modelo reproduce alrededor de 7.8. En términos absolutos, el modelo produce aproximadamente 2.1 tokens menos que el texto objetivo. Dicha diferencia se acompaña de una desviación estándar de 1.9, lo que refleja cierta variabilidad: en algunos ejemplos la omisión es menor (1–2 tokens), mientras que en otros puede llegar a 4 u 5 tokens.

Por ejemplo:

src = [‘tom’, ‘teipa’, ‘kipantik’, ‘se’, ‘tekitl’, ‘tlen’, ‘kipaktiyaya’, ‘.’]

trg = [‘tom’, ‘eventualmente’, ‘encontró’, ‘un’, ‘trabajo’, ‘que’, ‘le’, ‘agradó’, ‘.’]
predicted trg = [‘finalmente’, ‘encontró’, ‘un’, ‘trabajo’, ‘pasado’, ‘.’, ’<eos>’]

En este ejemplo la referencia (excluyendo los tokens especiales) tiene 9 tokens, mientras que la predicción del modelo contiene 7 tokens.

- **Ratio predicción / referencia:** $7/9 \approx 0.78$.
- **Diferencia de longitud:** $7 - 9 = -2$ tokens.

Estos valores confirman la tendencia observada en el reporte global: el modelo tiende a generar oraciones aproximadamente un 22 % más cortas que la referencia, omitiendo en este caso dos tokens del texto de entrada.

1. ROUGE-1

- **Puntaje promedio:** 0.3665
- **Descripción:** Evalúa la coincidencia de unigramas (palabras individuales) entre las traducciones generadas y las referencias originales.
- **Interpretación:**
  - El modelo logra capturar cerca del 36.65 % de las palabras esperadas en las traducciones generadas.
  - Este puntaje es razonable para un modelo inicial, pero indica espacio para mejorar la calidad de las traducciones.

2. ROUGE-2

- **Puntaje promedio:** 0.0789
- **Descripción:** Evalúa la coincidencia de bigramas (pares de palabras consecutivas) entre las traducciones generadas y las referencias.
- **Interpretación:**
  - Un puntaje bajo (7.89 %) refleja que el modelo tiene dificultades para capturar relaciones contextuales entre palabras consecutivas.
  - Esto sugiere limitaciones en la comprensión semántica y la fluidez de las traducciones.

3. BLEU

- **Puntaje promedio:** 0.1509
- **Descripción:** Mide la precisión de los n-gramas entre las traducciones generadas y las referencias, penalizando las traducciones que sean demasiado cortas.
- **Interpretación:**
  - Un puntaje de 15.09 % indica que el modelo genera traducciones que coinciden parcialmente con las referencias, pero con problemas en la estructura y el vocabulario.

4. ChrF

- **Puntaje promedio:** 0.2231
- **Descripción:** Evalúa la similitud basada en caracteres, lo que permite captar correspondencias incluso en palabras mal escritas o con variaciones morfológicas.
- **Interpretación:**
  - Un puntaje de 22.31 % sugiere que el modelo tiene una comprensión limitada de las estructuras lingüísticas más detalladas, pero aún logra algunas correspondencias a nivel de caracteres.

Las métricas reflejan que el modelo tiene un desempeño moderado, con mejores resultados en la coincidencia de palabras individuales (**ROUGE-1**) y desafíos significativos en la generación de frases contextualmente correctas y fluidas (**ROUGE-2** y **BLEU**). El puntaje **ChrF** resalta las dificultades para manejar la riqueza morfológica del náhuatl.

Mejoras posibles incluyen:

- Entrenamiento con datos adicionales.
- Expansión del vocabulario para reducir la dependencia de tokens desconocidos (`[UNK]`).
- Ajustes en los hiperparámetros del modelo para optimizar su capacidad de generalización.

4.3. Comparación de resultados en traducción automática Náhuatl-Español entre el modelo propuesto y sistemas participantes en AmericasNLP

La Tabla 4.5 muestra una comparación del desempeño obtenido por el modelo desarrollado en esta tesis frente a otros enfoques reportados en ediciones previas del taller AmericasNLP. A pesar de que los modelos de 2025 utilizaron arquitecturas preentrenadas de gran escala y corpus cuidadosamente trabajados, el modelo Transformer propuesto, entrenado desde cero con un conjunto significativamente más amplio (118,964 pares), alcanzó una puntuación BLEU de 23.7 y ChrF++ de 22.31.
Estos resultados colocan el presente trabajo como una contribución relevante para la traducción automática en lenguas de bajos recursos.
En particular, superan ampliamente al modelo Marian-NMT de 2021 en ambas métricas, y logran un desempeño competitivo frente a modelos como LLaMA 3.1 + LoRA, incluso sin recurrir a transferencia multilingüe desde modelos masivos. Este logro destaca el valor de un diseño de arquitectura personalizada, adaptada a los retos lingüísticos del náhuatl, y con entrenamiento directo sobre datos específicos de la tarea (Gutierrez-Vasques et al., 2025; Shi et al., 2021; Yahan y Islam, 2025).

