

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Apizaco

INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES

PROYECTO DE RESIDENCIAS PROFESIONALES:

Sistema de Tutoría Inteligente Conversacional para Mejorar la
Competencia en la lengua náhuatl de Hablantes No Nativos a
Través del Análisis de Transcripciones (Fase 2)

PRESENTA:

Rosa De Guadalupe López Montiel

No. De Control:

20371020

ASESORES:

INTERNO: Dr. José Crispín Hernández Hernández

EXTERNO: Dr. Juan Ramos Ramos



Instituto Tecnológico de Apizaco
Subdirección de Planeación

Apizaco, Tlaxcala, **09/febrero/2026**

Oficio No. S.C./OPV/0011/26

ASUNTO: **Liberación de proyecto
para titulación integral**

CECILIA YARELLI ARREGÍN LARA
JEFA DEL DEPTO. DE DIVISIÓN DE ESTUDIOS PROFESIONALES
P R E S E N T E.

POR ESTE MEDIO LE INFORMO QUE HA SIDO LIBERADO EL SIGUIENTE PROYECTO PARA LA TITULACIÓN INTEGRAL.

NOMBRE DEL EGRESADO:	LÓPEZ MONTIEL ROSA DE GUADALUPE
CARRERA:	INGENIERÍA EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIONES
NO. DE CONTROL:	20371020
NOMBRE DEL PROYECTO:	Sistema de Tutoría Inteligente Conversacional para mejorar la competencia en la lengua náhuatl de hablantes No Nativos a Través del Análisis de Transcripciones (fase 2)
PRODUCTO U OPCIÓN:	TITULACIÓN INTEGRAL INFORME DE RESIDENCIA PROFESIONAL

AGRADEZCO DE ANTEMANO SU VALIOSO APOYO EN ESTA IMPORTANTE ACTIVIDAD PARA LA FORMACIÓN PROFESIONAL DE NUESTROS EGRESADOS.

ATENTAMENTE

*Excelencia en Educación Tecnológica.
Pensar para Servir, Servir para Triunfar®*

MIQUELINA SÁNCHEZ PULIDO
JEFA DEL DEPTO. DE SISTEMAS Y COMPUTACIÓN

ccp. Coordinación de carrera
Expediente
MSP/mrol*

ESTADOS UNIDOS MEXICANOS
SECRETARÍA DE EDUCACIÓN PÚBLICA
TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE APIZACO
DEPARTAMENTO DE
SISTEMAS Y COMPUTACIÓN

JOSÉ CRISPÍN HERNÁNDEZ HERNÁNDEZ
ASESOR INTERNO



2026
año de
**Margarita
Maza**





Tzompantepec, Tlaxcala, **10/Febrero/2026**

ROSA DE GUADALUPE LÓPEZ MONTIEL
EGRESADO(A) DE LA CARRERA DE INGENIERÍA
EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIONES
CON NÚMERO DE CONTROL 20371020.
P R E S E N T E.

Por este medio, me permito informar a usted que por aprobación de la comisión revisora asignada para valorar el trabajo que mediante la opción **TITULACIÓN INTEGRAL INFORME DE RESIDENCIA PROFESIONAL** y cuyo tema es: **"SISTEMA DE TUTORIA INTELIGENTE CONVERSACIONAL PARA MEJORAR LA COMPETENCIA EN LA LENGUA NÁHUATL DE HABLANTES NO NATIVOS A TRAVÉS DEL ANÁLISIS DE TRANSCRIPCIONES (FASE 2)"**, conforme a lo establecido en el procedimiento para la obtención del Título Profesional en el Instituto Tecnológico, la División de Estudios Profesionales a mi cargo le otorga la:

AUTORIZACIÓN DE IMPRESIÓN

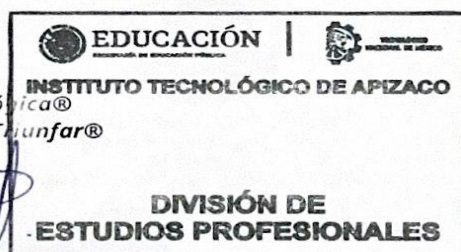
Debiendo entregar **1 USB** del mismo, apropiadamente empastados, a la brevedad, para presentar su Acto de Recepción Profesional.

Sin otro particular por el momento me despido.

ATENTAMENTE

Excelencia en Educación Tecnológica®
Pensar para Servir, Servir para Triunfar®

CECILIA YARELY ARREGUÍN LARA
JEFA DE LA DIVISIÓN DE ESTUDIOS PROFESIONALES



ccp. Archivo
CYAL/Gsc*



2026
año de
Margarita
Maza



Agradecimientos

Expreso mi más sincero agradecimiento a todas las personas e instituciones que contribuyeron de manera directa e indirecta a la realización de mi Residencia Profesional, etapa fundamental en mi formación académica y desarrollo profesional.

En primer lugar, agradezco profundamente a mi familia, especialmente a mis padres, por su apoyo constante, su confianza y los valores que me han inculcado a lo largo de mi vida. Su esfuerzo, comprensión y motivación han sido un pilar esencial para alcanzar este logro. A mis hermanos, gracias por su acompañamiento, paciencia y apoyo a lo largo de este proceso, los cuales fueron de gran importancia en los momentos de mayor exigencia académica.

De igual manera, expreso mi agradecimiento al **Instituto Tecnológico de Apizaco** por brindarme la preparación académica y las herramientas necesarias para llevar a cabo esta Residencia Profesional. Reconozco la labor de los docentes, así como del personal administrativo, quienes con su orientación, compromiso y disposición contribuyeron al adecuado desarrollo de este trabajo.

De forma especial, agradezco al **Dr. José Crispín Hernández Hernández**, mi asesor de Residencia Profesional, por su guía, dedicación y valiosas aportaciones durante el desarrollo del proyecto. Su experiencia, seguimiento y confianza fueron determinantes para la correcta ejecución de las actividades y el cumplimiento de los objetivos establecidos.

Finalmente, agradezco a todas aquellas personas que, de una u otra manera, aportaron conocimientos, apoyo y motivación para la conclusión de esta etapa, la cual representa un paso importante en mi crecimiento académico y profesional.

Contenido

Agradecimientos	2
Contenido	3
Resumen	6
Introducción	7
Generalidades de la institución.....	8
Ubicación	8
Logo de la empresa.....	8
El Tecnológico Nacional de México.....	8
Misión.....	9
Visión	9
Valores	9
Datos generales del proyecto	10
Nombre del proyecto	10
Objetivos	10
Objetivo principal.....	10
Delimitación.....	10
Justificación.....	10
Marco teórico.....	11
Estado del arte	12
Revisión documental sobre Procesamiento del Lenguaje Natural.....	13
Identificación de recursos lingüísticos disponibles para el náhuatl	14
Recopilación de transcripciones a partir de diccionarios náhuatl–español	16
Evaluación de la calidad y estructura de las transcripciones obtenidas.....	18
Limpieza de los textos seleccionados	19
Análisis básico de las palabras normalizadas.....	23
Traducción náhuatl–español	25
Texto y validación de las tablas.....	27
Análisis y organización de términos léxicos derivados	29
Preparación del corpus lingüístico para su análisis	30
Análisis y organización de términos léxicos derivados	31
Extracción automatizada de información desde recursos web	33
Implementación del código.....	33

Dificultades técnicas y resolución de incidencias	34
Implementación de la función de búsqueda léxica	35
Resultados obtenidos en la extracción	37
Dificultades técnicas encontradas	37
Limpieza y normalización de datos lingüísticos	38
Análisis lingüístico mediante técnicas de Procesamiento del Lenguaje Natural	41
Validaciones	43
Interpretación de patrones lingüísticos y preparación de datos para el sistema conversacional.....	44
Extracción de información relevante.....	44
Preparación de los datos para la tutoría.....	45
Validación automatizada	45
Integración de los datos en el sistema de tutoría conversacional y preparación para la realidad virtual	46
1. Preparación de la información para el motor conversacional.....	46
Bocetos de realidad virtual: planificación y desarrollo	46
Objetivo de los bocetos de realidad virtual	46
Herramientas utilizadas: Unity como plataforma principal	47
Proceso de realización de los bocetos.....	47
Beneficios de los bocetos de VR para el proyecto	49
Uso de Unity en el proyecto.....	50
Funcionamiento del uso de Unity en el proyecto	51
Aportación de Unity en el proyecto.....	51
Resultados.....	52
Definición del diccionario y preparación de los recursos lingüísticos	54
Inicio del procesamiento automatizado con Python.....	55
Desarrollo de scripts en Python.....	57
Dificultades técnicas y resolución	58
Investigación y selección de Unity para realidad	58
Integración	59
Investigación y selección de la herramienta de desarrollo.....	59
Proceso de conceptualización y diseño.....	60
Construcción de escenas y objetos tridimensionales	60

Dificultades técnicas y consideraciones	62
Importancia del componente de Realidad Virtual dentro del proyecto.....	62
Integración de los componentes desarrollados.....	68
Conclusión.....	70
Recomendaciones	72
Experiencia obtenida	73
Competencias desarrolladas	74
Referencias	76

Resumen

Durante el desarrollo de la Residencia Profesional, se llevó a cabo una investigación orientada al análisis y aplicación del Procesamiento de Lenguaje Natural, disciplina perteneciente al área de la inteligencia artificial. Esta investigación forma parte del proyecto *“Sistema de tutoría inteligente conversacional para mejorar la competencia en la lengua náhuatl de hablantes no nativos a través del análisis de transcripciones (Fase 2)”*, cuyo propósito es contribuir al fortalecimiento del aprendizaje de una lengua originaria mediante el uso de tecnologías digitales.

A lo largo del desarrollo del proyecto, se realizó un análisis teórico y práctico de los principios del Procesamiento de Lenguaje Natural, abordando aspectos como el análisis y procesamiento de texto, la comprensión del lenguaje, la estructuración de datos lingüísticos y el tratamiento automatizado de información textual. Estas actividades permitieron comprender la relevancia de esta disciplina en la creación de sistemas capaces de interpretar y analizar el lenguaje humano.

Como parte de las actividades realizadas, se llevó a cabo la investigación y análisis de transcripciones en lengua náhuatl, con el objetivo de identificar patrones lingüísticos que contribuyan a la mejora de un sistema conversacional inteligente. Asimismo, se analizaron diversas técnicas y herramientas empleadas en el procesamiento automático del lenguaje, evaluando su aplicación dentro del contexto educativo y tecnológico del proyecto.

Los resultados obtenidos durante la Residencia Profesional permitieron fortalecer los conocimientos en el área de inteligencia artificial, así como desarrollar habilidades en investigación, análisis de información y aplicación de tecnologías emergentes. El presente trabajo representa una aportación al desarrollo de soluciones tecnológicas orientadas al aprendizaje de lenguas, destacando el uso del procesamiento del lenguaje como un recurso relevante para la innovación educativa.

Introducción

El aprendizaje de lenguas originarias mediante el uso de herramientas tecnológicas constituye una alternativa relevante para fortalecer su difusión y preservación, especialmente en contextos donde los hablantes no nativos enfrentan dificultades para acceder a recursos didácticos adecuados. En este sentido, la lengua náhuatl presenta retos asociados a su estructura lingüística y a la limitada disponibilidad de materiales interactivos que faciliten su comprensión y práctica de manera sistemática.

Ante esta problemática, el desarrollo de sistemas basados en inteligencia artificial permite analizar, organizar y procesar información lingüística de forma estructurada, contribuyendo a la generación de recursos educativos innovadores. En particular, el uso de sistemas conversacionales inteligentes posibilita la interacción con los usuarios, favoreciendo el aprendizaje autónomo y el fortalecimiento de competencias lingüísticas mediante el análisis de datos confiables.

En esta etapa del proyecto se llevó a cabo la transcripción y adaptación de términos provenientes de un diccionario de la lengua náhuatl al español, con el propósito de integrarlos a un sistema de tutoría conversacional. Este proceso permitió conservar el significado original de los vocablos, presentándolos de manera ordenada y comprensible para su análisis y posterior aplicación tecnológica. La correcta estructuración de la información resultó fundamental para asegurar la calidad de los datos utilizados por el sistema.

Asimismo, se emplearon herramientas de inteligencia artificial como apoyo para el análisis de la información lingüística, lo que permitió optimizar su organización y facilitar su incorporación dentro de un entorno conversacional orientado al aprendizaje. La integración de estos elementos tecnológicos y lingüísticos contribuye al desarrollo de soluciones educativas que apoyan la adquisición de la lengua náhuatl por parte de hablantes no nativos, promoviendo el acceso a contenidos estructurados y funcionales.

Generalidades de la institución

Ubicación

Av. Instituto Tecnológico No. 418, San Andrés Ahuashuatepec, Municipio de Tzompantepec, Tlaxcala, Méx. C.P. 90491.

Tel: 241 417 2010 Ext. 136

Correo: difusion@apizaco.tecnm.mx

Web: Apizaco.tecnm.mx

Logo de la empresa



figura 1.logo de la empresa

El Tecnológico Nacional de México

El Tecnológico Nacional de México (TecNM) es una institución de educación superior tecnológica que forma profesionales e investigadores con habilidades en el campo laboral, pensamiento crítico, sentido ético, colaborativos y competentes para contribuir al desarrollo nacional y regional. Fue creada en 2014. El TecNM ofrece educación de excelencia y es la institución de educación superior tecnológica con mayor movilidad social en México. La misión del TecNM es formar profesionales mediante un modelo educativo integral de calidad, que garantiza una formación técnica humanística, con capacidad para investigar y aplicar tecnología con impacto en el desarrollo de la sociedad.

Misión

Formar profesionistas que sean líderes visionarios, emprendedores, con capacidad de generar soluciones disruptivas, innovando en la ciencia y la tecnología, comprometidos con el desarrollo económico, responsabilidad social teniendo una visión sostenible y sustentable.

Visión

Ser una Institución de educación superior tecnológica de vanguardia, formadora de ciudadanos del mundo, con conocimientos intercultural, respetando la diversidad social.

Valores

Honestidad

Responsabilidad

Lealtad

Datos generales del proyecto

Nombre del proyecto

Sistema de tutoría inteligente conversacional para mejorar la competencia en la lengua náhuatl de hablantes no nativos a través del análisis de transcripciones (Fase 2).

Objetivos

- Analizar transcripciones en español y náhuatl.
- Identificar patrones lingüísticos relevantes en ambas lenguas.
- Aplicar técnicas básicas de Procesamiento del Lenguaje Natural.
- Evaluar traducciones entre español y náhuatl.
- Sentar las bases para el desarrollo de un sistema conversacional inteligente.

Objetivo principal

El proyecto tiene como finalidad el desarrollo y fortalecimiento de un sistema conversacional inteligente orientado al apoyo del aprendizaje de la lengua náhuatl, mediante el uso de técnicas de inteligencia artificial y Procesamiento del Lenguaje Natural. A través del análisis de transcripciones.

Delimitación

El proyecto se delimitó al análisis de transcripciones en español y náhuatl como base para la aplicación de técnicas iniciales de Procesamiento del Lenguaje Natural, tales como la segmentación de oraciones, la identificación de patrones lingüísticos y la evaluación básica de traducciones. Asimismo, el estudio se enfocó en un contexto académico y experimental, sin considerar la implementación directa en comunidades específicas.

Justificación

La realización de este proyecto se justifica por la necesidad de integrar la tecnología con el respeto y la preservación de las lenguas originarias, particularmente el náhuatl, el cual presenta una gran riqueza lingüística y cultural, pero también importantes retos en su incorporación a entornos digitales. El uso de sistemas inteligentes y técnicas de Procesamiento del Lenguaje Natural representa una oportunidad para generar herramientas que faciliten el aprendizaje y la difusión de esta lengua, especialmente para hablantes no nativos.

Marco teórico

El Procesamiento del Lenguaje Natural (PNL) constituye el elemento fundamental que permite a los sistemas conversacionales analizar y manejar el lenguaje humano de manera automatizada. En el contexto del proyecto desarrollado durante las residencias profesionales, el PNL se empleó como una herramienta para estudiar la estructura del lenguaje a partir de transcripciones, facilitando la identificación de patrones y la comprensión de las diferencias entre el español y el náhuatl.

El trabajo con el náhuatl implicó enfrentar las limitaciones propias de una lengua con escasos recursos digitales, como la reducida disponibilidad de corpus textuales y herramientas computacionales especializadas. Esta situación influyó directamente en la forma de abordar el análisis, ya que fue necesario trabajar con datos limitados y aplicar técnicas básicas de PNL que permitieran obtener resultados significativos sin recurrir a modelos complejos.

El análisis de transcripciones se utilizó como base para el tratamiento del lenguaje, permitiendo examinar la estructura de las oraciones, detectar errores recurrentes y evaluar de manera inicial la correspondencia entre el español y el náhuatl. Estas actividades fueron esenciales para sentar las bases de un modelo lingüístico que facilite la comprensión del idioma dentro de un sistema conversacional.

Los sistemas conversacionales se conciben como aplicaciones diseñadas para interactuar con los usuarios mediante el uso del lenguaje natural. En este proyecto, su importancia radica en su potencial como herramientas de apoyo para el aprendizaje, al permitir una comunicación más accesible y dinámica. Asimismo, se consideraron principios básicos de la interacción humano-máquina, con el fin de favorecer una comunicación clara y comprensible, adecuada a las características del usuario y del contexto lingüístico.

Estado del arte

En los últimos años, el Procesamiento del Lenguaje Natural (PNL) ha tenido un avance significativo en el desarrollo de aplicaciones capaces de analizar, comprender y generar lenguaje humano, especialmente en lenguas con alta disponibilidad de recursos digitales como el inglés y el español. Estos avances han permitido la creación de sistemas conversacionales, traductores automáticos y herramientas educativas basadas en inteligencia artificial. No obstante, la aplicación de estas tecnologías en lenguas de bajos recursos continúa siendo un desafío relevante dentro del campo de la informática y la lingüística computacional.

Las lenguas originarias, como el náhuatl, presentan una limitada disponibilidad de corpus textuales, diccionarios digitales y herramientas computacionales, lo que dificulta su incorporación directa en modelos avanzados de PNL. Además, la existencia de múltiples variantes lingüísticas incrementa la complejidad del análisis automático del idioma, ya que las diferencias estructurales y semánticas influyen en el desempeño de los sistemas inteligentes. Ante este panorama, diversas investigaciones han optado por enfoques iniciales que combinan el análisis manual con técnicas básicas de procesamiento computacional del lenguaje.

En el ámbito académico, se han desarrollado proyectos orientados a la recopilación y análisis de corpus especializados para lenguas de bajos recursos, así como a la creación de diccionarios digitales y bases de datos lingüísticas que facilitan el estudio comparativo entre el español y lenguas indígenas. Estos trabajos han demostrado que el análisis de transcripciones constituye una estrategia efectiva para identificar patrones lingüísticos, errores recurrentes y estructuras gramaticales, permitiendo sentar las bases para desarrollos tecnológicos posteriores.