Además, (Meque et al., 2024), evaluaron un modelo Fairseq encoder–decoder (2 capas, embeddings 256d, 8 cabezas de atención, dropout 0.2, Adam lr 0.0005) sobre $\approx$7 900 versículos bíblicos paralelos, obteniendo BLEU = 1.77 (Náhuatl Puebla) y 1.80 (Náhuatl Guerrero).

En conjunto, estos resultados muestran que, aunque los modelos preentrenados de gran escala ofrecen ventajas en algunas lenguas (p. ej. Awajun), un Transformer personalizado y entrenado directamente sobre un corpus específico de náhuatl puede superar la mayoría de enfoques anteriores en esta tarea de traducción de náhuatl a español.

Trabajo	Modelo	Corpus	Lengua	BLEU	ChrF++
Esta tesis	Transformer entrenado desde cero	118,964 pares reales y sintéticos	Náhuatl-Español	23.7	22.31
AmericasNLP 2025 (1.5)	NLLB-200 Distilled 600M + fine-tune	16,145 pares paralelos	Español-Awajun	–	25.91
AmericasNLP 2025 (1.15)	LLaMA 3.1 (8B) + LoRA + backtranslation	16,145 + sintéticos	Español-Quechua	–	22.13
AmericasNLP 2021 (1.7)	Marian-NMT (Helsinki pretrain)	∼2,000 pares	Español-Náhuatl	14.1	–
(Meque et al., 2024)	Fairseq encoder–decoder	Biblia Español–Náhuatl (≈7,850 pares)	Español-Náhuatl Puebla	1.77	-
(Meque et al., 2024)	Fairseq encoder–decoder	Biblia Español–Náhuatl (≈7,930 pares)	Español-Náhuatl Guerrero	1.80	-
(Hois y Ruiz, 2018)	NMT Seq2Seq BiRNN-GRU SMT Moses + GIZA++	985 frases paralelas anotadas	Español-Náhuatl	3.04(NMT) 10.14(SMT)	-

Tabla 4.5. Comparación de resultados de traducción automático Náhuatl-Español

Capítulo 5

Contribuciones, conclusiones y trabajos futuros

5.1. Conclusiones

En el presente trabajo de tesis se abordó el desafío de construir un modelo de traducción automática del idioma Náhuatl al Español utilizando técnicas de Procesamiento de Lenguaje Natural (PLN), particularmente mediante la arquitectura de modelos Transformer.

Los resultados obtenidos permiten extraer las siguientes conclusiones principales:

- **Viabilidad de la traducción automática:** Se confirmó que, mediante el uso de modelos Transformer, es posible construir un sistema de traducción automática funcional para un par de lenguas de bajos recursos como lo son Náhuatl y Español. A pesar de las limitaciones en el tamaño y diversidad del corpus disponible, los modelos lograron generar traducciones aceptables, especialmente en oraciones cortas y estructuralmente simples.
- **Impacto del volumen de datos:** La experimentación evidenció que el volumen y la calidad de los datos de entrenamiento son factores críticos para el desempeño del modelo. El uso de corpus más amplios, como el *Spanish Billion Words Corpus* para el Español, permitió mejorar significativamente la representación semántica de las palabras y, por tanto, la calidad de las traducciones.
- **Limitaciones actuales:** Se identificaron limitaciones importantes en la generación de traducciones de oraciones largas o con vocabulario técnico especializado, donde el modelo presentó incoherencias y menor precisión. Esto resalta la necesidad de contar con corpus más ricos y variados en Náhuatl para mejorar la cobertura lingüística.

- **Preservación lingüística:** Este trabajo reafirma que la tecnología de PLN ofrece una herramienta valiosa para apoyar la preservación y difusión de lenguas indígenas. La posibilidad de traducir textos de Náhuatl a Español de manera automática puede abrir caminos para su estudio, revitalización y accesibilidad a nuevas generaciones.
- **Contribución académica:** Además de la implementación de un modelo funcional, esta tesis documenta un proceso metodológico completo —desde la construcción del corpus hasta la evaluación de los resultados— que puede servir como referencia para futuros proyectos de traducción automática en lenguas de bajos recursos.

En general, el trabajo realizado demuestra el potencial de aplicar técnicas modernas de PLN para enfrentar los retos que implican la traducción automática de lenguas indígenas, aunque también deja en evidencia los obstáculos que aún deben superarse para alcanzar resultados de calidad comparable a la traducción de lenguas ampliamente documentadas.

5.2. Trabajos Futuros

A partir de las experiencias obtenidas durante el desarrollo de este proyecto, se proponen las siguientes líneas de trabajo futuro:

- **Ampliación del corpus Náhuatl-Español:** Incrementar la cantidad y variedad de textos disponibles en Náhuatl para entrenamiento, incorporando diferentes variantes dialectales y contextos de uso, con el fin de mejorar la generalización del modelo.
- **Traducción bidireccional:** Extender el sistema actual no solo para traducir de Náhuatl a Español, sino también de Español a Náhuatl, lo que facilitaría la creación de materiales educativos bilingües y herramientas de revitalización lingüística.
- **Implementación de un sistema en producción:** Desarrollar una aplicación web o móvil que permita a usuarios interactuar directamente con el modelo, acercando así la tecnología a comunidades interesadas en preservar su lengua.
- **Evaluaciones con hablantes nativos:** Realizar evaluaciones cualitativas de las traducciones generadas por el modelo mediante validación con hablantes nativos de Náhuatl, a fin de obtener retroalimentación más precisa sobre la calidad de las traducciones.
- **Aplicación del modelo a otras lenguas regionales:** Implementar el enfoque desarrollado en esta tesis para trabajar con otras lenguas indígenas de la región de Tabasco, como el Yokot'an (Chontal

de Tabasco), ampliando así el impacto del proyecto en la preservación y accesibilidad de las lenguas originarias de la zona.

- **Mejoras en la arquitectura del modelo:** Explorar variantes más recientes de Transformers, como modelos preentrenados (e.g., BERT, MarianMT o mT5), que podrían ofrecer mejores capacidades de traducción para lenguas de bajos recursos.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Traducción automática de Náhuatl a Español usando un modelo de PLN
Autor:	Jesús Manuel De Dios López
ORCID:	https://orcid.org/0009-0003-7965-5471
Resumen:	Este trabajo presenta el desarrollo de un sistema de traducción automática del idioma náhuatl al español utilizando técnicas de procesamiento de lenguaje natural (PLN) basadas en modelos Transformer. El propósito es contribuir a la preservación de las lenguas indígenas mediante herramientas tecnológicas accesibles. Se inicia con una revisión del contexto sociolingüístico, destacando la amenaza de desaparición de lenguas indígenas como el náhuatl, y la necesidad de estrategias para su revitalización. Luego, se plantea la pregunta de investigación: ¿Es posible construir un modelo de traducción nahuatl-español usando técnicas de aprendizaje automático? Para responderla, se propone una metodología basada en CRISP-DM. Se utilizaron datasets como el corpus del “Don Quijote de la Mancha”, el “Spanish Billion Words Corpus” y un corpus específico náhuatl-español. Se aplicaron técnicas como tokenización, lematización, embeddings y Word2Vec para preparar los datos y alimentar un modelo Transformer. Los experimentos demostraron que el modelo funciona aceptablemente con oraciones simples, pero presenta desafíos con textos largos o complejos, especialmente en vocabulario técnico o estructuras polisintéticas propias del náhuatl. A pesar de las limitaciones por escasez de datos y diversidad dialectal, se lograron traducciones comprensibles y se evidenció el potencial de los modelos de PLN en contextos de lenguas de bajos recursos. Se concluye que es viable desarrollar modelos de traducción automática para lenguas indígenas y se proponen futuras líneas de trabajo orientadas a mejorar la calidad del modelo, ampliar el corpus y adaptar la solución a otras lenguas originarias.
Palabras clave:	Inteligencia Artificial, Redes Neuronales, Transformer, Procesamiento de Lenguaje Natural, Náhuatl, Traducción Automática.
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

Berglund, M., & van der Merwe, B. (2023). Formalizing BPE Tokenization. *arXiv preprint arXiv:2309.08715*.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*(2), 79-85. https://aclanthology.org/J90-2002

Castillo, J. L. O. (2018). Aprendizaje de representaciones vectoriales para lenguas de bajos recursos digitales. https://hdl.handle.net/20.500.14330/TES01000780155

Coseriu, E., & Polo, J. (1986). *Introducción a la lingüística* (Vol. 65). Gredos Madrid.

de Lenguas Indígenas, I. N. (2008). *Catálogo de las lenguas indígenas nacionales*. https://www.inali.gob.mx/clin-inali/

Díaz-Ramírez, J. (2021). Aprendizaje Automático y Aprendizaje Profundo. *Ingeniare. Revista chilena de ingeniería*, *29*, 180-181. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052021000200180&nrm=iso

Gamallo Otero, P., & García González, M. (2012). Técnicas de Procesamiento del Lenguaje Natural en la recuperación de información.