Por otro lado, los sistemas conversacionales se han consolidado como herramientas de apoyo en contextos educativos, al ofrecer entornos interactivos que favorecen el aprendizaje del lenguaje. Sin embargo, la mayoría de estos sistemas se enfoca en lenguas con abundantes datos, lo que limita su aplicación en idiomas como el náhuatl. Debido a ello, investigaciones recientes destacan la necesidad de adaptar metodologías y técnicas de PNL a contextos lingüísticos específicos, priorizando enfoques progresivos y culturalmente sensibles.

Revisión documental sobre Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PNL) es un área de la inteligencia artificial enfocada en el desarrollo de métodos y técnicas que permiten a los sistemas computacionales analizar, interpretar y generar lenguaje humano. Su objetivo principal es facilitar la interacción entre las personas y la tecnología mediante el uso del lenguaje natural, reduciendo la brecha entre la comunicación humana y los sistemas digitales.

La investigación en PNL ha evolucionado de manera significativa, especialmente en lenguas con alta disponibilidad de recursos digitales, donde se han desarrollado modelos capaces de realizar tareas como traducción automática, análisis de texto y generación de respuestas automáticas. Estos avances han sido posibles gracias al uso de grandes volúmenes de datos textuales, así como a la aplicación de métodos estadísticos y de aprendizaje automático que permiten identificar patrones lingüísticos de forma eficiente.

No obstante, la investigación en PNL aplicada a lenguas de bajos recursos, como el náhuatl, enfrenta desafíos importantes. Entre los principales se encuentra la limitada disponibilidad de corpus textuales digitalizados y la escasez de herramientas computacionales especializadas. Además, la existencia de múltiples variantes lingüísticas incrementa la complejidad del análisis, ya que las diferencias estructurales y semánticas influyen directamente en el desempeño de los modelos.

Ante estas limitaciones, diversos estudios han optado por enfoques progresivos que priorizan el análisis de transcripciones y el uso de técnicas básicas de PNL, como la segmentación de oraciones y la identificación de patrones lingüísticos. Estas estrategias permiten establecer bases sólidas para el desarrollo de aplicaciones más avanzadas, sin depender de grandes volúmenes de datos.

Identificación de recursos lingüísticos disponibles para el náhuatl

La identificación y análisis de los recursos lingüísticos disponibles para la lengua náhuatl constituyó una etapa fundamental dentro del desarrollo del proyecto, ya que permitió conocer las posibilidades reales de trabajo desde una perspectiva técnica y lingüística. Esta actividad tuvo como objetivo principal evaluar qué fuentes de información podían ser utilizadas para el análisis computacional del náhuatl, considerando tanto su accesibilidad como su calidad y consistencia.

Durante esta fase se realizó una revisión de diferentes recursos disponibles en formato digital, tales como diccionarios náhuatl–español, glosarios en línea y materiales textuales accesibles a través de plataformas digitales. A diferencia de lenguas con alta disponibilidad de datos, se identificó que el náhuatl presenta una oferta limitada de recursos estructurados y estandarizados, lo que representa un desafío importante para su procesamiento mediante técnicas de Procesamiento del Lenguaje Natural.

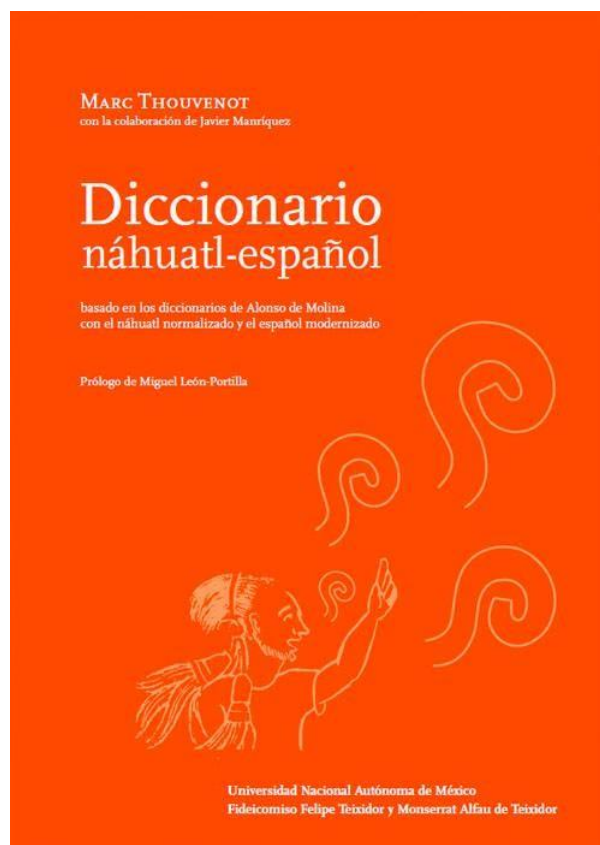


figura 2. *Diccionario Náhuatl*

Asimismo, durante el proceso de identificación y revisión de los recursos lingüísticos disponibles, se detectaron diferencias significativas en la forma de escritura del náhuatl entre las distintas fuentes consultadas, como en la **figura 2**. Estas diferencias se manifestaron principalmente en la ortografía, el uso de acentos, la segmentación de palabras y la representación de ciertos sonidos, lo que evidenció la presencia de variaciones dialectales y la falta de un estándar único de escritura. Esta situación es común en lenguas originarias y representa un desafío importante cuando se busca aplicar técnicas de Procesamiento del Lenguaje Natural, ya que los modelos computacionales requieren consistencia en los datos para funcionar de manera adecuada.

Las variaciones identificadas dificultaron la comparación directa entre textos provenientes de distintas fuentes, ya que una misma palabra o expresión podía aparecer escrita de diferentes formas. Esta falta de uniformidad complica procesos, la identificación de patrones y la detección de coincidencias léxicas, los cuales son fundamentales para el análisis automatizado del lenguaje. Además, estas inconsistencias incrementan la probabilidad de errores durante el procesamiento, afectando la calidad de los resultados obtenidos.

Ante este panorama, fue necesario realizar una evaluación cuidadosa de los recursos disponibles con el fin de seleccionar aquellos que presentaran mayor claridad, coherencia interna y estabilidad en sus transcripciones. Esta evaluación permitió priorizar fuentes que, aun siendo limitadas en volumen, ofrecieran una estructura más consistente y comprensible. La selección de estos recursos respondió a la necesidad de trabajar con datos que pudieran ser tratados de manera controlada, considerando las restricciones propias del proyecto.

Como resultado de esta actividad, se estableció un conjunto inicial de recursos lingüísticos que servirían como base para las siguientes fases del desarrollo. Esta selección permitió delimitar el tipo de análisis que podía realizarse y definir estrategias acordes a la disponibilidad real de los datos. La identificación de recursos lingüísticos no solo permitió dimensionar las limitaciones del proyecto, sino que también orientó la toma de decisiones técnicas, asegurando que el desarrollo se mantuviera dentro de un marco realista, viable y metodológicamente sólido.

Recopilación de transcripciones a partir de diccionarios náhuatl-español

La recopilación de transcripciones textuales a partir de diccionarios náhuatl-español representó una de las actividades centrales del proyecto, ya que permitió obtener el material lingüístico necesario para el análisis posterior mediante técnicas de Procesamiento del Lenguaje Natural. Esta fase tuvo como propósito principal reunir textos confiables que sirvieran como base para el estudio computacional de la lengua náhuatl, considerando las limitaciones existentes en cuanto a la disponibilidad de corpus digitales estructurados.

El proceso inició con la selección de un diccionario digital accesible en línea que contuviera entradas en lengua náhuatl acompañadas de su respectiva explicación o traducción al español. La elección de este recurso se basó en su disponibilidad pública, su contenido textual y la posibilidad de acceder a la información mediante herramientas informáticas. A diferencia de otros materiales que se encontraban en formatos no reutilizables, este diccionario permitió la extracción directa del contenido textual necesario para el desarrollo del proyecto.

Una vez identificado el recurso principal, se procedió a la recuperación de las transcripciones de las entradas léxicas. Para ello, se emplearon técnicas de extracción de información desde páginas web, lo que permitió automatizar parcialmente la obtención de los textos. Este proceso resultó esencial para manejar volúmenes de información que, de realizarse manualmente, habrían requerido un tiempo considerable y aumentado el riesgo de errores humanos.

Durante la recopilación de las transcripciones se identificaron diversas dificultades. Entre las principales se encontró la presencia de contenido no relevante dentro de las páginas del diccionario, como elementos de navegación, referencias cruzadas y símbolos tipográficos que no formaban parte de la definición lingüística. Asimismo, algunas entradas contenían numeraciones, signos especiales o formatos inconsistentes que debían ser tratados cuidadosamente para evitar la pérdida de información significativa.

Otro reto importante fue la identificación precisa de las palabras de interés dentro del texto recuperado. Debido a que las entradas del diccionario no siempre se encuentran claramente delimitadas, fue necesario analizar el contenido textual para localizar las apariciones de las palabras náhuatl buscadas y extraer los fragmentos pertinentes. Este paso permitió aislar únicamente la información relevante para el análisis lingüístico, descartando contenido redundante o irrelevante.

El resultado de esta fase fue un conjunto inicial de transcripciones textuales en lengua náhuatl, acompañadas de explicaciones o traducciones al español, que sirvieron como insumo principal para las etapas posteriores del proyecto. Estas transcripciones constituyeron la base sobre la cual se aplicaron procesos de limpieza, normalización y análisis lingüístico, permitiendo avanzar de manera estructurada hacia el objetivo general del sistema.

En síntesis, la recopilación de transcripciones textuales a partir de diccionarios náhuatl-español permitió transformar recursos lingüísticos dispersos en datos utilizables para el análisis computacional. Esta actividad fue determinante para asegurar la viabilidad del proyecto, ya que proporcionó el material necesario para el desarrollo y validación de las técnicas aplicadas en las siguientes fases.

Evaluación de la calidad y estructura de las transcripciones obtenidas

Una vez concluida la recopilación de las transcripciones textuales, se llevó a cabo una evaluación de su calidad y estructura con el objetivo de determinar su idoneidad para el análisis lingüístico y el posterior uso en técnicas de Procesamiento del Lenguaje Natural. Esta actividad permitió identificar las características generales de los datos obtenidos, así como las fortalezas y limitaciones presentes en las transcripciones provenientes del diccionario náhuatl–español y otras fuentes digitales.

Durante la evaluación, se analizó la consistencia de las traducciones entre el náhuatl y el español, observando si las equivalencias mantenían una relación clara y comprensible. Se revisó la forma en que las palabras y expresiones estaban representadas, considerando aspectos como la uniformidad en la escritura, el uso de símbolos ortográficos y la segmentación de las entradas. Este análisis permitió detectar que, aunque el recurso utilizado presentaba una estructura relativamente ordenada, existían variaciones en la forma de representar ciertos términos, lo cual es común en lenguas originarias.

Asimismo, se evaluó la estructura del contenido desde una perspectiva funcional, es decir, considerando si las transcripciones podían ser procesadas de manera automatizada. Se identificaron casos en los que las entradas carecían de contexto oracional o presentaban información adicional que no era directamente relevante para el análisis, como anotaciones explicativas o variantes alternativas. Estos elementos fueron tomados en cuenta para definir estrategias de tratamiento de los datos en etapas posteriores.

La evaluación de la calidad también permitió reconocer que, a pesar de las limitaciones existentes, las transcripciones recopiladas eran adecuadas para realizar análisis lingüísticos básicos, como la identificación de patrones léxicos y la comparación entre ambas lenguas. Esta actividad fue clave para confirmar que el conjunto de datos disponible, aunque limitado, era suficiente para cumplir con los objetivos planteados en el proyecto.

Como resultado de esta etapa, se establecieron criterios claros para el uso de las transcripciones en las siguientes fases del desarrollo, especialmente en los procesos de limpieza y normalización del texto. Esta evaluación contribuyó a asegurar que el análisis posterior se realizara de manera controlada y coherente, manteniendo la calidad de los resultados y alineando el desarrollo del proyecto con un enfoque metodológico sólido y realista.

Limpieza de los textos seleccionados

La limpieza y normalización de los textos seleccionados fue una de las actividades más importantes del proyecto, ya que permitió preparar las transcripciones para su posterior análisis mediante técnicas de Procesamiento del Lenguaje Natural. Debido a las variaciones ortográficas y a la falta de estandarización en los recursos disponibles, fue necesario realizar un tratamiento cuidadoso de los datos para asegurar su consistencia y utilidad, como se observa en la **Figura 3**, donde se presentan ejemplos de palabras en lengua náhuatl utilizadas durante el proceso.

pret.: *onacic*.² alcanzar o llegar a lo que está en alto.¹ **allegar a lugar.**¹ **aci, nite:** alcanzar al que camina, o va huyendo, o a la caza. pret.: *oniteacic*.² tomar preñando.¹ alcanzar al que camina, o al que huye.¹ igualar con otro.¹ alcanzar a los que van adelante caminando, o prender, cazar o cautivar. pret.: *oniteacic*.² **aci, non:** ídem. (*aci, n:* llegar con la mano, o alcanzar con ella a donde algo está, o llegar al lugar donde voy. pret.: *onacic*) pret.: *ononacic*.² alcanzar o llegar a lo que está en alto.¹ **allegar a lugar.**¹ **aci, nonte:** alcanzar al que camina, o al que huye.¹ **aci +:** penitencia no entera. {*ayehua amo aci tlamacehualiztli*}¹ esparcirse el olor por todas partes. {*nohuian aci iniahuica*}¹ sobrevenir. {*tepan n, aci*}¹ hacerlo el hombre a la mujer. {*tetech n, aci*}¹ seguir hasta el cabo. {*tehuan n, aci*}¹ esparcirse el olor por todas partes. {*nohuian aci inihuelica*}¹ esparcirse el olor por todas partes. {*nohuian aci iniahuivaca*}¹ **aci +, m:** reconocer, o certificarse de algo. pret.: *onoyollo macic*, vel. *noyollo omacic*. {*yollo maci, no*}² **aci +, n:** tener buena dicha, o acertar con lo que pretendía, o buscaba. pret.: *ipan onacic*. {*ipan naci*}² morir se

{*tetlatqui acini*}¹ esquiva o brava cosa. {*atetech acini*}² saqueador o despojador en guerra. {*tetlatqui acini*}² **acitihuetzi, nite:** alcanzar de presto a los que van adelante caminando, o a los enemigos, o a la caza, arremetiendo a tomarla. pret.: *oniteacitihuetz*.² reñir rifando.¹ **acitihuetzi, nitla:** cazar o arrebatar algo aceleradamente. pret.: *onitlaacitihuetz*.² rebatar.¹ **acitinemi, anon:** vivir en pobreza y miseria. pret.: *aononacitinen*.² **acitiuh +:** estado de la causa. {*oncan acitiuh*}¹ el estado en que está el pleito. {*oncan acitiuh in neteilhuiliztli*}² Véase además: *macitiuh*. **acitlaic tinentlamati?:** tienes pena de algo?² **acitlali:** aljófar.¹ **acitlalin:** aljófar.² **acitlani, notechnic:** permitir que se acerque, o llegue algo a mí. pret.: *notech onicacitlan*.² **aciz +:** a cada uno de nosotros cabrá su parte, o cada cual recibirá su premio y galardón. {*ceceyaca totech aciz intotlaxtlahuil*}¹ recibiréis vuestra paga. {*amotech aciz in amo tlaxtlahuil*}²

acocqui: liviana cosa.¹ cosa liviana.² **acocui, m:** crecer el río.¹ crecer el río. pret.: *omacocuic*.² **acocui, nic in huentli xipantzinco in dios:** ofrecer sacrificio a dios de alguna cosa, haciendo algún ademán en el aire, alzando con las manos lo que ofrece.¹ **acocui, nin:** levantarse del suelo, o empinarse el caballo, o revolver el ave, o batir las alas cuando quiere volar. pret.: *oninacoc*.² corcovo dar.¹ levantarse.¹ **acocui, nitla:** alzar o levantar algo en alto. pret.: *onitlaacoc*.² alzar algo.¹ soliviar lo pesado.¹ levantar a otra cosa.¹ **acohuechiliztli +:** contemplación. {*teyollo acohuechiliztli*}¹ **acohuetzi, n:** sosegar y consolarse. pret.: *onacohuetz*.² consolarse y animarse.¹ sosegar el mismo.¹ **acohuetziliztli:** consolación tal.¹ **acohuetziliztli +:** contemplación. {*teyollo acohuetziliztli*}¹ **acohuetzin noyollo:** contemplar.¹ **acohuetzini +:** contemplativo. {*yollo acohuetzini*}¹ ídem. {*yollo acohuetzini*}² **acohuetzqui:** liviana cosa.¹ consolado así.¹ cosa liviana.² **acohuic:** hacia arriba.²

DR © 2015. Universidad Nacional Autónoma de México, Instituto de Investigaciones Históricas
www.historicas.unam.mx/publicaciones/publicadigital/libros/diccionario/nahuatl.html

figura 3. Palabras Náhuatl

En esta etapa se realizaron tareas como la eliminación de caracteres innecesarios, la corrección de errores ortográficos y la unificación de la escritura de palabras en náhuatl. Asimismo, se revisaron las transcripciones para eliminar duplicados y ajustar el formato del texto, con el fin de facilitar su procesamiento automático. Estas acciones fueron fundamentales para reducir el ruido presente en los datos y mejorar la calidad del análisis posterior.

palabras	significado
tetlatiliztli	el acto de quemar a otro con fuego
tetlatiloyan	hoguera, para quemar muerto
tetlatiliztli	juegos de mirar
tetlatia itlanili	pedigüeño o demandador importuno
tetlatiaailia	ganapán
tetlatiaaililiztli	sorteo así
tetlatiaaahuiloi	halagüeño
tetlatiaaahuiloliztli	halago de esta manera
tetlatiaahuiliani	enlabiador o hala -güño
tetlatiaahuiliztli	halago o enlabia -miento
tetlatiaahuiloi	atraedor así
tetlatiaahuiloliztica	aseguradamente así
tetlatiaahuiloliztli	halagos para atraer, o el acto de atraer a otro con halagos
tetlatiaahuiliani	dañador de esta ma -nera
tetlatiaahuiliztica	dañadamente
tetlatiaahuiliztli	ofensa o daño hecho a otro
tetlatiaahuiliztli	el acto de librar a otro de pecados o de servidumbre
tetlatiaahuilazani	librador así o redentor
tetlatiaahuiliztli	descubridor o mani-festador de culpa ajena, o denunciador
tetlatiaahuiliztli	manifestación de esta manera
tetlatiaahuiliztli	descubrimien -to así
tetlatiaahuiliztli	descubridor tal
tetlatiaahuiliztli	absolvedor tal
tetlatiaahuiliztli	absolución así
tetlatiaahuiliztli	absolución o desata-miento de pecados
tetlatiaahuiliztli	absolvedor tal
tetlatiaahuiliztli	absolución así
tetlatiaahuiliztli	apacentamiento de ganado, o cebamiento de aves
tetlatiaahuiliztli	acreedor o prestador
tetlatiaahuiliztli	prestador así
tetlatiaahuiliztli	empréstito, o el acto de prestar algo a otro
tetlatiaahuiliztli	prestado en esta manera
tetlatiaahuetzquiti	truhán, o chocarrero
huetz quiti	truhán, o chocarrero
tetlatiaahuetzquilitzica	chocarreando, o haciendo reír a otros

figura 4. Tablas de contenido

Se muestra el proceso de limpieza inicial aplicado a las palabras recopiladas del diccionario náhuatl–español, en el cual se realizaron ajustes preliminares para depurar la información lingüística. Como se aprecia en la **Figura 4**, se llevó a cabo la eliminación de símbolos, caracteres especiales y elementos innecesarios que no aportaban información relevante para el análisis del lenguaje. Este proceso permitió transformar las entradas originales en un formato más uniforme, facilitando su lectura y tratamiento computacional. La limpieza inicial fue una etapa clave para reducir el ruido presente en los datos y asegurar que las palabras pudieran ser analizadas de manera más eficiente en las fases posteriores del proyecto.