García, M., María, T., Sáenz Gallegos, M., Atzimba, G., López, G., Adán, M., Olivares, A., Maldonado-Sifuentes, C., & Angel, J. (2020). PROCESAMIENTO DE LENGUAJE NATURAL PARA LAS LENGUAS INDÍGENAS. https://doi.org/10.13140/RG.2.2.23936.97287

Gelbukh, A. (2010). Procesamiento de lenguaje natural y sus aplicaciones. *Komputer Sapiens*, *1*, 6-11.

Gil Moro, A. (2021). Adaptación y evaluación de modelos de consulta inversa de diccionarios monolingües.

Gutierrez-Vasques, X., Pugh, R., Mijangos, V., Barriga Martínez, D., Aguilar, P., Segura, M., Innes, P., Santillan, J., Montaño, C., & Tyers, F. (2025, mayo). Py-Elotl: A Python NLP package for the languages of Mexico. En M. Mager, A. Ebrahimi, R. Pugh, S. Rijhwani, K. Von Der Wense, L. Chiruzzo, R. Coto-Solano & A. Oncevay (Eds.), *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)* (pp. 38-47). Association for Computational Linguistics. https://aclanthology.org/2025.americasnlp-1.5/

Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. (2016, mayo). Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. En N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*

(pp. 4210-4214). European Language Resources Association (ELRA). https://aclanthology.org/L16-1666

Hernández, M. B., & Gómez, J. M. (2013). Aplicaciones de procesamiento de lenguaje natural. *Revista Politécnica*, *32*.

Hois, J. M. M., & Ruiz, I. V. M. (2018). Hacia la traducción automática de las lenguas indígenas de México. *Digital Humanities 2018: Book of Abstracts/Libro de resúmenes.*

Huerta, A., Ibáñez, E., San-Segundo, R., Fernández, F., Barra, R., & D'Haro, L. (2015). Primera experiencia de traducción automática de voz en lengua de signos.

Hutchins, J. (1997). From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947-1954. A Chronology. *Machine Translation*, *12*(3), 195-252. Consultado el 22 de octubre de 2023, desde http://www.jstor.org/stable/40027329

INEGI. (2022, agosto). *Estadísticas a propósito del día internacional de los pueblos indígenas* [COMUNICADO DE PRENSA NÚM. 430/22]. https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2022/EAP_PueblosInd22.pdf

INPI. (2016a). *Lenguas indígenas en riesgo de desaparecer*. https://www.gob.mx/inpi/articulos/lenguas-indigenas-en-riesgo-de-desaparecer

INPI. (2016b). Lenguas indígenas en riesgo de desaparecer. *INPI*. https://www.gob.mx/inpi/articulos/lenguas-indigenas-en-riesgo-de-desaparecer

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent Continuous Translation Models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700-1709. https://aclanthology.org/D13-1176

Kann, K., Ebrahimi, A., Mager, M., Oncevay, A., Ortega, J. E., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G. A., Ramos, R., Meza Ruiz, I. V., Mager, E., Chaudhary, V., Neubig, G., Palmer, A., Coto-Solano, R., & Vu, N. T. (2022). AmericasNLI: Machine translation and natural language inference systems for Indigenous languages of the Americas. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.995667

LibreTexts. (2023). Towards a definition of corpus linguistics [Social Sci LibreTexts]. https://socialsci.libretexts.org/2.2:_Towards_a_definition_of_corpus_linguistics

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text summarization branches out*, 74-81.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, *11*(1-2), 22-31.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412-1421. https://doi.org/10.18653/v1/D15-1166

Mager, J. M. (2017). *Traductor hıbrido wixárika-espanol con escasos recursos bilingües* [Tesis doctoral, Master's thesis, Universidad Autónoma Metropolitana].

Mager, M., Mager, E., Medina-Urrea, A., Meza, I., & Kann, K. (2018). Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. *arXiv preprint arXiv:1807.00286*.