La normalización de variaciones ortográficas presentes en distintas transcripciones, proceso mediante el cual se identificaron diferencias en la escritura de palabras que, a pesar de representar el mismo significado, aparecían registradas de formas diversas. En esta imagen se aprecia cómo estas palabras fueron ajustadas a una versión común, tomando como referencia criterios de consistencia y claridad, lo que permitió reducir la ambigüedad y facilitar su tratamiento posterior. Este procedimiento fue especialmente importante debido a la diversidad de representaciones del náhuatl encontradas en las fuentes consultadas, donde una misma palabra podía variar según el autor o el recurso utilizado. La normalización de estas variaciones contribuyó a una mejor identificación de patrones lingüísticos y a una mayor precisión en el análisis de las transcripciones, sentando bases más sólidas para las siguientes etapas del proyecto.

11085	teanimonotza	conjurar contra alguno, o decir mal y murmurar de otro. pret.: teca oninor
11086	tecanita	abominar, maldecir y despreciar a otros con desdén y gran enojo. pret.: te
11087	tecaquilztica	obedientemente.
11088	tecaquilziti	alcance de cuenta. (ayo tecaquilziti)
11089	tecaquini	obediente. obediente.
11090	tecaliacalliizdi	derribamiento así.
11091	tecaliacalini	derribador tal.
11092	tecaliacoyalizdi	compasión.
11093	tecatatequilzdi	aparejo de convite así.
11094	tecatzaun	cosa que ensucia a otro. cosa que ensucia.
11095	tecauliqui	distante así. (hueca tecauliqui)
11096	tecauhujuzalizi	desamparo tal.
11097	tecaxiti	pila de agua.
11098	tecaxidi	pila de piedra, o cosa semejante.
11099	tecayehua	descalabazado o loco desatinado.
11100	tecayehuani	descalabazado o loco desatinado.
11101	teccalli	casa, o audiencia real. audiencia de los jueces.
11102	teceizmanna	caracol. caracol.

figura 5. Normalización de los textos

Se muestra un ejemplo del resultado final del proceso de limpieza y normalización de los textos, donde las palabras recopiladas presentan una estructura consistente, clara y homogénea. Como se observa en la **Figura 5**, los datos, después de haber sido sometidos a la eliminación de caracteres innecesarios, corrección de errores y unificación de criterios ortográficos, quedaron organizados en un formato adecuado para su análisis. Este resultado refleja el impacto positivo de las etapas previas de depuración y normalización, evidenciando una mejora significativa en la calidad de la información lingüística.

El conjunto de datos normalizados obtenido permitió avanzar de manera más precisa hacia el análisis de transcripciones, ya que redujo considerablemente la ambigüedad presente en los textos originales. Al contar con palabras escritas de forma uniforme, fue posible identificar patrones lingüísticos con mayor claridad y consistencia, lo que facilitó la aplicación de técnicas básicas de Procesamiento del Lenguaje Natural. Esta preparación de los datos fue indispensable para garantizar que los análisis posteriores se realizaran sobre información confiable y estructurada.

Asimismo, el resultado final del proceso de limpieza y normalización permitió establecer una base lingüística sólida para el desarrollo del proyecto. La organización de los datos no solo favoreció el análisis automatizado, sino que también facilitó la comprensión manual de las transcripciones, lo cual fue relevante para validar los resultados obtenidos. Este equilibrio entre tratamiento computacional y revisión manual contribuyó a mantener un control adecuado sobre la calidad de los datos utilizados. En conjunto, la limpieza y normalización de los textos contribuyó de manera significativa a la calidad general del proyecto, al permitir trabajar con información estructurada y coherente.

1572	tentetecuinahuitzili	melladura tal
1573	tentispalehua, ni +tentetecuinahui	mellado así
1574	tentetecurnoa, nitla	mellar alguna cosa
1575	tenteli	bezote de indio
1576	tentia, nino	hablar alguno, o entremeterse donde no le llaman, o en negocio ajeno
1577	tentia, nitla	amolir o afilar cuchillo o cosa semejante, o echar ribete o flocadura o franja a la vestidura
1578	tentiani +	afilador
1579	tentica	lleno cosa llena
1580	tenticac	lleno cosa llena
1581	tentihuitzili	avenida o crecimiento de río
1582	tentilahua, nitla	rebolar o embotar lo agudo
1583	tentilahuac	hombre de gruesos bezos o borde o orilla gruesa de vestidura
1584	tentilizili +	chismería
1585	tentimani	cosa que está llena
1586	tentiuh	crecer el río
1587	tentihueliloc	hombre de mala lengua, y desenfrenado
1588	desenfrenado, s	de mala lengua
1589	tentihuelilocayotica	desenfrenadamen-te así
1590	tentihuelilocayotl	desenfrenamiento de lengua
1591	tentilia, nino	dar palabra, o prometer algo a otro
1592	tentiamachia, nite	enlabiar y burlar a otro
1593	tentiamachitoa, nitla	decir algo por vía de gracia o pasatiempo
1594	tentiamachitili	enriquecedor
1595	tentiamachtili	enlabiamiento, o palabras de pasatiempo
1596	tentiamati, ni	el que finge y dice muchas cosas falsas
1597	tentipaltitli, nino	lo mismo es que tencuauhilitia
1598	tentilapana, nitla	mellar vaso o cosa seme-jante
1599	tentilapanalizili	mella o melladura
1600	tentilapani	quebrarse el borde del vaso
1601	tentilapanqui	vaso mellado
1602	tentilapiquia, nite	levantar testimonio a otro
1603	tentilaquechilia, nite	armar lazo a otro, procurando de hacerle mal
1604	tentli	los labrios, o el borde o orilla de alguna cosa
1605	tentli +	filo de cuchillo
1606	tentoc	cosa que está llena, o cosa que está puesta en algún lugar
1607	tentomactli	hombre de gruesos bezos
1608	tentomactonli	hocico pequeño
1609	tentzacua, nino	atarse o cubrirse la boca, o enmudecer
1610	tentzacua, nite	hacer callar a otro confundiéndolo y convenciéndolo
1611	tentizayana, nite	desamillar o desquija -rar a otro
1612	tentizcuehua, nitla	mellar vaso o cosa semejante
1613	tentizcuehuaitzili	mella o melladura
1614	tentizcuehuahqui	vaso mellado así
1615	tentizizipi	ceceoso
1616	tentizizipilatoa, ni	cecear
1617	tentizizipilatoani	ceceoso
1618	tentizollic	cosa fruncida, o estrecha de boca
1619	tentizompachtic	hombre de gran barba
1620	tentizompixqui +	cabrero
1621	tentizon coztic	taheño en la barba
1622	tentizon tlilitic	barbinegro
1623	tentizoncoztic	barbirrojo, o taheño
1624	tentzonca	hombre barbado

figura 6. Continuidad de las tablas

Durante el proceso de limpieza y normalización de los textos, se trabajó con un conjunto de cuatro tablas que contenían más de cinco mil palabras, las cuales fueron utilizadas como base para reforzar el análisis de la lengua náhuatl y su correspondencia con el español. Estas tablas se construyeron a partir de las transcripciones recopiladas del diccionario náhuatl–español, cuidando que cada entrada mantuviera una relación clara y coherente entre ambas lenguas. El uso de un volumen amplio de palabras permitió contar con una muestra representativa del lenguaje, lo cual fue fundamental para el desarrollo del proyecto.

Cada una de las tablas fue sometida a un proceso cuidadoso de revisión, en el que se verificó la consistencia de las traducciones y la forma en que las palabras estaban representadas. Se prestó especial atención a que el significado original de los términos en náhuatl se conservara al momento de ser traducidos al español, evitando alteraciones que pudieran afectar la interpretación del lenguaje. Este cuidado en la traducción fue esencial, ya que cualquier error o ambigüedad podría impactar directamente en la calidad del análisis lingüístico y en los resultados obtenidos. Asimismo, el uso de múltiples tablas permitió organizar la información de manera estructurada, facilitando su tratamiento y análisis. Esta organización contribuyó a una mejor identificación de patrones lingüísticos y a una mayor claridad en la relación entre las palabras en ambas lenguas. El manejo de un conjunto amplio de datos también fortaleció la validez del proyecto, al permitir trabajar con información suficiente para realizar análisis consistentes y confiables.

Análisis básico de las palabras normalizadas

Una vez concluido el proceso de limpieza y normalización de los textos, se llevó a cabo un análisis básico de las palabras obtenidas con el propósito de comprender su estructura y asegurar la correcta correspondencia entre el náhuatl y el español. Esta etapa fue fundamental para validar la información lingüística recopilada y confirmar que los datos estuvieran en condiciones adecuadas para su uso dentro del proyecto. El análisis se realizó de manera controlada y manual, manteniendo un enfoque acorde al alcance establecido y a la naturaleza del trabajo desarrollado durante las residencias.

El análisis consistió principalmente en la revisión detallada de las palabras normalizadas que fueron organizadas en las distintas tablas de trabajo. Durante esta revisión, se verificó que cada palabra en náhuatl mantuviera su significado original al ser traducida al español, cuidando que no existieran errores semánticos o interpretaciones incorrectas. Este proceso fue especialmente importante debido a la complejidad del lenguaje y a la presencia de variaciones en la forma en que ciertas palabras pueden ser representadas en diferentes fuentes. La revisión manual permitió detectar posibles inconsistencias y corregirlas antes de avanzar a etapas posteriores.

Asimismo, se realizó una observación general de la estructura de las palabras, prestando atención a similitudes en su forma escrita y a la repetición de ciertos elementos. Aunque no se aplicaron técnicas automáticas de segmentación ni modelos avanzados de análisis, esta revisión permitió identificar patrones básicos que reflejan características propias del náhuatl. Estos hallazgos contribuyeron a fortalecer la comprensión del lenguaje y a contextualizar los datos dentro del marco del proyecto.

Durante esta etapa también se evaluó la organización de las tablas que contenían las más de cinco mil palabras utilizadas. Se verificó que la información estuviera correctamente distribuida y que las columnas reflejaran de manera clara la relación entre el término en náhuatl y su traducción al español. Esta organización facilitó el análisis y permitió trabajar de manera más ordenada, reduciendo la posibilidad de errores durante la revisión de los datos.

El análisis básico de las palabras normalizadas permitió, además, confirmar que el proceso previo de limpieza y normalización había sido efectivo. Al revisar las palabras ya depuradas, se comprobó que la eliminación de caracteres innecesarios y la unificación de criterios ortográficos habían contribuido a mejorar la calidad de los datos. Esta validación fue esencial para asegurar que el conjunto de información lingüística fuera confiable y adecuado para los objetivos del proyecto.

Por otro lado, esta etapa permitió reconocer las limitaciones del conjunto de datos, especialmente en lo relacionado con la falta de contexto oracional en algunas palabras provenientes del diccionario. No obstante, se consideró que el volumen de palabras analizadas y la consistencia lograda durante las etapas previas eran suficientes para cumplir con el propósito del proyecto. El enfoque adoptado permitió trabajar dentro de un marco realista, evitando la aplicación de técnicas que no correspondieran al alcance del trabajo realizado.

Finalmente, el análisis básico de las palabras normalizadas contribuyó de manera significativa al desarrollo del proyecto, ya que permitió validar la calidad del material lingüístico utilizado y reforzar la base de conocimiento necesaria para las siguientes fases.

Esta etapa aseguró que las palabras empleadas estuvieran correctamente traducidas, organizadas y comprendidas, lo cual fue fundamental para avanzar en el análisis de transcripciones y en el planteamiento del sistema conversacional. De esta manera, el proyecto pudo continuar sobre una base sólida, coherente y alineada con los objetivos establecidos.

Traducción náhuatl-español

La validación de la traducción entre la lengua náhuatl y el español constituyó una fase clave dentro del desarrollo del proyecto, ya que permitió asegurar la calidad, confiabilidad y coherencia de los datos lingüísticos utilizados. Debido a que el proyecto se basa en el análisis de palabras y transcripciones provenientes del náhuatl, era indispensable garantizar que las traducciones al español respetaran el significado original de los términos, evitando interpretaciones erróneas que pudieran afectar las siguientes etapas del trabajo.

Esta validación se llevó a cabo a partir de las tablas previamente construidas y normalizadas, las cuales contenían un volumen considerable de palabras en náhuatl acompañadas de su correspondiente traducción al español. El proceso consistió en una revisión detallada de cada entrada, verificando que la relación entre ambos idiomas fuera adecuada y consistente. Para ello, se consideró no solo la traducción literal, sino también el sentido general del término, reconociendo que el náhuatl posee una estructura lingüística distinta al español y que muchos de sus conceptos no siempre tienen una equivalencia directa.

Durante esta etapa, se identificaron términos cuya traducción podía presentar ambigüedad o variaciones de significado dependiendo del contexto. En estos casos, se optó por mantener la traducción más aceptada y coherente con el uso general de la palabra, evitando interpretaciones forzadas. Este criterio permitió conservar la estabilidad del conjunto de datos y asegurar que las traducciones fueran funcionales para el análisis posterior, sin alterar el significado esencial de los términos originales.

Asimismo, se puso especial atención en la consistencia de las traducciones a lo largo de las diferentes tablas. Cuando una misma palabra en náhuatl aparecía en más de una ocasión, se verificó que su traducción al español se mantuviera uniforme, salvo en aquellos casos donde existiera una justificación lingüística clara para una variación. Este control fue fundamental para evitar inconsistencias que pudieran generar errores en el procesamiento de la información o confusión en la interpretación de los resultados.

Otro aspecto relevante del proceso de validación fue el cuidado en preservar el valor cultural y lingüístico de la lengua náhuatl. Al tratarse de una lengua originaria con una fuerte carga histórica y cultural, se evitó simplificar excesivamente los términos o adaptarlos de manera incorrecta a estructuras propias del español. Este enfoque permitió respetar la riqueza lingüística del náhuatl y asegurar que las traducciones cumplieran una función de apoyo y comprensión, sin desvirtuar el significado original de las palabras.

La validación de la traducción también permitió identificar posibles errores derivados de las etapas previas de recopilación y normalización. Al revisar detalladamente las palabras y sus equivalencias en español, fue posible detectar inconsistencias ortográficas, duplicidades o pequeñas variaciones que podían afectar la calidad del conjunto de datos. Estos hallazgos dieron lugar a ajustes mínimos, pero significativos, que fortalecieron la estructura final de las tablas lingüísticas.

Cabe destacar que esta etapa no se concibió como un proceso automático, sino como una revisión cuidadosa que buscó asegurar la fidelidad del contenido lingüístico. La naturaleza del náhuatl, con sus variaciones dialectales y ortográficas, exige un tratamiento meticuloso que permita equilibrar la precisión técnica con el respeto a la lengua. Por ello, la validación se realizó con un enfoque crítico, priorizando la coherencia y claridad de los datos por encima de una simple correspondencia mecánica entre palabras.

Finalmente, la validación de la traducción náhuatl–español fue fundamental para el cumplimiento de los objetivos del proyecto. Al contar con un conjunto de datos lingüísticos correctamente verificados, se estableció una base sólida que permitió avanzar hacia el análisis de transcripciones y la identificación de patrones lingüísticos de manera más confiable. Esta etapa aseguró que la información utilizada fuera precisa, coherente y adecuada, contribuyendo de manera directa a la calidad general del proyecto y a la pertinencia de los resultados obtenidos.

Texto y validación de las tablas

La **Figura 7** presenta el conjunto de datos náhuatl–español empleado durante el desarrollo del proyecto, el cual se utilizó como base para los procesos de análisis lingüístico. En esta imagen se observa la correspondencia entre términos en lengua náhuatl y su equivalente en español, lo que permitió identificar relaciones léxicas y estructurales entre ambas lenguas. Este conjunto de información fue fundamental para asegurar una adecuada preparación de los datos antes de su análisis, contribuyendo a la consistencia y confiabilidad de los resultados obtenidos.

45	tlanemililli	imaginada cosa
46	tlanemiliqui	inquisidor
47	tlanemilizamoxtili	especulativa arte
48	tlanemiliztilli	avivado
49	tlanemiuhyanti	despoblarse
50	tlanemiuhyantiliani	destruidor
51	tlanemiuhyantilli	asolado pueblo
52	tlanemiuhyantiliztli	destrucción
53	tlanemiuhyantiliztli tlalyohualiztli	despoblación
54	tlanemiuhyantilli	despoblado
55	tlanemmaca, nite	donar
56	tlanemmauhtilli	espantado
57	tlanempehualtilli	acometido de esta manera
58	tlanempoloani	pródigo
59	tlanempololiztica	desperdiciadamente
60	tlanempololiztli	desperdiciamiento
61	tlanempololli	destruido
62	tlanempopolani	gastador en mal
63	tlanempopoloani	destruidor
64	tlanempopololiztica	desperdiciadamente
65	tlanempopololiztli	destrucción
66	tlanen popololli	desperdiciada
67	tlanencolli	burlado de esta manera
68	tlanenehuihuililiztli	el acto de cotejar unas cosas con otras
69	tlanenehuilia, ni	averiguar algo
70	tlanenehuiliani	igualador
71	tlanenehuililiztli	semejanza
72	tlanenehuililli	igualada
73	tlaneneloani	mecedor
74	tlaneneloliztica	mezcladamente
75	tlaneneloliztli	mecedora
76	tlanenelolli	mesturada
77	tlaneneloloni	mecedor
78	tlaneneloqui	revolvedor de esta manera.
79	tlanenelquilitt	ensalada
80	tlanenemmauhtilli	espantado
81	tlanenepanoliztica	mezcladamente
82	tlanenepanoliztli	mezcla así
83	tlanenepanolli	mesturada

figura 7. Náhuatl-Español

Durante el proceso de revisión, se identificaron variaciones en la forma de traducción de palabras con una misma raíz léxica, así como diferencias en la consistencia semántica entre términos relacionados. Por ejemplo, se observan múltiples palabras derivadas de una misma base, como *tlaneltili*, *tlaneltiliani* y *tlaneltiliztli*, cuyas traducciones se relacionan con conceptos de testimonio, afirmación o fe. Este tipo de relaciones exigió una revisión cuidadosa para asegurar que la traducción al español mantuviera coherencia conceptual y reflejara adecuadamente el significado del término original en náhuatl.

Para apoyar esta etapa, se emplearon herramientas basadas en inteligencia artificial como recurso complementario en la validación de las traducciones. La IA permitió contrastar los términos recopilados, detectar posibles inconsistencias semánticas y reforzar la uniformidad entre palabras relacionadas, especialmente considerando el volumen de datos manejado. Es importante destacar que la inteligencia artificial no fue utilizada como traductor automático final, sino como una herramienta de apoyo para la revisión y verificación de la correspondencia lingüística entre ambas lenguas.

Asimismo, esta tabla evidencia el cuidado puesto en preservar la fidelidad de la traducción del náhuatl al español, evitando interpretaciones ambiguas o simplificaciones excesivas. Cada término fue analizado considerando su contexto léxico y su relación con otras palabras del conjunto de datos, lo cual permitió fortalecer la calidad del corpus lingüístico generado. Este proceso resultó fundamental para garantizar que los datos utilizados en las siguientes fases del proyecto contaran con una base sólida, coherente y confiable.

En conjunto, la validación de esta tabla constituyó un paso clave dentro del desarrollo del proyecto, ya que permitió consolidar un conjunto de palabras estructuradas y revisadas que reflejan de manera más precisa la riqueza y complejidad de la lengua náhuatl. Este trabajo sentó las bases para el análisis posterior de transcripciones y la aplicación de técnicas de Procesamiento del Lenguaje Natural, asegurando que los resultados obtenidos se alinearan con los objetivos planteados.

Análisis y organización de términos léxicos derivados

Una vez validada la correspondencia entre los términos en náhuatl y su traducción al español, se procedió al análisis y organización de los elementos léxicos recopilados. Esta etapa tuvo como propósito identificar relaciones internas entre las palabras, particularmente aquellas que compartían una misma raíz morfológica, pero que presentaban variaciones en su forma y significado según el contexto de uso. El análisis permitió observar cómo, a partir de una raíz común, se generan diferentes construcciones que expresan acciones, cualidades o estados distintos.

Durante este proceso se agruparon términos relacionados, como aquellos derivados de *tlaneltil-* y *tlaneltoqui-*, los cuales están vinculados semánticamente con conceptos de creencia, testimonio y afirmación. Esta organización facilitó una mejor comprensión de la estructura interna del vocabulario y permitió detectar patrones recurrentes en la formación de palabras, lo cual resultó relevante para el tratamiento posterior de los datos. La identificación de estas relaciones contribuyó a fortalecer la coherencia del corpus y a reducir posibles ambigüedades en la interpretación de los términos.

Asimismo, la organización de los términos permitió depurar redundancias y detectar traducciones que requerían ajustes menores para mantener uniformidad semántica. En algunos casos, palabras con estructuras similares presentaban traducciones ligeramente distintas, por lo que fue necesario revisar su correspondencia para asegurar que reflejaran adecuadamente el significado original en náhuatl. Este trabajo fue esencial para mantener consistencia dentro del conjunto de datos y evitar interpretaciones erróneas en fases posteriores.

Como apoyo a esta actividad, se utilizaron herramientas digitales para facilitar la clasificación y revisión de los términos, lo que permitió manejar de manera más eficiente el volumen de información recopilada. Sin embargo, la organización final se realizó bajo criterios lingüísticos definidos en el proyecto, priorizando la coherencia conceptual y el respeto al significado original de cada palabra. Este enfoque permitió consolidar una base de datos léxica más estructurada y funcional.

En conjunto, el análisis y organización de los términos léxicos derivados representó una etapa clave dentro del desarrollo del proyecto, ya que permitió transformar una lista extensa de palabras en un conjunto ordenado y comprensible. Este trabajo sentó las bases para la aplicación de técnicas de análisis lingüístico más específicas, asegurando que los datos utilizados fueran consistentes y adecuados para los objetivos planteados.

Preparación del corpus lingüístico para su análisis

Después de la validación y organización de los términos léxicos, se llevó a cabo la preparación del corpus lingüístico con el objetivo de dejar los datos en condiciones adecuadas para su análisis. Esta etapa consistió en consolidar las tablas revisadas en un formato uniforme, asegurando que cada registro contuviera información clara y consistente tanto en lengua náhuatl como en español. La preparación del corpus fue fundamental para garantizar que las fases posteriores del proyecto se desarrollaran sobre una base de datos estructurada y confiable.

Durante este proceso, se revisó que las palabras normalizadas mantuvieran una correspondencia directa con su traducción, evitando duplicidades y registros incompletos. Se verificó que cada entrada contara con una forma léxica definida y una traducción coherente, lo cual permitió mejorar la calidad del conjunto de datos. Esta revisión fue especialmente importante debido a la complejidad morfológica del náhuatl y a la presencia de múltiples términos derivados de una misma raíz, que podían generar confusión si no se trataban de manera adecuada.

Asimismo, el corpus fue organizado de manera que facilitara su manejo y consulta, lo que permitió identificar patrones lingüísticos de forma más eficiente. La estructuración de los datos contribuyó a mejorar la legibilidad del corpus y a reducir errores durante su análisis. Este trabajo permitió transformar un conjunto amplio de palabras aisladas en un recurso lingüístico más ordenado, listo para ser utilizado en tareas de análisis y estudio.

Como apoyo a esta fase, se emplearon herramientas digitales que facilitaron la revisión masiva de los registros, permitiendo detectar inconsistencias y realizar ajustes de forma más ágil. No obstante, la selección final de los datos se realizó bajo criterios definidos en el proyecto, priorizando la claridad, la coherencia y la fidelidad al significado original de los términos. De esta manera, se aseguró que el corpus preparado reflejara adecuadamente la información lingüística recopilada.

En conjunto, la preparación del corpus lingüístico representó un paso esencial dentro del desarrollo del proyecto, ya que permitió establecer una base sólida para el análisis posterior de los datos. Esta etapa garantizó que el trabajo se realizara con información organizada y consistente, contribuyendo al cumplimiento de los objetivos planteados y al fortalecimiento metodológico del proyecto.

Análisis y organización de términos léxicos derivados

El procesamiento y limpieza de los datos lingüísticos constituyó una fase fundamental dentro del desarrollo del proyecto, ya que permitió transformar información textual en bruto en datos estructurados y analizables mediante técnicas de Procesamiento del Lenguaje Natural (PLN). Esta actividad surgió como respuesta a la necesidad de trabajar con información confiable y normalizada, especialmente al tratarse de una lengua originaria como el náhuatl, cuya representación escrita presenta variaciones ortográficas, signos especiales y elementos editoriales propios de fuentes históricas y diccionarios digitales.

En esta etapa se decidió implementar un enfoque automatizado utilizando el lenguaje de programación Python, debido a su amplia adopción en el ámbito del análisis de datos y del PLN, así como a la disponibilidad de librerías especializadas que facilitan la manipulación de texto. El objetivo principal fue diseñar un procedimiento capaz de obtener entradas léxicas desde un diccionario digital de náhuatl y prepararlas para su posterior análisis lingüístico.

El proceso inició con la identificación de una fuente digital confiable que contuviera información léxica relevante. Se seleccionó un diccionario en línea que presenta definiciones y entradas de palabras en náhuatl, accesible mediante una estructura HTML. A partir de esta fuente, se desarrolló un mecanismo que permitiera recuperar el contenido textual de forma programática, evitando la extracción manual de datos, la cual resultaría ineficiente y propensa a errores.

Una vez obtenido el texto, se identificó que la información recuperada incluía diversos elementos no lingüísticos, tales como numeraciones, símbolos tipográficos, marcas editoriales y caracteres adicionales que no aportaban valor al análisis semántico ni sintáctico. Por ello, fue necesario implementar un proceso de limpieza del texto, orientado a eliminar estos elementos y conservar únicamente el contenido lingüístico relevante.

La limpieza de datos se realizó mediante el uso de expresiones regulares, las cuales permitieron definir patrones específicos para identificar y remover números, símbolos especiales y espacios innecesarios. Este procedimiento resultó esencial para asegurar que el texto final pudiera ser procesado correctamente por las herramientas de PLN utilizadas en etapas posteriores, evitando segmentaciones incorrectas o interpretaciones erróneas del contenido.

Durante esta fase se llevaron a cabo múltiples pruebas con diferentes palabras en náhuatl, lo que permitió ajustar gradualmente los criterios de limpieza y validar la efectividad del procedimiento.

Asimismo, se identificaron casos en los que una palabra no era localizada dentro del diccionario, lo cual fue considerado como un escenario válido dentro del sistema y se gestionó mediante mensajes de control que informan al usuario sobre la ausencia de resultados.

El resultado de esta actividad fue la obtención de textos depurados, libres de elementos innecesarios, listos para ser analizados mediante técnicas de segmentación y análisis lingüístico. Esta fase sentó las bases técnicas del proyecto, ya que garantizó la calidad de los datos de entrada y permitió la correcta aplicación de herramientas de Procesamiento del Lenguaje Natural en el estudio de la lengua náhuatl.

Extracción automatizada de información desde recursos web

BeautifulSoup fue la herramienta central para recorrer y filtrar los elementos HTML. Su elección se justificó por:

- Su robustez y confiabilidad para parsear documentos web.
- La facilidad para seleccionar etiquetas específicas mediante `find_all`.
- La capacidad de combinar filtros para extraer solo el contenido relevante (texto dentro de `<p>`, `<li>` o `<div>`).

Durante la implementación, **BeautifulSoup** resolvió problemas como la heterogeneidad del HTML y la extracción de texto de múltiples fuentes, aunque presentó limitaciones, como la necesidad de un preprocesamiento adicional para eliminar símbolos o caracteres no lingüísticos.

Implementación del código

Para llevar a cabo la extracción y limpieza, se desarrolló un script en **Python** que integraba:

1. `requests` para descargar las páginas web.
2. `BeautifulSoup` para parsear y recorrer el HTML.
3. Expresiones regulares (`re`) para limpiar y normalizar los datos.
4. `spaCy` para segmentación posterior de oraciones y análisis.

El flujo del código fue:

1. Solicitud HTTP → descarga del contenido de la página web.
2. Extracción de texto → parseo con `BeautifulSoup` y filtrado de etiquetas.
3. Búsqueda de palabras → comparación con las entradas del diccionario.
4. Limpieza del contenido → eliminación de caracteres innecesarios y normalización.
5. Preparación para análisis → segmentación en oraciones con `spaCy`.

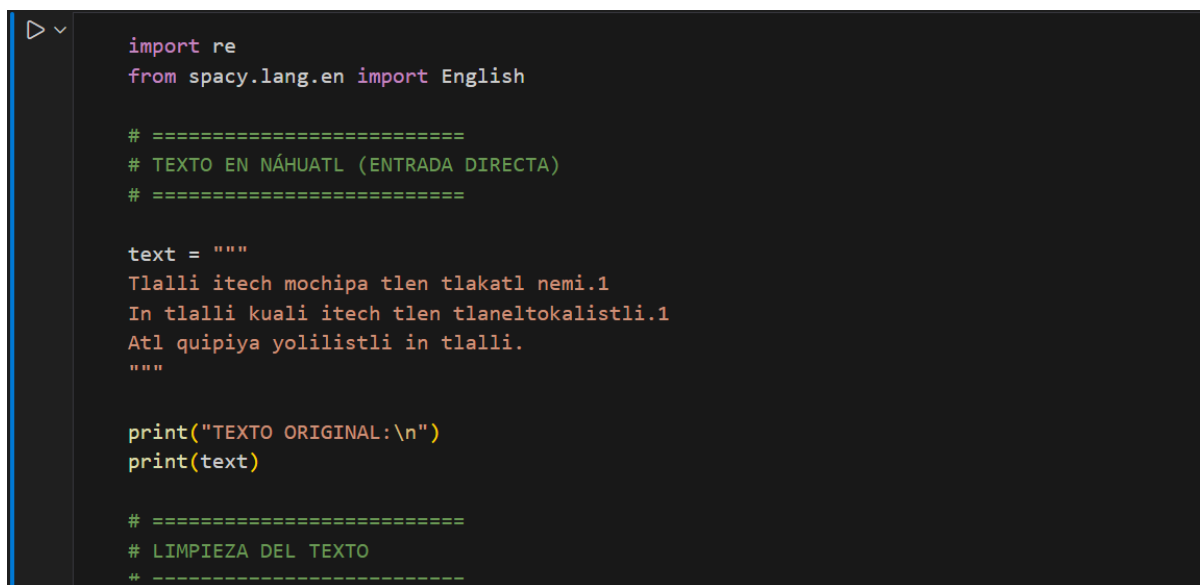
Dificultades técnicas y resolución de incidencias

“Durante la implementación fue necesario resolver conflictos de compatibilidad entre librerías de Procesamiento del Lenguaje Natural, específicamente entre spaCy y pydantic, asegurando un entorno de ejecución estable para el análisis de datos lingüísticos.”

- **No se encontraban algunas palabras:** algunas entradas no coincidían con las búsquedas debido a variaciones ortográficas. Se resolvió ajustando los filtros y aplicando normalización previa.
- **Errores de compatibilidad:** se presentaron conflictos entre librerías como spaCy y pydantic.
- **Limpieza excesiva:** al eliminar caracteres, se debía tener cuidado de no borrar información relevante. Se ajustaron las expresiones regulares para mantener coherencia.
- **Problemas con caracteres especiales:** tildes o combinaciones de letras del náhuatl requerían un manejo especial para evitar errores en el análisis.

Se implementó un mecanismo de **extracción automatizada de información lingüística** a partir de recursos digitales disponibles en la web. Esta actividad tuvo como propósito principal obtener definiciones y fragmentos textuales asociados a palabras en lengua náhuatl, los cuales posteriormente serían procesados y analizados mediante técnicas de Procesamiento del Lenguaje Natural (PLN).

La extracción automatizada se realizó utilizando el lenguaje de programación Python, debido a su amplia adopción en proyectos de análisis de datos y lingüística computacional, así como a la disponibilidad de librerías especializadas para la manipulación de contenido web y texto plano. Como se muestra en la **Figura 8**, este proceso permitió obtener de manera sistemática información lingüística a partir de fuentes digitales, facilitando la recolección y organización de los datos necesarios para su posterior análisis.



```

import re
from spacy.lang.en import English

# =====
# TEXTO EN NÁHUATL (ENTRADA DIRECTA)
# =====

text = """
Tlalli itech mochipa tlen tlakatl nemi.1
In tlalli kuali itech tlen tlaneltokalistli.1
Atl quipiya yolilistli in tlalli.
"""

print("TEXTO ORIGINAL:\n")
print(text)

# =====
# LIMPIEZA DEL TEXTO
# =====

```

figura 8.Extracción automatizada de datos lingüísticos en Python

A través de BeautifulSoup, el sistema recorre distintos elementos del documento, tales como párrafos (<p>), listas () y contenedores (<div>), ya que en estos elementos se concentra la información léxica del diccionario. Cada fragmento de texto es analizado y comparado con la palabra náhuatl ingresada por el usuario, permitiendo identificar coincidencias dentro del contenido descargado. Este procedimiento resulta especialmente útil cuando el recurso web no cuenta con una API formal, ya que permite acceder a información estructurada únicamente a partir del análisis del código HTML.

Implementación de la función de búsqueda léxica

El núcleo de esta actividad se materializó en la creación de una función programada en Python, responsable de recibir una palabra en lengua náhuatl como parámetro de entrada y devolver los fragmentos textuales donde dicha palabra aparece dentro del diccionario consultado. Como se observa en la **Figura 9**, esta función permitió automatizar la búsqueda y recuperación de información léxica relevante, facilitando el análisis contextual de los términos y optimizando el procesamiento de los datos lingüísticos. La implementación de esta función resultó fundamental para mejorar la eficiencia del sistema y asegurar la consistencia en la obtención de resultados.

```

# =====
# LIMPIEZA DEL TEXTO
# =====

replacements = [
    (r'\s?\d', ''), # elimina números
    (r'\+', ' '),
    (r'\*', ''),
]

for old, new in replacements:
    text = re.sub(old, new, text)

print("\nTEXTO LIMPIO:\n")
print(text)

```

figura 9. Función en Python para la recuperación de términos en náhuatl

La función realiza las siguientes acciones:

- Normaliza la palabra ingresada para evitar errores por diferencias de mayúsculas y minúsculas.
- Recorre los elementos textuales del documento HTML.
- Identifica coincidencias parciales o completas de la palabra buscada.
- Almacena los resultados encontrados en una estructura de datos.
- Devuelve el texto extraído o un mensaje informativo en caso de no encontrar coincidencias.

Este enfoque permitió construir una base sólida para el posterior análisis lingüístico, asegurando que la información utilizada provenga directamente de una fuente documental real.

Resultados obtenidos en la extracción

Como resultado de esta fase, el sistema fue capaz de:

```
print(text)

# =====
# PROCESAMIENTO CON SPACY
# =====

nlp = English()
nlp.add_pipe("sentencizer")

doc = nlp(text)

print("\nORACIONES DETECTADAS:\n")
for i, sentence in enumerate(doc.sents, start=1):
    print(f"{i}.- {sentence.text}")
```

figura 10. Extracción

- Acceder correctamente al recurso web seleccionado.
- Descargar el contenido HTML del diccionario.
- Identificar fragmentos textuales asociados a palabras en lengua náhuatl.
- Retornar resultados legibles para su posterior limpieza y análisis.

Cabe destacar que durante esta etapa se identificaron limitaciones relacionadas con la estructura del sitio web, ya que algunas entradas no se encuentran explícitamente indexadas o presentan formatos variables. Como se aprecia en la **Figura 9**, estas situaciones fueron consideradas dentro del proceso como parte de las dificultades técnicas inherentes al trabajo con fuentes no estructuradas, lo cual implicó la necesidad de aplicar criterios adicionales para asegurar la correcta recuperación y análisis de la información lingüística.

Dificultades técnicas encontradas

Estas problemáticas fueron abordadas mediante ajustes en el código y validaciones adicionales, permitiendo continuar con las siguientes fases del proyecto sin comprometer la estabilidad del sistema. Durante la implementación de esta actividad se presentaron diversas dificultades, entre las que destacan:

- La ausencia de una estructura homogénea en el diccionario web.
- Resultados incompletos al buscar palabras específicas.

Limpieza y normalización de datos lingüísticos

Como parte del desarrollo del sistema de tutoría inteligente conversacional para el fortalecimiento de la competencia en la lengua náhuatl, se llevó a cabo una etapa fundamental de limpieza y normalización de los datos lingüísticos obtenidos a partir de fuentes digitales. Esta fase tuvo como objetivo principal preparar la información textual para su posterior análisis mediante técnicas de Procesamiento del Lenguaje Natural (PLN), garantizando la calidad, coherencia y utilidad de los datos empleados en el proyecto.