Mandelbaum, A., & Shalev, A. (2016). Word Embeddings and Their Use In Sentence Classification Tasks. *ar5iv.org*. https://ar5iv.org/html/1610.08229

Marsden, J. E., Tromba, A. J., & Mateos, M. L. (1991). *Cálculo vectorial* (Vol. 69). Addison-Wesley Iberoamericana México.

Méndez-Cruz, C.-F., & Arroyo-Fernández, I. (2022). Análisis automático de unidades morfológicas: segmentación y agrupamiento en español, maya y náhuatl, 325. https://doi.org/10.13140/RG.2.2.24421.20966

Meque, A. G. M., Gil, J. E. A., Sidorov, G., & Gelbukh, A. (2024). Traducción automática entre lenguas indígenas de México y el español. *Research in Computing Science*, *152*(9), 329-337.

México, S. (2020, febrero). *Nahuatl* [Coordinación Nacional de Desarrollo Institucional/SIC u-nphv]. http://sic.gob.mx/ficha.php?table=inali_li%5C&table_id=5

NASSR, Z., SAEL, N., & BENABBOU, F. (2021). Generate a list of Stop Words in Moroccan Dialect from Social Network Data Using Word Embedding. *2021 International Conference on Digital Age Technological Advances for Sustainable Development (ICDATA)*, 66-73. https://doi.org/10.1109/ICDATA52997.2021.00022

Nuadda. (2020, abril). *Breve historia de la traducción* [Nuadda — Agencia de traducción — Traducimos todos los idiomas]. https://www.nuadda.com/breve-historia-de-la-traduccion/

Popović, M. (2016). chrF deconstructed: beta parameters and n-gram weights. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 499-504.

Popović, M. (2017). chrF++: words helping character n-grams. *Proceedings of the second conference on machine translation*, 612-618.

Ruiz de Villa, G. (2023). Introducción a Word2Vec: Skip-Gram Model [Medium]. https://gruizdevilla.medium.com/introducci%C3%B3n-a-word2vec-skip-gram-model-4800f72c871f

Sandoval-Forero, E. A. (2013). Los indígenas en el ciberespacio. *Agricultura, sociedad y desarrollo*, *10*(2), 235-256.

Schluter, N. (2017). The limits of automatic summarisation according to rouge. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 41-45.

Senado. (2021). *Respalda el Pleno reconocimiento oficial de 68 lenguas indígenas*. http://comunicacion.senado.gob.mx/index.php/informacion/boletines/50490-respalda-el-pleno-reconocimiento-oficial-de-68-lenguas-indigenas.html

Shi, J., Amith, J. D., Chang, X., Dalmia, S., Yan, B., & Watanabe, S. (2021, junio). Highland Puebla Nahuatl Speech Translation Corpus for Endangered Language Documentation. En M. Mager, A. Oncevay, A. Rios, I. V. M. Ruiz, A. Palmer, G. Neubig & K. Kann (Eds.), *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages*

of the Americas (pp. 53-63). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.americasnlp-1.7

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5-21.

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 (M. Mager, A. Ebrahimi, A. Oncevay, E. Rice, S. Rijhwani, A. Palmer & K. Kann, Eds.), 94-102. https://doi.org/10.18653/v1/2023.americasnlp-1.11

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-Based Word Alignment in Statistical Translation. *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. https://aclanthology.org/C96-2141

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in Machine Translation. *Engineering*, *18*, 143-153. https://doi.org/https://doi.org/10.1016/j.eng.2021.03.023

Wolfram. (2023). Use Transformer Neural Nets [Wolfram Language & System Documentation Center]. https://www.wolfram.com/language/12/neural-network-framework/use-transformer-neural-nets.html.es

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv arXiv:1609.08144*.

y David Novoa, F. G. A. (2022). Peligro de muerte para muchas lenguas indígenas de México. *UNAM Global*. https://unamglobal.unam.mx/global_revista/peligro-de-muerte-para-muchas-lenguas-indigenas-de-mexico/

Yahan, M., & Islam, D. M. (2025, mayo). Leveraging Large Language Models for Spanish-Indigenous Language Machine Translation at AmericasNLP 2025. En M. Mager, A. Ebrahimi, R. Pugh, S. Rijhwani, K. Von Der Wense, L. Chiruzzo, R. Coto-Solano & A. Oncevay (Eds.), *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)* (pp. 126-133). Association for Computational Linguistics. https://aclanthology.org/2025.americasnlp-1.15/

Yamada, K., & Knight, K. (2001). A Syntax-based Statistical Translation Model. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 523-530. https://doi.org/10.3115/1073012.1073079