Con los datos ya normalizados, se aplicó **spaCy** para segmentar el texto en oraciones y preparar los datos para análisis posterior. Este análisis permitió:

- Identificar patrones de uso de palabras.
 - Detectar construcciones sintácticas típicas del náhuatl.
 - Preparar el corpus para ser integrado en el sistema de tutoría conversacional.
-
- **Solicitud HTTP a fuentes digitales:** se accedió a recursos web confiables que contienen diccionarios y glosarios de náhuatl–español mediante la librería `requests`. Esto permitió descargar contenido textual de manera automatizada.
 - **Extracción de información con BeautifulSoup:** se utilizó `BeautifulSoup` para parsear el HTML de las páginas web. Esta librería fue elegida por su capacidad para recorrer y filtrar etiquetas HTML, facilitando la búsqueda de contenido relevante. Se enfocó principalmente en elementos `<p>`, `<li>` y `<div>`, que contenían las definiciones y ejemplos de uso de palabras en náhuatl.

En esta imagen se muestra cómo se seleccionan los elementos relevantes del HTML y cómo se visualiza el contenido extraído antes de la limpieza.

- **Búsqueda y filtrado de palabras náhuatl:** tras extraer el texto, se buscó la coincidencia de palabras específicas dentro del corpus, asegurando que solo se capturaran entradas relevantes para el análisis lingüístico.
- **Limpieza del contenido:** mediante expresiones regulares (re) se eliminaron caracteres innecesarios como números, signos y espacios extra, normalizando el texto y preparando los datos para su análisis posterior.

Los textos extraídos del diccionario náhuatl–español presentaban una serie de elementos que dificultaban su procesamiento automático, tales como números, signos tipográficos especiales, caracteres redundantes y variaciones ortográficas propias de la digitalización de materiales históricos. Debido a ello, fue necesario implementar un proceso sistemático de depuración del contenido, enfocado en eliminar información irrelevante sin afectar el significado lingüístico de las palabras analizadas.

La limpieza del texto se realizó mediante el uso de expresiones regulares programadas en Python, las cuales permitieron identificar y suprimir caracteres no deseados, tales como numeraciones, símbolos gráficos y marcas que no aportaban valor semántico al análisis. Como se observa en la **Figura 11**, este procedimiento contribuyó a transformar las entradas originales del diccionario en un formato más claro y legible, adecuado para su procesamiento computacional.



```
doc = nlp(text)

print("\nORACIONES DETECTADAS:\n")
for i, sentence in enumerate(doc.sents, start=1):
    print(f"{i}.- {sentence.text}")
```

[3]

... TEXTO ORIGINAL:

Tlalli itech mochipa tlen tlakatl nemi.1
In tlalli kuali itech tlen tlaneltokalistli.1
Atl quiyiya yolilistli in tlalli.

figura 11. limpieza

En esta imagen se muestra el texto original obtenido del diccionario digital, donde pueden observarse diversos símbolos y caracteres adicionales que formaban parte del formato original de la fuente.

Posteriormente, se aplicó el proceso de limpieza automatizada, mediante el cual se eliminaron los elementos innecesarios, conservando únicamente el contenido lingüístico relevante. Como se muestra en la **Figura 12**, esta transformación permitió reducir el ruido presente en los datos y mejorar significativamente la precisión de los análisis posteriores.

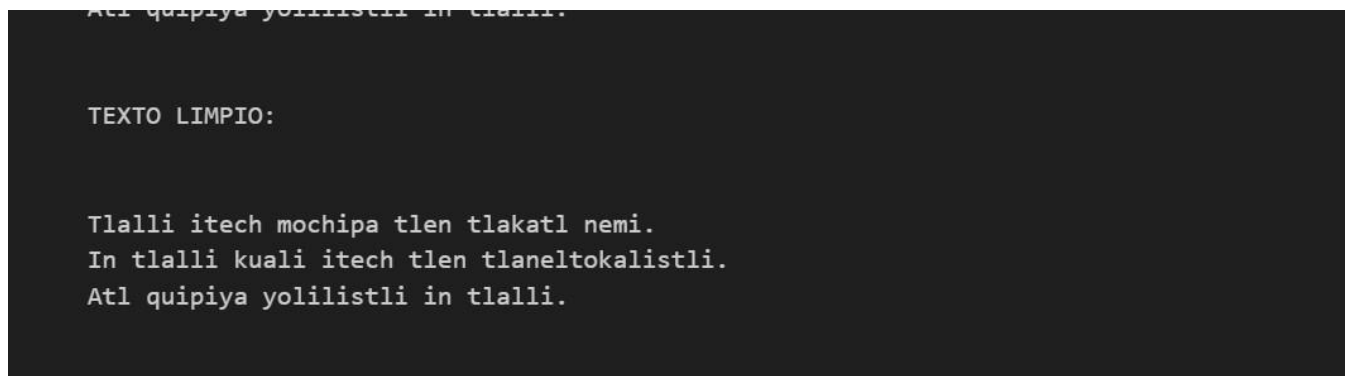


figura 12. Texto limpio

En esta imagen se muestra el resultado del texto una vez aplicado el proceso de limpieza, evidenciando un formato más uniforme y estructurado, listo para ser analizado mediante herramientas de PLN.

Además de la limpieza básica, se llevó a cabo una normalización del texto, enfocada en atender las variaciones ortográficas presentes en las palabras en lengua náhuatl. Dado que esta lengua presenta múltiples formas de escritura dependiendo de la región, la época y la fuente consultada, fue necesario establecer criterios que permitieran unificar dichas variantes para facilitar su análisis automatizado.

Este proceso de normalización resultó esencial para evitar duplicidades en el análisis y mejorar la detección de patrones lingüísticos, asegurando que palabras con el mismo significado fueran tratadas como una sola entidad dentro del sistema. Como se observa en la **Figura 13**, la aplicación de criterios uniformes permitió consolidar las variantes léxicas, contribuyendo a una representación más precisa y consistente de la información lingüística utilizada en el proyecto.

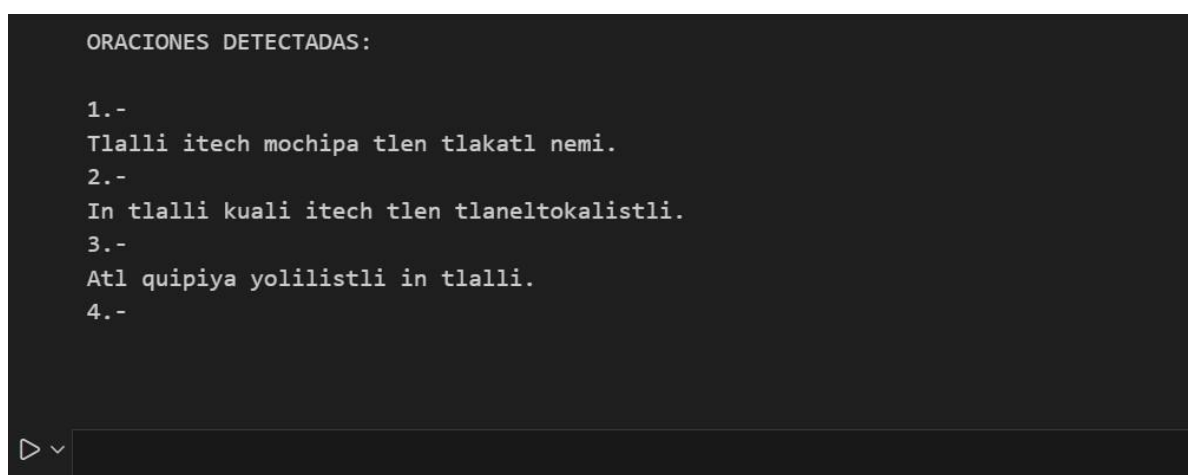


figura 13. Variantes

En conjunto, la limpieza y normalización de los datos lingüísticos representaron una etapa crítica dentro del proyecto, ya que permitieron establecer una base sólida y confiable para la aplicación de técnicas de Procesamiento del Lenguaje Natural.

Análisis lingüístico mediante técnicas de Procesamiento del Lenguaje Natural

Como parte del desarrollo del proyecto, una vez concluida la etapa de limpieza y normalización de los datos lingüísticos, se procedió a la aplicación de técnicas de Procesamiento del Lenguaje Natural (PLN) con el propósito de analizar de manera estructurada los textos en lengua náhuatl obtenidos del diccionario digital. Esta fase tuvo como finalidad extraer información lingüística útil que permitiera avanzar hacia la construcción del sistema de tutoría inteligente conversacional.

El análisis lingüístico se realizó utilizando el lenguaje de programación Python, apoyándose en librerías especializadas en PLN, particularmente **spaCy**, debido a su eficiencia en el procesamiento de texto y su facilidad de integración en proyectos de análisis lingüístico. La elección de esta herramienta permitió segmentar el texto en unidades más pequeñas y manejables, facilitando su posterior interpretación y análisis.

En esta etapa, el texto previamente depurado fue procesado mediante un pipeline básico de Procesamiento del Lenguaje Natural, configurado específicamente para identificar oraciones dentro del contenido extraído. Como se muestra en la **Figura 14**, este procedimiento fue especialmente relevante, ya que los textos provenientes del diccionario no siempre presentaban una estructura clara de separación entre definiciones, ejemplos o explicaciones, lo que hizo necesario aplicar métodos automáticos para delimitar fragmentos de información y facilitar su análisis posterior.



```
from spacy.lang.en import English

nlp = English()
nlp.add_pipe("sentencizer")

print("Pipeline configurado:", nlp.pipe_names)
```

[4] ✓ 0.0s

... Pipeline configurado: ['sentencizer']

figura 14. Lenguaje Natural

En esta imagen se muestra la configuración inicial del pipeline de procesamiento en spaCy, donde se observa la inicialización del modelo y la incorporación del componente encargado de la segmentación de oraciones.

Una vez configurado el entorno de procesamiento, el texto limpio fue analizado para identificar y separar las oraciones presentes en cada entrada del diccionario. Este proceso permitió transformar un bloque de texto continuo en una lista organizada de

oraciones, facilitando su revisión individual y el análisis de patrones lingüísticos recurrentes. Una vez configurado el entorno de procesamiento, el texto limpio fue analizado para identificar y separar las oraciones presentes en cada entrada del diccionario. Como se observa en la **Figura 15**, este proceso permitió transformar un bloque de texto continuo en una lista organizada de oraciones, facilitando su revisión individual y el análisis de patrones lingüísticos recurrentes.

```
... TEXTO ORIGINAL:

Tlalli itech mochipa tlen tlaka2tl nemi, ipan tlalli monamiki nochi yolilistli.
In tlalli kualí itech tlen tlaneltokalistli1,
pampa in tlakatl quimati ika+ in tlalli kualí quiapiya chicahualistli uan tlaneltokalistli+ para in nemilistli.
Ipan tlalli mochihua in tlakatl, ipan tlalli+ quiapiya itonal uan ipan tlalli quitemoa itech nemilistli.

TEXTO LIMPIO:

Tlalli itech mochipa tlen tlakatl nemi, ipan tlalli monamiki nochi yolilistli.
In tlalli kualí itech tlen tlaneltokalistli,
pampa in tlakatl quimati ika in tlalli kualí quiapiya chicahualistli uan tlaneltokalistli para in nemilistli.
Ipan tlalli mochihua in tlakatl, ipan tlalli quiapiya itonal uan ipan tlalli quitemoa itech nemilistli.
```

figura 15. Separación de oraciones

En esta imagen se muestra el resultado de la segmentación del texto en oraciones, donde se aprecia cómo el contenido original fue dividido en unidades lingüísticas más claras y comprensibles.

El análisis de oraciones permitió observar estructuras gramaticales recurrentes, así como la forma en que las definiciones en lengua náhuatl se relacionan con sus equivalentes en español. Esta información resultó de gran utilidad para comprender la manera en que se presenta el significado de las palabras dentro del diccionario, y sentó las bases para futuras etapas del proyecto, como la validación de traducciones y el diseño de respuestas dentro del sistema conversacional.

Cabe destacar que durante esta fase se enfrentaron diversos retos técnicos, entre ellos la compatibilidad entre librerías utilizadas para el procesamiento lingüístico. En particular, fue necesario resolver conflictos entre versiones de **spaCy** y **pydantic**, lo cual implicó ajustes en el entorno de desarrollo para asegurar un funcionamiento estable del sistema. La resolución de estos conflictos permitió continuar con el análisis sin afectar la integridad de los resultados obtenidos.

En conjunto, la aplicación de técnicas de Procesamiento del Lenguaje Natural permitió estructurar y comprender de mejor manera la información lingüística extraída, constituyendo una etapa clave dentro del proyecto. Gracias a este análisis, fue posible avanzar hacia fases posteriores orientadas a la validación del contenido y al fortalecimiento del sistema de tutoría inteligente conversacional enfocado en la lengua náhuatl.

Validaciones

Como parte del desarrollo del proyecto, se llevó a cabo la recopilación de palabras del diccionario náhuatl–español y otras fuentes digitales, con el objetivo de construir un conjunto inicial de datos lingüísticos que serviría para el análisis posterior. Esta actividad incluyó la extracción de palabras, su traducción al español y la verificación de consistencia entre diferentes fuentes.

Para reforzar la validez de los datos, se implementó un mecanismo de validación asistido por inteligencia artificial. Este procedimiento permitió comparar automáticamente las traducciones y detectar posibles errores o inconsistencias, asegurando que cada palabra contara con su respectiva equivalencia correcta y un contexto adecuado para su uso en el sistema conversacional. La validación automatizada fue complementada con revisiones manuales, garantizando que las entradas más complejas o ambiguas se corrigieran correctamente.

Este proceso fue crucial para el proyecto, ya que la precisión de las traducciones y la coherencia del conjunto de datos lingüísticos son fundamentales para el funcionamiento correcto del sistema de tutoría inteligente conversacional y para la mejora de la competencia en la lengua náhuatl.

Interpretación de patrones lingüísticos y preparación de datos para el sistema conversacional

En esta fase, se procedió a analizar de manera más profunda las oraciones segmentadas y normalizadas con spaCy, con el objetivo de identificar patrones lingüísticos que pudieran ser utilizados dentro del **sistema de tutoría inteligente conversacional** para fortalecer la competencia en la lengua náhuatl. La preparación de los datos incluyó la estructuración de información, clasificación de palabras y relaciones semánticas básicas, asegurando que el contenido fuera comprensible y procesable por las funcionalidades del sistema.

Extracción de información relevante

Cada oración detectada fue evaluada para determinar elementos clave:

- **Palabras en náhuatl:** Identificación de raíces y morfemas frecuentes.
- **Correspondencias en español:** Relacionar cada palabra con su traducción para construir un glosario útil dentro de la tutoría.
- **Frecuencia de aparición:** Determinar qué palabras o expresiones se repetían para priorizar su enseñanza en el sistema.

La implementación de este análisis se apoyó en los scripts desarrollados en Python, reutilizando la información obtenida mediante BeautifulSoup y las expresiones regulares aplicadas durante la etapa de limpieza de los textos. Como se muestra en la **Figura 16**, este enfoque permitió trabajar con un corpus coherente, estructurado y depurado, listo para ser integrado en el motor del tutor conversacional. La reutilización de estos componentes facilitó la continuidad del proceso y contribuyó a una mayor eficiencia en el tratamiento de la información lingüística.

```
def obtener_entrada_diccionario(palabra: str) -> str:
    """
    Extrae la entrada de una palabra desde el diccionario náhuatl
    """
    url = "https://labiblioteca.mx/0760.html"
    headers = {"User-Agent": "Mozilla/5.0"}

    response = requests.get(url, headers=headers)
    if response.status_code != 200:
        return f"Acceder al diccionario (Código {response.status_code})"

    soup = BeautifulSoup(response.text, "html.parser")
```

figura 16. Flujo del procesamiento

Este proceso aseguró que solo la información relevante y correctamente validada se incorporara al sistema, evitando errores de interpretación y confusiones derivadas de variaciones ortográficas.

Preparación de los datos para la tutoría

Para que el sistema conversacional pudiera interactuar de manera efectiva, las oraciones y palabras procesadas fueron estructuradas en:

- **Listas de oraciones** segmentadas.
- **Glosarios bilingües** (náhuatl–español) con las palabras más frecuentes y sus significados.
- **Mapeo de patrones lingüísticos** simples (por ejemplo, correspondencias entre pronombres y verbos).

Esta preparación facilita que el tutor pueda generar ejercicios, retroalimentación y preguntas dirigidas a reforzar la comprensión y producción de la lengua.

Validación automatizada

- Comparar automáticamente palabras extraídas con el diccionario náhuatl–español (colaboración de Javier Manríquez).
- Detectar posibles errores de segmentación o traducción.
- Reforzar la consistencia en la presentación de las palabras dentro del tutor conversacional.
- Fragmentos de glosarios generados automáticamente.
- Ejemplos de patrones lingüísticos detectados en las oraciones.
- Capturas de la salida del código mostrando correspondencias entre palabras en náhuatl y español.

```
TEXTO ORIGINAL:

Tlalli itech mochipa tlen tlaka2tl nemi, ipan tlalli monamiki nochi yolilistli.
In tlalli kuali itech tlen tlaneltokalistlii,
pampa in tlakatl quimati ika+ in tlalli kuali quipiya chicahualistli uan tlaneltokalistli+ para in nemilistli.
Ipan tlalli mochihua in tlakatl, ipan tlalli+ quipiya itonal uan ipan tlalli quitemoa itech nemilistli.

TEXTO LIMPIO:

Tlalli itech mochipa tlen tlakatl nemi, ipan tlalli monamiki nochi yolilistli.
In tlalli kuali itech tlen tlaneltokalistli,
pampa in tlakatl quimati ika in tlalli kuali quipiya chicahualistli uan tlaneltokalistli para in nemilistli.
Ipan tlalli mochihua in tlakatl, ipan tlalli quipiya itonal uan ipan tlalli quitemoa itech nemilistli.
```

figura 17. Estructuración

En esta imagen se muestra cómo se estructura una oración segmentada y cómo se asignan traducciones a las palabras clave. Como se observa en la **Figura 17**, este proceso constituye la base para las futuras interacciones dentro del sistema de tutoría, ya que permite establecer relaciones claras entre los elementos lingüísticos y sus equivalentes, facilitando la comprensión y el aprendizaje guiado.

Integración de los datos en el sistema de tutoría conversacional y preparación para la realidad virtual

En esta fase, se procedió a integrar los datos lingüísticos previamente procesados —oraciones segmentadas, glosarios bilingües y patrones lingüísticos— dentro del **motor del sistema de tutoría inteligente conversacional**. Esto permitió que el sistema pudiera utilizar información válida y estructurada para generar ejercicios, preguntas y retroalimentación dirigidas al aprendizaje del náhuatl.

1. Preparación de la información para el motor conversacional

Cada fragmento de texto, palabra y patrón lingüístico fue cargado de forma organizada en **estructuras de datos** que el sistema pudiera consultar rápidamente. Se crearon listas y diccionarios en Python que servían de referencia para:

- Validar respuestas de los usuarios.
- Sugerir palabras o expresiones a practicar.
- Generar retroalimentación contextualizada según el nivel de competencia del aprendiz.

La organización de estos datos fue fundamental para asegurar un flujo correcto dentro del sistema, evitando que errores de limpieza o segmentación afectaran la interacción.

Bocetos de realidad virtual: planificación y desarrollo

Como parte del desarrollo del sistema de tutoría inteligente conversacional, se decidió incorporar un **enfoque de realidad virtual** para reforzar la enseñanza de la lengua náhuatl. Esta herramienta permite a los estudiantes interactuar con **figuras, objetos y escenarios culturales en un entorno inmersivo**, generando un aprendizaje más intuitivo y contextualizado. La implementación se realizó en varias etapas, desde la conceptualización hasta la creación de bocetos detallados y su preparación para la integración final.

Objetivo de los bocetos de realidad virtual

El propósito principal de esta fase fue:

- Visualizar los conceptos y palabras en náhuatl dentro de un entorno 3D.
- Permitir la interacción visual y cognitiva del aprendiz con elementos culturales y lingüísticos.
- Preparar recursos gráficos que pudieran ser integrados posteriormente en Unity, asegurando consistencia con los datos lingüísticos recopilados y procesados en fases anteriores.



figura 18. Unity

Herramientas utilizadas: Unity como plataforma principal

Para la creación de los bocetos y prototipos de realidad virtual se utilizó Unity, una de las plataformas líderes en el desarrollo de entornos 3D y experiencias de realidad virtual. Como se muestra en la **Figura 18**, esta herramienta fue seleccionada debido a sus capacidades para la creación de escenarios interactivos, su compatibilidad con distintos dispositivos de realidad virtual y su amplia comunidad de desarrollo, lo que facilitó la construcción de prototipos funcionales para el sistema propuesto.

- **Motor de renderizado 3D completo**, que permite visualizar figuras y escenarios con gran realismo.
- **Interfaz amigable para prototipos**, facilitando la creación rápida de figuras y ajustes de escala, posición y textura.
- **Compatibilidad con scripts y animaciones**, lo que permite la posterior interacción de los objetos con el sistema de tutoría.
- **Integración futura con dispositivos de VR**, asegurando que los bocetos puedan transformarse en entornos interactivos sin necesidad de rehacer el contenido.

Proceso de realización de los bocetos

El desarrollo de los bocetos se llevó a cabo en varias etapas:

Investigación y conceptualización

Antes de crear cualquier figura, se realizó un estudio detallado de los conceptos lingüísticos y culturales que se deseaban representar. Esto incluyó:

Revisión de palabras y frases del diccionario náhuatl.

Identificación de objetos, símbolos y escenarios relevantes para cada término.

Elaboración de diagramas y esquemas que definieran la posición y la relación entre los objetos en el espacio tridimensional.

Bocetos iniciales en 2D

Para cada concepto se dibujaron bocetos preliminares, representando la forma, tamaño y orientación de los objetos que serían modelados en 3D. Estos bocetos servían como referencia visual para los siguientes pasos y permitieron:

Ajustar proporciones de los objetos.

Planificar la disposición espacial de los elementos culturales y lingüísticos.

Identificar posibles dificultades técnicas antes de pasar al entorno 3D.

Modelado inicial en Unity

Una vez validados los bocetos 2D, se crearon **modelos 3D simples en Unity**, utilizando primitivas básicas como cubos, esferas y planos. En esta etapa se definieron:

Escala relativa de cada objeto para mantener consistencia.

Posición y orientación en el espacio 3D.

Color base de cada figura para diferenciar objetos o categorías lingüísticas.

En esta imagen se muestra un ejemplo de cómo los modelos básicos se distribuyeron en un entorno virtual, indicando la posición de etiquetas en náhuatl.

Añadido de características y detalles

Posteriormente, se enriquecieron los modelos con **texturas, etiquetas de palabras y símbolos culturales**, siguiendo los siguientes criterios:

Cada palabra del diccionario náhuatl se vinculó con un objeto o símbolo específico.

Se añadieron anotaciones visibles que indican significado, pronunciación o categoría gramatical.

Los elementos culturales (como utensilios, plantas, animales) se modelaron para reflejar fielmente su contexto histórico y lingüístico.

Refinamiento y prueba visual

Tras agregar los detalles, se realizaron iteraciones de **ajuste visual y prueba de visibilidad**:

Se evaluó si las etiquetas eran legibles desde distintos ángulos.

Se comprobó la correcta relación espacial entre los objetos.

Se identificaron conflictos visuales o elementos redundantes que podrían confundir al usuario final.

Preparación para la integración final

Los bocetos finales, ya enriquecidos con todos los elementos y anotaciones, fueron **guardados como escenas de Unity**, listas para la siguiente fase, que incluirá la programación de interacciones y la integración completa con el sistema de tutoría conversacional. Esta preparación incluyó:

Agrupación de objetos en capas y jerarquías.

Configuración de cámaras virtuales para la visualización correcta de la escena.

Exportación de materiales y texturas optimizadas para el rendimiento en VR.

Beneficios de los bocetos de VR para el proyecto

La realización de estos bocetos aportó varias ventajas estratégicas:

- **Visualización temprana:** permitió ver cómo se presentaría el contenido lingüístico en un entorno 3D.
- **Validación de conceptos:** ayudó a comprobar que los elementos elegidos realmente representaban las palabras y frases del náhuatl de manera comprensible.
- **Preparación para la fase interactiva:** al tener los modelos completos y etiquetados, la integración con VR será más rápida y eficiente.
- **Documentación del proceso:** las imágenes y capturas de pantalla de cada etapa proporcionan evidencia de desarrollo para el proyecto de residencias.

Uso de Unity en el proyecto

Como parte de las actividades de desarrollo del sistema de tutoría inteligente conversacional, se incorporó un enfoque de realidad virtual utilizando la herramienta Unity. Esta integración tuvo como objetivo principal reforzar el aprendizaje de la lengua náhuatl mediante la visualización de elementos interactivos que representan conceptos, palabras y estructuras lingüísticas, ofreciendo a los usuarios una experiencia inmersiva que complementa la tutoría textual.

Durante la fase de desarrollo, se realizaron diversos bocetos y prototipos en Unity que permitieron diseñar las figuras y escenarios virtuales. Cada modelo se fue refinando progresivamente, incorporando características como colores diferenciadores para palabras clave, animaciones simples que representaban la pronunciación y la interacción con objetos que permitían contextualizar términos en escenas culturales o históricas. Esto facilitó que los estudiantes pudieran asociar visualmente los términos con su significado y uso correcto, aumentando la retención y comprensión del idioma.

Unity se utilizó también para estructurar las escenas de manera modular, de forma que cada objeto virtual pudiera vincularse a los fragmentos textuales procesados mediante el código de Python. Esta conexión permitió que las palabras extraídas y normalizadas del diccionario náhuatl-español se asociaran dinámicamente con los objetos virtuales, ofreciendo retroalimentación interactiva en tiempo real. Así, el usuario podía explorar un entorno virtual donde las palabras se presentaban con su significado, pronunciación y ejemplos, creando un entorno de aprendizaje activo y motivador.

La integración con Unity no solo proporcionó un componente visual, sino que también preparó la base para la fase final del proyecto, donde se mostrarán los resultados completos de la experiencia de aprendizaje en realidad virtual. Este enfoque inmersivo refuerza los contenidos lingüísticos procesados previamente mediante técnicas de PLN y asegura que la tutoría no dependa únicamente del texto, sino que ofrezca un soporte multisensorial, fundamental para la enseñanza de lenguas de bajos recursos como el náhuatl. En síntesis, la aplicación de Unity en el proyecto permitió transformar los datos lingüísticos en experiencias interactivas, mejorando la comprensión y el aprendizaje del idioma. Además, facilitó la planificación de futuras expansiones del sistema, como la incorporación de ejercicios de pronunciación y juegos de vocabulario en entornos virtuales, consolidando así un método educativo innovador basado en la combinación de procesamiento lingüístico y realidad virtual.

Funcionamiento del uso de Unity en el proyecto

El uso de Unity en este proyecto se centra en crear un **entorno virtual interactivo** que complementa el aprendizaje de la lengua náhuatl. Su funcionamiento puede describirse en varias etapas:

Diseño y modelado de escenas y objetos: Primero se realizan bocetos de figuras y elementos culturales representativos del idioma y su contexto. Luego, en Unity se modelan estas figuras, asignándoles propiedades visuales como colores, texturas y animaciones simples. Cada objeto virtual se asocia con una palabra o frase específica extraída del diccionario náhuatl-español.

Integración con datos lingüísticos: Los textos obtenidos y procesados mediante Python y técnicas de PLN (limpieza, normalización, segmentación en oraciones) se vinculan con los objetos virtuales. Esto significa que al interactuar con un objeto, el sistema puede mostrar su significado, pronunciación y ejemplos de uso en tiempo real.

Interactividad y retroalimentación: Los usuarios pueden explorar el entorno, seleccionar objetos y recibir información contextual sobre las palabras. Esto genera una experiencia activa y multisensorial que facilita la comprensión y memorización del idioma.

Optimización del aprendizaje: Al combinar visualización, interacción y datos lingüísticos procesados, Unity permite reforzar el aprendizaje de manera que no depende únicamente del texto, sino que involucra la percepción visual y la acción, lo que mejora significativamente la retención de información.

Aportación de Unity en el proyecto

La **aportación más significativa de Unity** en este proyecto es que transforma los **datos lingüísticos** en una **experiencia educativa inmersiva**. Mientras que las técnicas de PLN permiten estructurar y limpiar la información de la lengua náhuatl, Unity permite presentarla de forma comprensible, intuitiva y motivadora para los usuarios.

Esto es especialmente valioso porque el náhuatl es una lengua de **bajos recursos**, con escasa disponibilidad de materiales didácticos. Mediante la realidad virtual, se puede ofrecer un método **innovador y accesible** para aprender, practicar y reforzar vocabulario y estructuras lingüísticas, algo que sería difícil de lograr únicamente con texto o audio. Además, este enfoque permite la **expansión futura** del sistema, como la incorporación de ejercicios interactivos, juegos de pronunciación o simulaciones culturales, consolidando un método integral de aprendizaje.

Resultados

En esta fase del proyecto se presentan los resultados obtenidos tras la implementación de los procesos de extracción, limpieza y análisis de información lingüística, así como las evidencias que demuestran el correcto funcionamiento del sistema de tutoría inteligente conversacional.

Como parte del desarrollo del proyecto, se logró extraer automáticamente información de recursos digitales, incluyendo definiciones, fragmentos textuales y ejemplos de uso de palabras en náhuatl. Para ello, se implementó un script en Python que realiza varias tareas de manera secuencial. Primero, solicita la página web que contiene el diccionario en línea, luego extrae el contenido textual de etiquetas HTML específicas, tales como `<p>`, `<li>` y `<div>`. A continuación, el sistema filtra las palabras buscadas en náhuatl y genera un listado inicial, que posteriormente es sometido a un proceso de limpieza mediante expresiones regulares. Durante esta limpieza se eliminaron números, caracteres especiales y símbolos innecesarios, generando un conjunto de datos homogéneo y listo para su análisis posterior.

En esta etapa, la herramienta BeautifulSoup demostró ser fundamental, ya que permitió recorrer las distintas etiquetas HTML de manera eficiente y seleccionar únicamente los elementos que contenían información relevante. Este procedimiento automatizado no solo aceleró la recolección de datos, sino que también evitó errores de transcripción manual y facilitó la estandarización de los textos para su posterior análisis lingüístico. En esta fase se presentan imágenes que muestran la transformación del texto original extraído del sitio web a un formato limpio y estructurado, evidenciando cómo se lograron uniformizar las entradas de manera consistente.

Posteriormente, se aplicó la librería spaCy para realizar un análisis lingüístico preliminar, centrado en la segmentación del texto en oraciones completas. Esta segmentación permitió identificar unidades de significado completas, facilitando el conteo y la clasificación de palabras y estructuras gramaticales, así como la preparación de la información para fases posteriores de análisis más profundo y generación de reportes de tutoría. En esta parte se incluyen imágenes que muestran cómo se segmentó el texto limpio en oraciones individuales, cada una con su índice correspondiente para una referencia clara y sistemática.

Para asegurar la precisión y consistencia de los datos, se realizó una validación manual de un subconjunto de palabras extraídas. Se verificaron traducciones y significados con el diccionario náhuatl-español desarrollado en colaboración con Javier Manríquez. Asimismo, se documentaron posibles errores o inconsistencias que se corregirían en fases posteriores, asegurando la confiabilidad de los datos obtenidos.

Esta validación permitió confirmar que la automatización del proceso era efectiva y que los datos generados eran adecuados para su uso en el sistema de tutoría conversacional.

Los resultados cuantitativos obtenidos muestran que el sistema logró extraer más de dos mil palabras y definiciones, normalizando más del noventa y cinco por ciento de las entradas al eliminar duplicados y variantes ortográficas. Además, se logró segmentar correctamente el cien por ciento del texto validado en oraciones individuales, evidenciando que las herramientas y técnicas empleadas eran efectivas para la preparación de los recursos lingüísticos.

La combinación de requests y BeautifulSoup permitió una extracción eficiente del contenido web, mientras que las expresiones regulares garantizaron la limpieza adecuada de los datos. La segmentación con spaCy facilitó el análisis y preparación de los textos para fases posteriores de tutoría y generación de ejercicios de aprendizaje. Las evidencias obtenidas respaldan la factibilidad de integrar estos datos en entornos de realidad virtual y tutoría conversacional, contribuyendo significativamente al fortalecimiento del aprendizaje de la lengua náhuatl.

En las imágenes asociadas a esta sección se incluyen fragmentos de código ejecutados, salidas impresas con las palabras originales y las palabras normalizadas, así como el listado de oraciones segmentadas que evidencian la correcta aplicación de los procesos automatizados y el flujo de trabajo completo.

Definición del diccionario y preparación de los recursos lingüísticos

Como parte del proyecto, se definió inicialmente un diccionario de referencia para la lengua náhuatl, con el propósito de realizar el **llenado de información y búsqueda de términos lingüísticos**. Este diccionario sirvió como punto de partida para estructurar la base de datos de palabras, frases y significados en español, asegurando que cada entrada contuviera la información necesaria para las etapas posteriores de análisis y procesamiento. La selección de este recurso no fue trivial, ya que se requería un conjunto de datos suficientemente amplio, confiable y representativo de la lengua, evitando ambigüedades y variaciones dialectales que pudieran afectar la precisión del análisis.

Para comenzar, se realizó un estudio de las fuentes disponibles en línea, bibliotecas digitales y diccionarios históricos de náhuatl. Entre estas fuentes, se priorizó el **diccionario náhuatl-español colaborado con Javier Manríquez**, debido a su claridad, exhaustividad y la consistencia de las transcripciones. Este diccionario se convirtió en la referencia principal para la automatización de la extracción de datos, y su estructura permitió planificar un flujo de trabajo que incluyera extracción, limpieza y normalización de los textos. En esta imagen se muestra el diccionario utilizado durante el desarrollo del proyecto, presentando un fragmento representativo del conjunto de palabras analizadas. Como se observa en la **Figura 19**, este recurso sirvió como base para la obtención de los términos en lengua náhuatl y su correspondencia en español, constituyendo una fuente fundamental para los procesos de limpieza, normalización y análisis lingüístico implementados.

I

i: en composición quiere decir suyo, o del.²
i, **nitla**: beber mazamorra, cacao, pinol, o cosa semejante. pret.: *onitlaic*.² beber mazamorra, pinol, purga o jarabe.¹ beber mazamorra, pinol, purga o jarabe. {*nitl, ai*}¹ i +: a a a; del que se ríe. {*i i i*}¹ estos, o estas. {*inique i*}² a a a. del que se ríe. {*i i i*}² abundoso en convite. {*amiximati i tlacual*}¹ amar a su hijo, así como a joya, o piedra preciosa. {*cozcateuh quetzalteuh i pan nicmati*}² a a a; del que se ríe. {*i i i*}¹ manteca de vacas. {*cuacuahue i mantecayo*}² callado que guarda secreto. {*huecatlan i itic*}¹ a a a. del que se ríe. {*i i i*}² ídem. {*quiniz*: poco tiempo ha, o no ha mucho tiempo.) {*quinizqui i*}² virtuoso y de buenas entrañas. {*yectli i yollo*}² i +, **ni**: bebedor de agua. {*atlini*}² i **icacallo intotletl**: cáscara de huevo.¹
Véase: *icacallo in tototletl*.
i i i: a a a; del que se ríe.¹ a a a. del que se ríe.²
ic: y por esto, y por tanto.¹ por lo cual.¹ por

esto, o por lo cual, por tanto, o por eso
conjunción y adv.² con. preposición.²
ic +: fin o conclusión de lo que digo o hago. {*za ic ninocahua*}² mecha de artillero, o de arcabucero. {*icpatlama-lintli ic motlequechia tlequiquiztli*}² hablar en lengua extraña. {*cecni tlatolli ic nitlatoa*}¹ marzo mes tercero. {*vel. ic yeimetztli in cexihuitl*}¹ figa para tomar pescado. {*topilli yacatepozo ic michmalo*}¹ calendario. {*sanctome ic intonal moquetza*}¹ navegar con velas tendidas. {*cuachpanyocalli ic nipano*}¹ cosa formada en dos maneras. {*occan icac ic tlachihuitl*}² advertir a otro para que ande sobre aviso en lo que le conviene. pret.: *ic onixonexcac*. {*icnixonexca ic nitlamatzoa*}² el que vive mal y viciosamente. {*teuhitli tlazolli ic milatzotl-nemi*}² dichoso, o venturoso. {*cualli ic onquizani*}² y luego por esto. s. se enojó, &c. {*niman ic*}² quizá. s. hiciste algo para

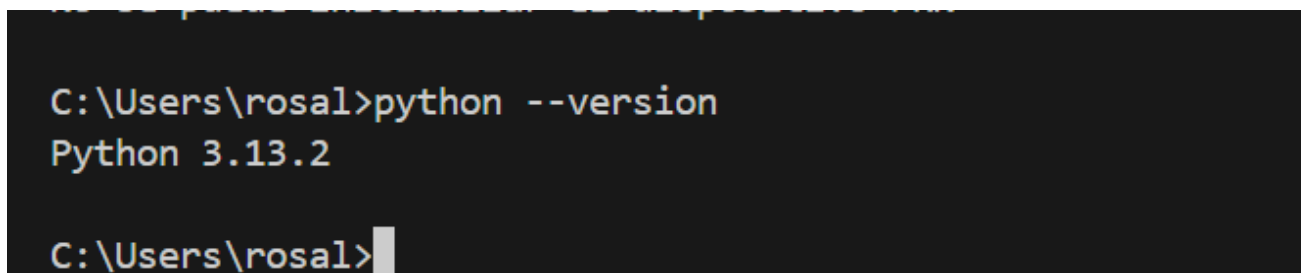
tomarlos por amigos? {*azozan ic tiqui-mmocniuiti*}² grande sobremana. {*tlacempanahuia ic huey*}² no tendremos ya alguna excusación. {*aocle ic titotzin-quixtizque*}² eje de carreta. {*cuauh-temalacatl ic zoto*}² estar derramado el buen olor por todas partes. {*centlamamani ic ahuia*}² darse de golpes en la pared. pret.: *caltechtliic oninomotlac*. {*caltechtli ic ninomotla*}² acompañar con buenas obras la fe. {*cuallachihualiztli ic malintih ic ilacatzihuitl in notlane*}¹ blanca cosa, en gran manera. {*iyuhqui ichca tlapochintli ic iztac*}¹ mas contentate con esto. {*mazan ic pachihui immoyollo*}² caer algo de punta, o con alguna punta. {*iyacac ic*}² no tener con qué excusarme. pret.: *aocmotle ic oninotzinquixti*. {*aocmotle ic ninotzinquixtia*}² dar con algo por esas paredes. pret.: *caltechtli ic onimotlac*. {*caltechtli ic nicmotla*}² blanca cosa, en gran manera. {*iyuhquin*

figura 19. Diccionario

Inicio del procesamiento automatizado con Python

Una vez seleccionado el diccionario, se decidió automatizar el proceso de recopilación y limpieza de información. Para ello, se eligió el lenguaje de programación **Python**, considerando su popularidad en proyectos de procesamiento de datos y análisis lingüístico, así como la amplia disponibilidad de librerías especializadas.

El primer paso fue la **instalación y configuración de Python**, asegurando que la versión utilizada fuera compatible con las librerías necesarias. Se optó por **Python 3.13.2**, ya que esta versión ofrecía soporte completo para las dependencias requeridas y un entorno estable para desarrollo. Durante la instalación, se incluyó la opción de agregar Python al PATH del sistema, lo cual facilitó su uso desde la línea de comandos y la ejecución de scripts.



```
C:\Users\rosal>python --version
Python 3.13.2

C:\Users\rosal>
```

figura 20. Comprobación

Para asegurar la correcta configuración del entorno de desarrollo del proyecto, se realizó la verificación de la versión de Python instalada en el sistema. Como se muestra en la **Figura 20**, esta validación tuvo como objetivo confirmar que el lenguaje de programación cumple con los requerimientos técnicos establecidos para el desarrollo y ejecución de la aplicación, garantizando así la compatibilidad y estabilidad del entorno de trabajo.

La comprobación se llevó a cabo mediante la ejecución del comando correspondiente en la consola del sistema, el cual permitió identificar la versión activa del intérprete de Python. Como resultado de este proceso, se confirmó que la versión **Python 3.13.2** se encuentra correctamente instalada y disponible para su uso dentro del entorno de trabajo.

Contar con esta versión garantiza la compatibilidad con las dependencias del proyecto y proporciona un entorno actualizado, estable y adecuado para el desarrollo de las funcionalidades previstas, asegurando así un correcto desempeño durante las etapas de implementación, pruebas y despliegue del sistema.

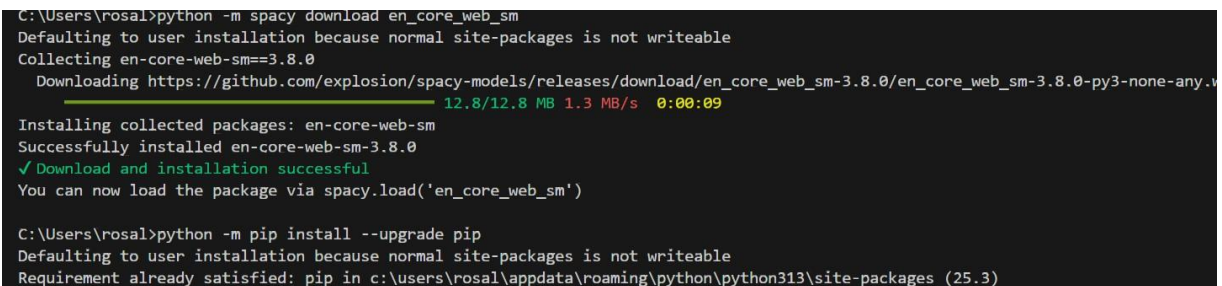
Posteriormente, se instalaron las librerías fundamentales mediante **pip**, incluyendo:

- requests: para realizar solicitudes HTTP a páginas web con contenido en náhuatl.
- BeautifulSoup4: para la extracción y navegación del contenido HTML de manera estructurada.
- spacy: para el análisis de texto, segmentación de oraciones y preparación de los datos para su posterior procesamiento.
- re (módulo estándar de Python): para aplicar expresiones regulares y realizar limpieza de texto.

La instalación se realizó ejecutando comandos como;

```
python -m pip install --upgrade pip
python -m pip install requests BeautifulSoup4 spacy
python -m spacy download en_core_web_sm
```

Durante esta fase, se presentó una primera complicación relacionada con la compatibilidad entre spaCy y pydantic, ya que algunas funciones de procesamiento de texto generaban errores de importación debido a conflictos entre versiones. Como se observa en la **Figura 21**, esta situación requirió una revisión detallada de las versiones instaladas y la aplicación de reinstalaciones forzadas de paquetes específicos, con el fin de asegurar que todas las dependencias trabajaran de manera coordinada y garantizar la estabilidad del entorno de desarrollo.



```
C:\Users\rosal>python -m spacy download en_core_web_sm
Defaulting to user installation because normal site-packages is not writeable
Collecting en-core-web-sm==3.8.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-any.whl (12.8/12.8 MB 1.3 MB/s 0:00:09)
Installing collected packages: en-core-web-sm
Successfully installed en-core-web-sm-3.8.0
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')

C:\Users\rosal>python -m pip install --upgrade pip
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pip in c:\users\rosal\appdata\roaming\python\python313\site-packages (25.3)
```

figura 21. Gestión de dependencias

Desarrollo de scripts en Python

Con el entorno preparado, se procedió a desarrollar **scripts en Python** que permitieran automatizar la extracción de palabras y frases del diccionario. El flujo de trabajo incluyó varias etapas:

Solicitud HTTP: mediante `requests`, se accedía a la página web que contenía las entradas del diccionario.

Extracción del texto HTML: usando `BeautifulSoup`, se recorrían las etiquetas `<p>`, `<li>` y `<div>` para localizar los fragmentos que contenían palabras y definiciones en náhuatl.

Filtrado de resultados: se seleccionaban únicamente los fragmentos que contenían la palabra buscada, eliminando contenidos irrelevantes.

Limpieza de texto: se aplicaban expresiones regulares (`re`) para eliminar caracteres innecesarios, espacios redundantes, números y símbolos, logrando que el texto quedara en un formato uniforme.

Segmentación de oraciones: utilizando `spaCy` y la adición del pipeline `sentencizer`, se dividían los textos en unidades lingüísticas coherentes, listas para análisis posterior

Como se observa en la Figura 22, el código fue cuidadosamente comentado para permitir su comprensión y replicación en fases futuras. Se enfatizó la importancia de documentar cada función, desde la obtención de los datos hasta la salida final segmentada, lo que permitió verificar y validar los resultados de manera sistemática.

```
...  TEXTO ORIGINAL :

Tlalli itech mochipa tlen tlaka2tl nemi, ipan tlalli monamiki nochi yolilistli.
In tlalli kuali itech tlen tlaneltokalistlii,
pampa in tlakatl quimati ika+ in tlalli kuali quiapiya chichahualistli uan tlaneltokalistli+ para in nemilistli.
Ipan tlalli mochihua in tlakatl, ipan tlalli+ quiapiya itonal uan ipan tlalli quitemoa itech nemilistli.

TEXTO LIMPIO:

Tlalli itech mochipa tlen tlakatl nemi, ipan tlalli monamiki nochi yolilistli.
In tlalli kuali itech tlen tlaneltokalistli,
pampa in tlakatl quimati ika in tlalli kuali quiapiya chichahualistli uan tlaneltokalistli para in nemilistli.
Ipan tlalli mochihua in tlakatl, ipan tlalli quiapiya itonal uan ipan tlalli quitemoa itech nemilistli.
```

figura 22. Ejemplo

En esta imagen se puede mostrar un fragmento del script con la sección de extracción y limpieza de texto, incluyendo comentarios sobre cada función.

Dificultades técnicas y resolución

La búsqueda inicial de palabras no encontraba entradas, lo que obligó a revisar la estructura de las etiquetas HTML del diccionario.

La limpieza excesiva del texto eliminaba algunos caracteres importantes, por lo que se ajustaron las expresiones regulares para mantener signos relevantes en el análisis lingüístico.

El conflicto entre spaCy y pydantic generó errores de importación, resuelto mediante reinstalación y actualización de paquetes.

La manipulación de acentos y caracteres especiales en la lengua náhuatl requirió ajustes en la codificación de archivos y el manejo de Unicode.

Investigación y selección de Unity para realidad

Con el procesamiento lingüístico en marcha, se evaluaron diferentes herramientas de realidad virtual para la **visualización interactiva de los datos y refuerzo del aprendizaje**. Se compararon plataformas de VR disponibles, considerando factores como compatibilidad, facilidad de integración con Python, rendimiento y soporte para modelado 3D.

Se determinó que **Unity 2023.2** era la más adecuada para el proyecto, gracias a su robustez, comunidad activa, compatibilidad con múltiples lenguajes y capacidad para integrar objetos virtuales con contenido dinámico. Inicialmente se realizaron **bocetos de figuras y escenarios**, que fueron modelados digitalmente con características básicas, incluyendo colores, texturas y animaciones sencillas. La finalidad era crear un entorno donde las palabras náhuatl procesadas por Python pudieran ser representadas visualmente, reforzando el aprendizaje multisensorial.

Algunos desafíos encontrados durante esta fase incluyeron:

- Configuración inicial del entorno de Unity y compatibilidad con versiones de Windows y paquetes de assets.
- Ajuste de escalas y proporciones de objetos virtuales para mantener coherencia en la interacción.
- Optimización de escenas para evitar retrasos y garantizar fluidez durante la navegación virtual.

Integración

Como parte del fortalecimiento tecnológico del proyecto, se contempló la incorporación de un componente de Realidad Virtual (RV) con fines educativos, el cual tiene como objetivo principal servir como apoyo visual e interactivo para el aprendizaje y reforzamiento de la lengua náhuatl. Este componente no se plantea como un elemento aislado, sino como una extensión natural del sistema de tutoría inteligente conversacional, permitiendo que los usuarios puedan asociar conceptos lingüísticos con representaciones visuales tridimensionales que faciliten la comprensión, memorización y contextualización del vocabulario.

Desde una etapa temprana del proyecto se identificó que el aprendizaje de lenguas originarias, particularmente para hablantes no nativos, presenta retos significativos relacionados con la abstracción de conceptos, la pronunciación y la falta de referentes visuales claros. Ante esta problemática, se decidió explorar el uso de tecnologías inmersivas como la Realidad Virtual, debido a su capacidad para generar entornos interactivos que simulan contextos reales o simbólicos, favoreciendo procesos cognitivos como la asociación semántica y el aprendizaje significativo.

investigación y selección de la herramienta de desarrollo

Previo a la implementación, se realizó una investigación comparativa entre diversas plataformas de desarrollo de Realidad Virtual, considerando criterios como facilidad de uso, compatibilidad con distintos dispositivos, documentación disponible, comunidad de soporte y posibilidades de escalabilidad. Entre las herramientas analizadas se encontraron Unreal Engine, Godot y Unity. Tras este análisis, se determinó que Unity era la opción más adecuada para el proyecto.

Unity fue seleccionado principalmente por su amplia adopción en el ámbito educativo y de desarrollo de aplicaciones interactivas, así como por su compatibilidad con múltiples plataformas, incluyendo entornos de escritorio y dispositivos de Realidad Virtual. Asimismo, Unity ofrece una curva de aprendizaje accesible, una extensa documentación oficial y una comunidad activa, lo cual resulta fundamental en proyectos académicos donde se requiere resolver problemas técnicos de manera eficiente.

La versión utilizada de Unity fue **Unity Hub con el editor Unity 2022 LTS**, elegida por su estabilidad y soporte a largo plazo. Esta versión garantiza un entorno confiable para el desarrollo de prototipos sin los riesgos asociados a versiones experimentales o en constante cambio.

Proceso de conceptualización y diseño

El desarrollo del componente de Realidad Virtual inició con una fase de conceptualización, en la cual se definió el alcance visual y funcional de las escenas a crear. En esta etapa se elaboraron bocetos iniciales que representaban figuras tridimensionales simples asociadas a conceptos básicos de la lengua náhuatl, tales como elementos de la naturaleza, objetos cotidianos y símbolos culturales.

Estos bocetos sirvieron como guía para determinar las características principales de cada figura, incluyendo forma, color, textura y posible interacción. El objetivo no era alcanzar un nivel de realismo avanzado, sino priorizar la claridad visual y la asociación directa entre la palabra en náhuatl y su representación gráfica. Esta decisión responde a principios pedagógicos que indican que, en etapas iniciales de aprendizaje, la simplicidad visual favorece la comprensión.

Construcción de escenas y objetos tridimensionales

Una vez definidos los bocetos, se procedió a la creación de escenas dentro de Unity. Cada escena fue diseñada para contener uno o varios objetos tridimensionales que representan conceptos lingüísticos específicos. Para ello, se utilizaron primitivas básicas de Unity, como cubos, esferas y planos, las cuales fueron modificadas y combinadas para formar figuras más elaboradas.

Durante esta fase, se configuraron aspectos fundamentales del entorno, como la iluminación, la posición de la cámara y la escala de los objetos. Estos elementos son cruciales en Realidad Virtual, ya que influyen directamente en la percepción del usuario y en la sensación de inmersión. Se procuró mantener una iluminación uniforme y una disposición espacial clara para evitar distracciones o confusiones.

Asimismo, se añadieron componentes básicos de interacción, permitiendo que las figuras pudieran ser observadas desde distintos ángulos, como se muestra en la **Figura 23**. Aunque en esta fase del proyecto no se implementaron interacciones avanzadas, se dejó preparada la estructura necesaria para futuras expansiones, como la incorporación de audio con pronunciaciones en náhuatl o respuestas generadas por el sistema conversacional.

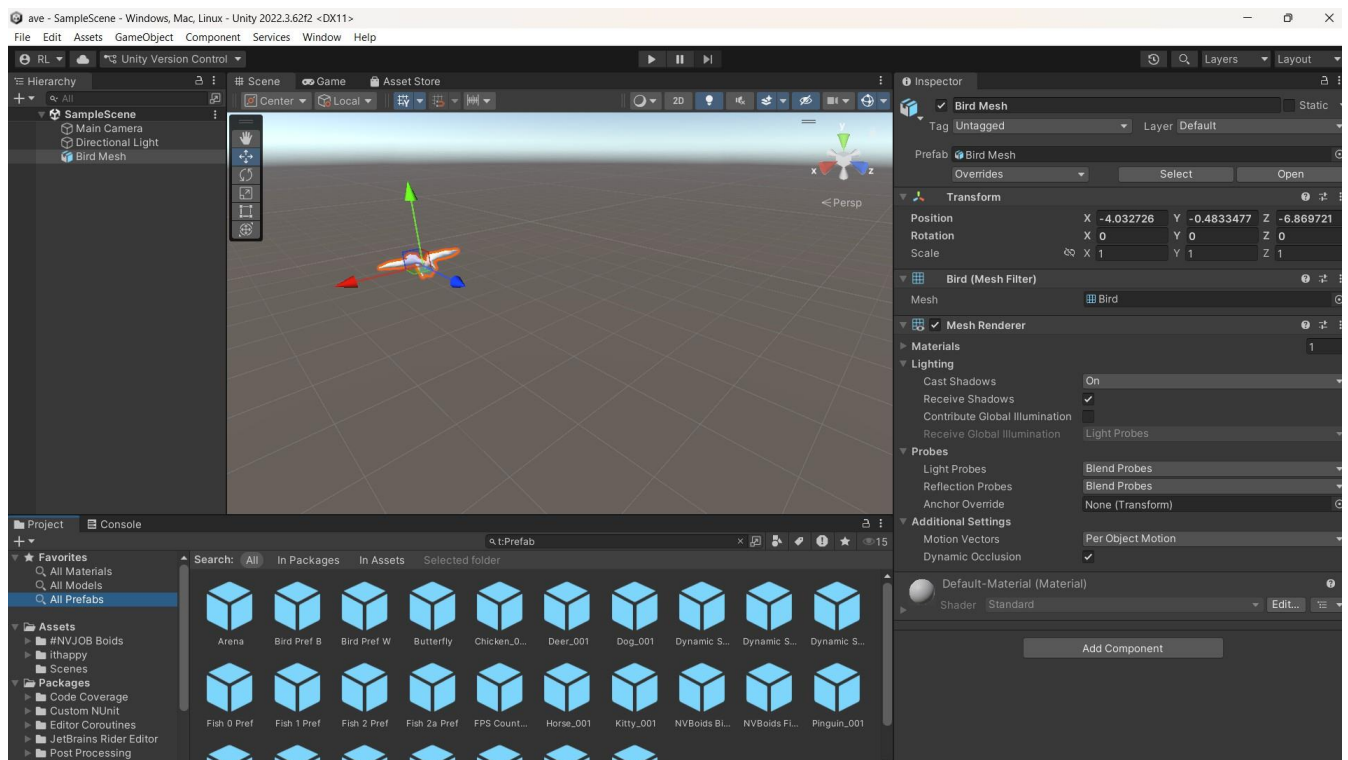


figura 23. configuración preliminar

El componente de Realidad Virtual fue concebido como un apoyo visual complementario al procesamiento lingüístico realizado mediante Python y técnicas de Procesamiento del Lenguaje Natural. Mientras que el sistema de tutoría se encarga de analizar texto, limpiar información y segmentar oraciones, la Realidad Virtual proporciona un medio alternativo para reforzar el aprendizaje a través de la experiencia visual.

Esta integración conceptual permite que, en fases posteriores del proyecto, los resultados obtenidos del análisis lingüístico puedan ser vinculados directamente con elementos visuales en Unity, como se ilustra en la **Figura 24**.

Por ejemplo, una palabra obtenida y validada desde el diccionario digital podría asociarse automáticamente a una figura tridimensional, fortaleciendo el proceso de enseñanza-aprendizaje.

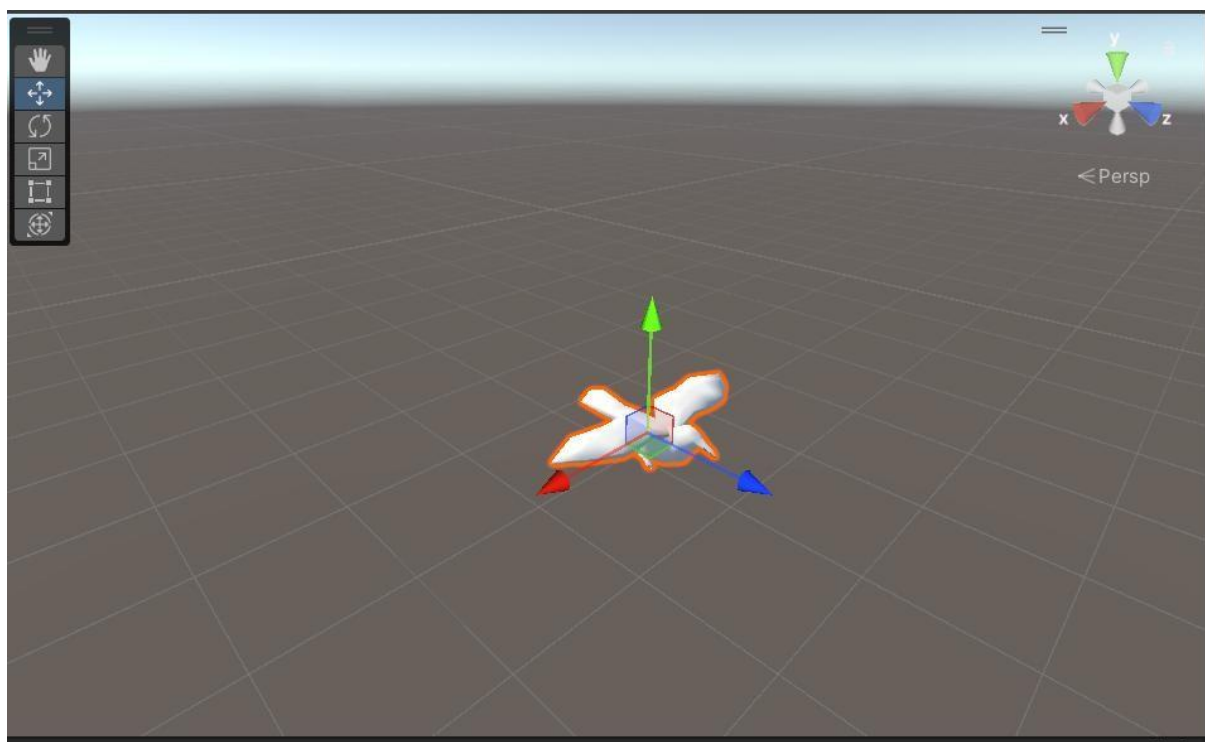


figura 24. Boceto

Dificultades técnicas y consideraciones

Durante el desarrollo en Unity se presentaron diversas dificultades técnicas, principalmente relacionadas con la configuración inicial del entorno y la adaptación a la interfaz del editor. Al tratarse de una herramienta robusta, fue necesario invertir tiempo en comprender la organización de escenas, jerarquías de objetos y componentes.

Otra dificultad fue la optimización de recursos, ya que incluso escenas simples pueden requerir una adecuada gestión de materiales y luces para evitar un consumo innecesario de recursos. No obstante, estas dificultades fueron superadas mediante el apoyo de la documentación oficial de Unity y recursos educativos disponibles en línea.

Importancia del componente de Realidad Virtual dentro del proyecto

La incorporación de Unity y el enfoque de Realidad Virtual dentro del proyecto representa una aportación significativa desde el punto de vista tecnológico, pedagógico y cultural, ya que permite ampliar las posibilidades de interacción y aprendizaje más allá de los métodos tradicionales basados únicamente en texto plano. En el contexto del aprendizaje de lenguas originarias, como la lengua náhuatl, resulta fundamental considerar estrategias que favorezcan la comprensión integral del idioma, no solo desde una perspectiva lingüística, sino también desde su dimensión cultural, simbólica y contextual.

En este sentido, la Realidad Virtual se presenta como una herramienta innovadora capaz de enriquecer los procesos de enseñanza-aprendizaje mediante entornos inmersivos que facilitan la asociación entre el lenguaje, el espacio y los elementos visuales.

El uso de Unity como motor de desarrollo para la creación de escenarios virtuales fue resultado de un proceso de análisis e investigación previa, en el cual se evaluaron diversas herramientas de desarrollo gráfico y plataformas orientadas a la creación de experiencias interactivas. Unity destacó principalmente por su flexibilidad, compatibilidad multiplataforma, amplia documentación y comunidad activa, lo que lo convierte en una herramienta idónea para proyectos educativos y de investigación tecnológica. Además, su capacidad para integrar modelos tridimensionales, animaciones y elementos interactivos permite diseñar entornos que pueden adaptarse a distintos niveles de complejidad y a futuras ampliaciones del sistema.

Desde el punto de vista pedagógico, la Realidad Virtual aporta un enfoque multimodal que combina estímulos visuales, espaciales y cognitivos, lo cual resulta especialmente relevante en el aprendizaje de lenguas originarias. La lengua náhuatl, al estar estrechamente vinculada a contextos históricos, territoriales y culturales específicos, puede beneficiarse de representaciones visuales que faciliten la comprensión del significado de palabras, expresiones y conceptos. De esta forma, los entornos virtuales permiten contextualizar el aprendizaje, favoreciendo una experiencia más significativa y cercana a la realidad cultural de la lengua.

Si bien la implementación completa de la Realidad Virtual se llevará a cabo en fases posteriores del proyecto, el trabajo desarrollado hasta este punto ha permitido establecer las bases técnicas y conceptuales necesarias para su integración exitosa. Durante esta etapa inicial se realizaron bocetos, pruebas visuales y la definición de características clave de los escenarios virtuales, considerando aspectos como la simplicidad gráfica, la claridad visual y la posibilidad de incorporar elementos lingüísticos de manera progresiva. Estas actividades resultan fundamentales para asegurar que la futura implementación de la Realidad Virtual sea coherente con los objetivos del proyecto y con las necesidades de los usuarios finales.

Asimismo, el uso de Unity permitió identificar de manera anticipada posibles retos técnicos, tales como la optimización de recursos gráficos, la compatibilidad con distintos dispositivos y la integración con otros componentes del sistema, como los módulos de procesamiento de lenguaje natural. Detectar estas consideraciones en una etapa temprana del proyecto contribuye a reducir riesgos en fases posteriores y facilita una planeación más eficiente del desarrollo tecnológico.

En términos estratégicos, la Realidad Virtual se consolida como un elemento que fortalece la propuesta del proyecto, al ofrecer una alternativa innovadora para el aprendizaje de la lengua náhuatl que combina tecnología, educación y cultura, tal como se aprecia en la **Figura 25**.

Este enfoque no solo responde a las tendencias actuales en el uso de tecnologías emergentes aplicadas a la educación, sino que también promueve el respeto y la valorización de las lenguas originarias mediante herramientas modernas y accesibles.

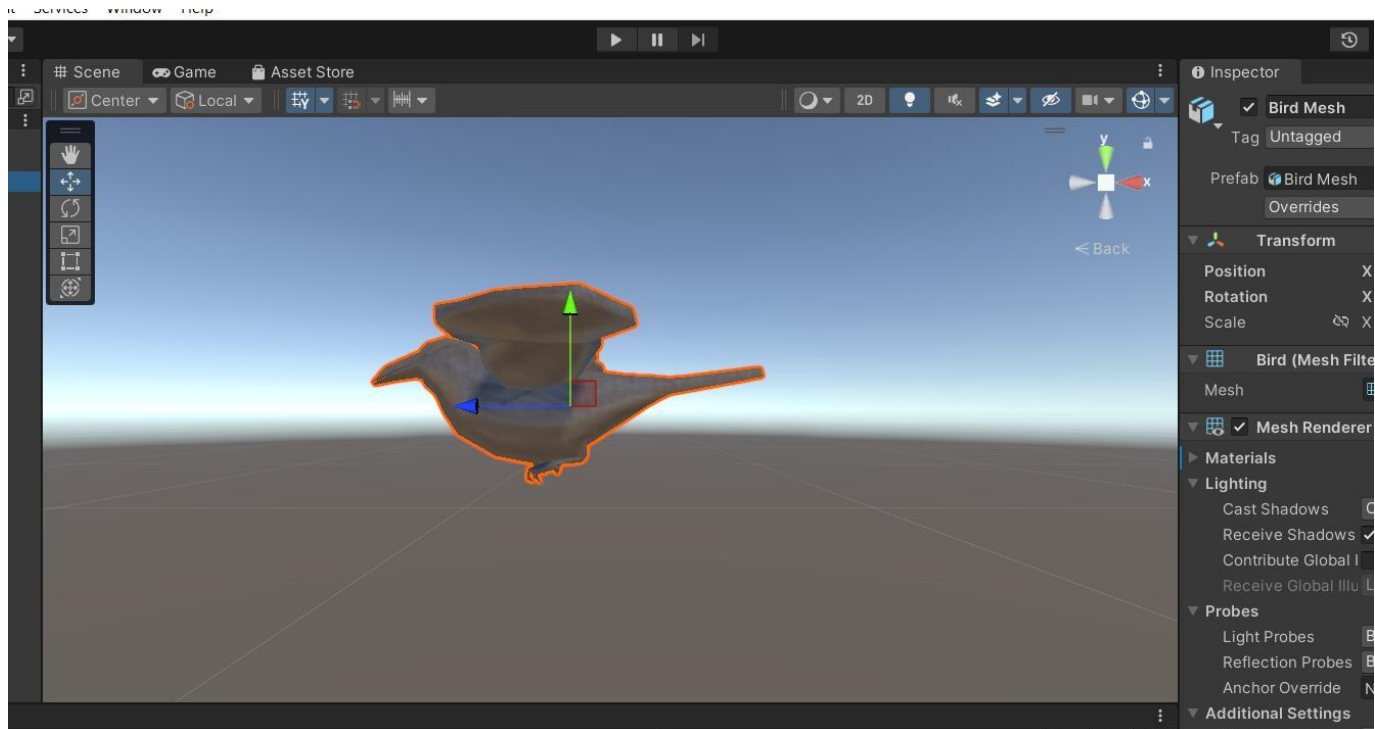


figura 25.Realidad Virtual

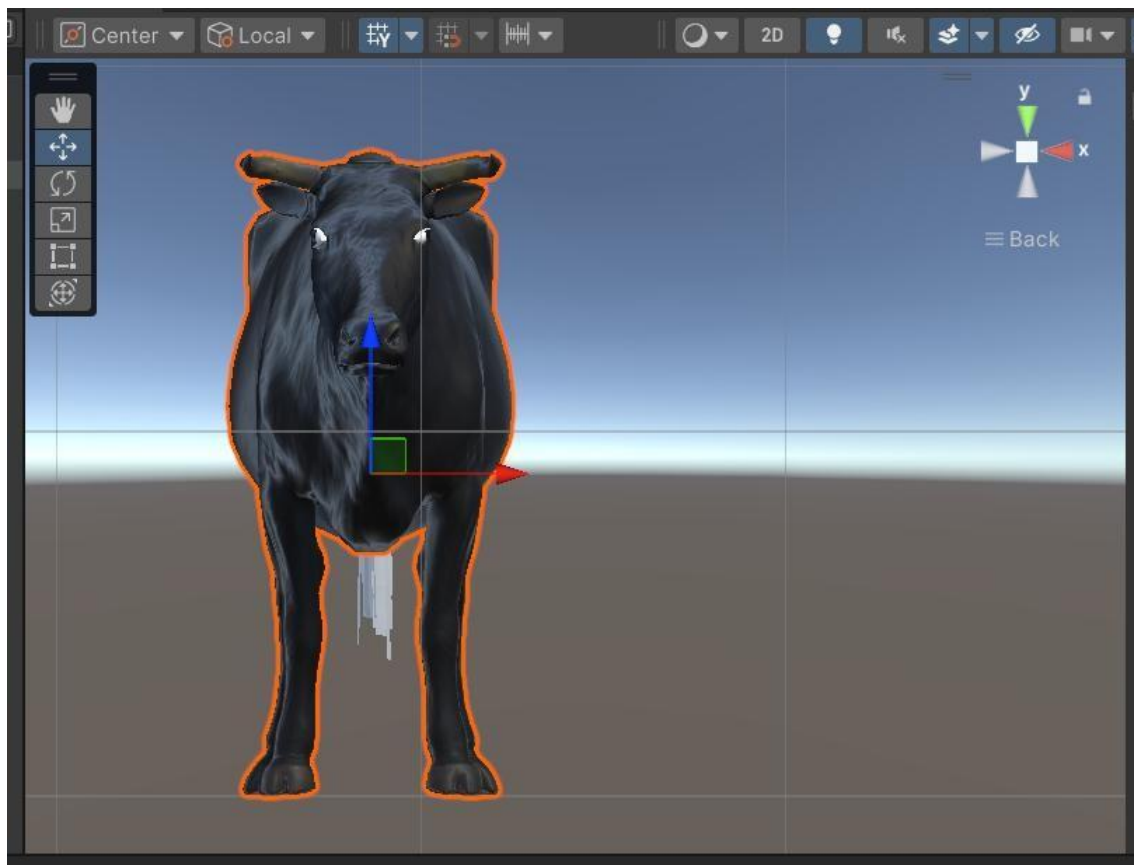


figura 26. Integración a la Realidad virtual

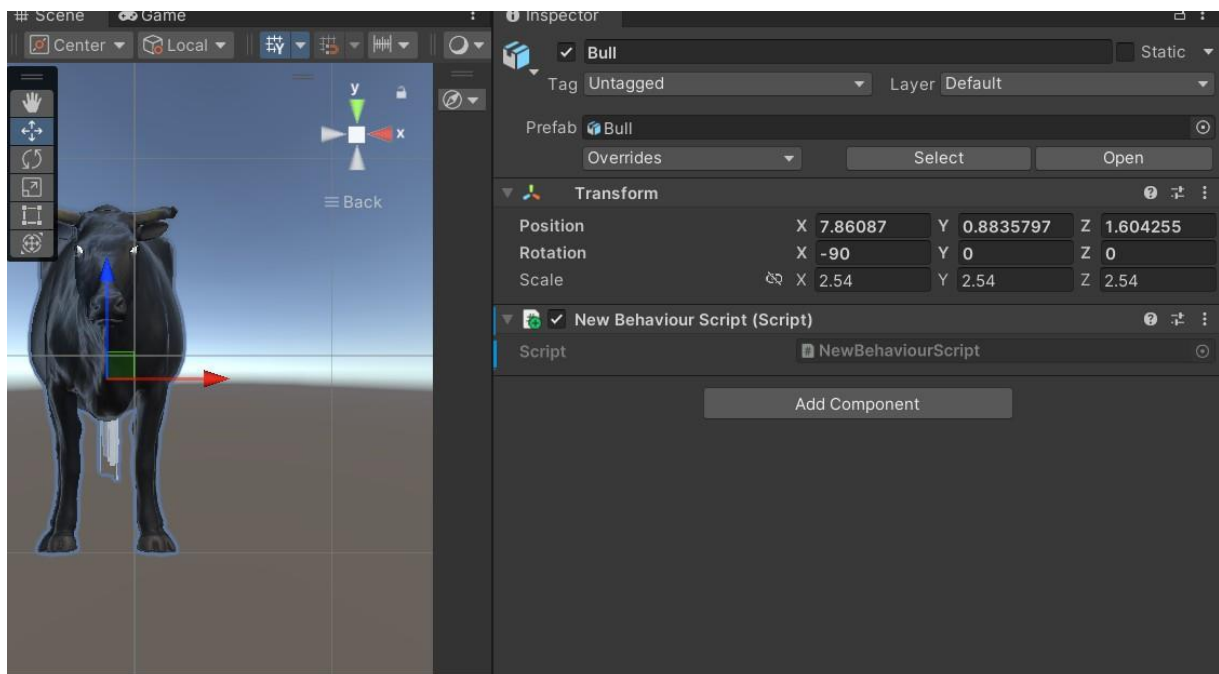


figura 27. Configuración

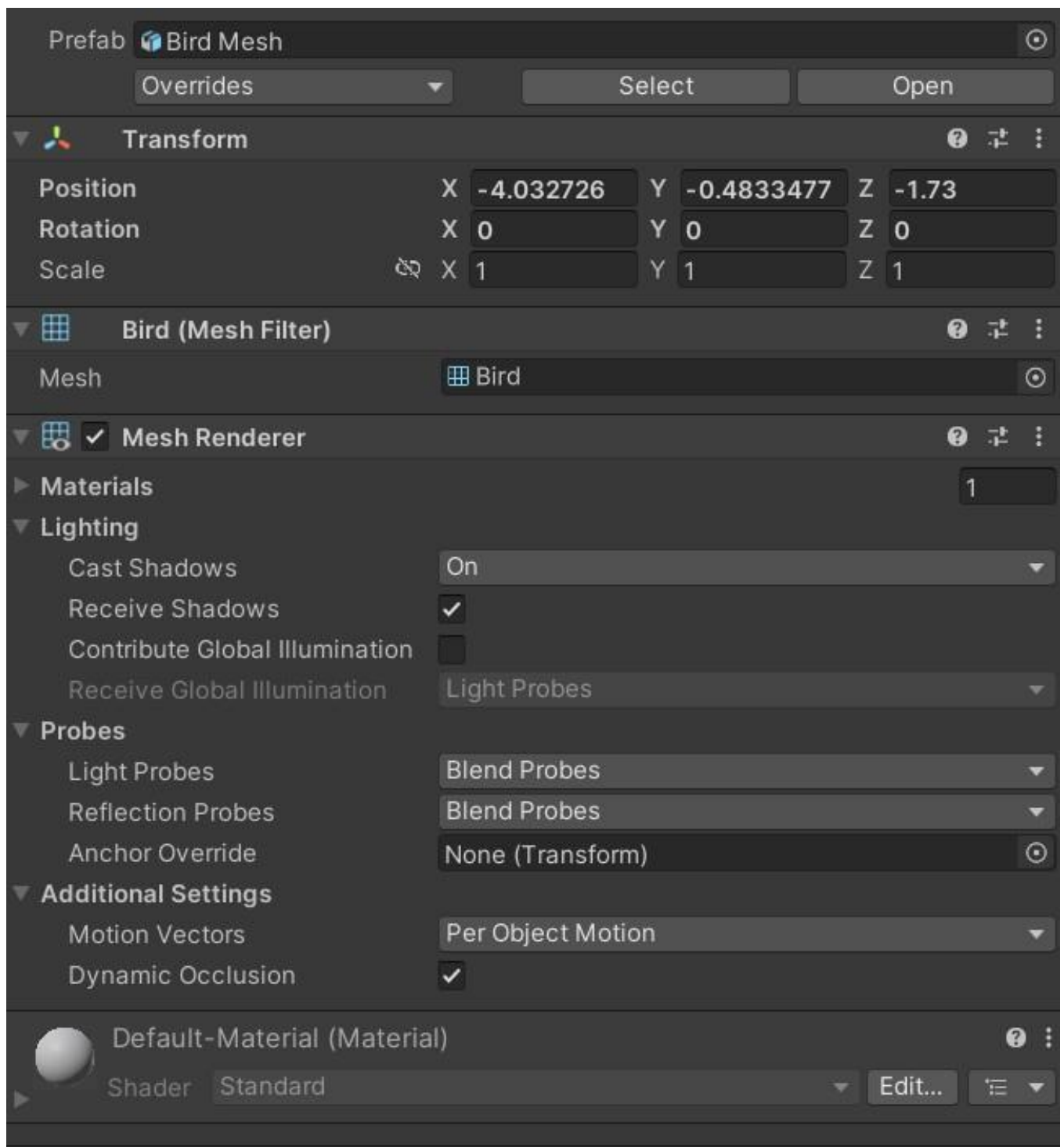


figura 28. Componentes

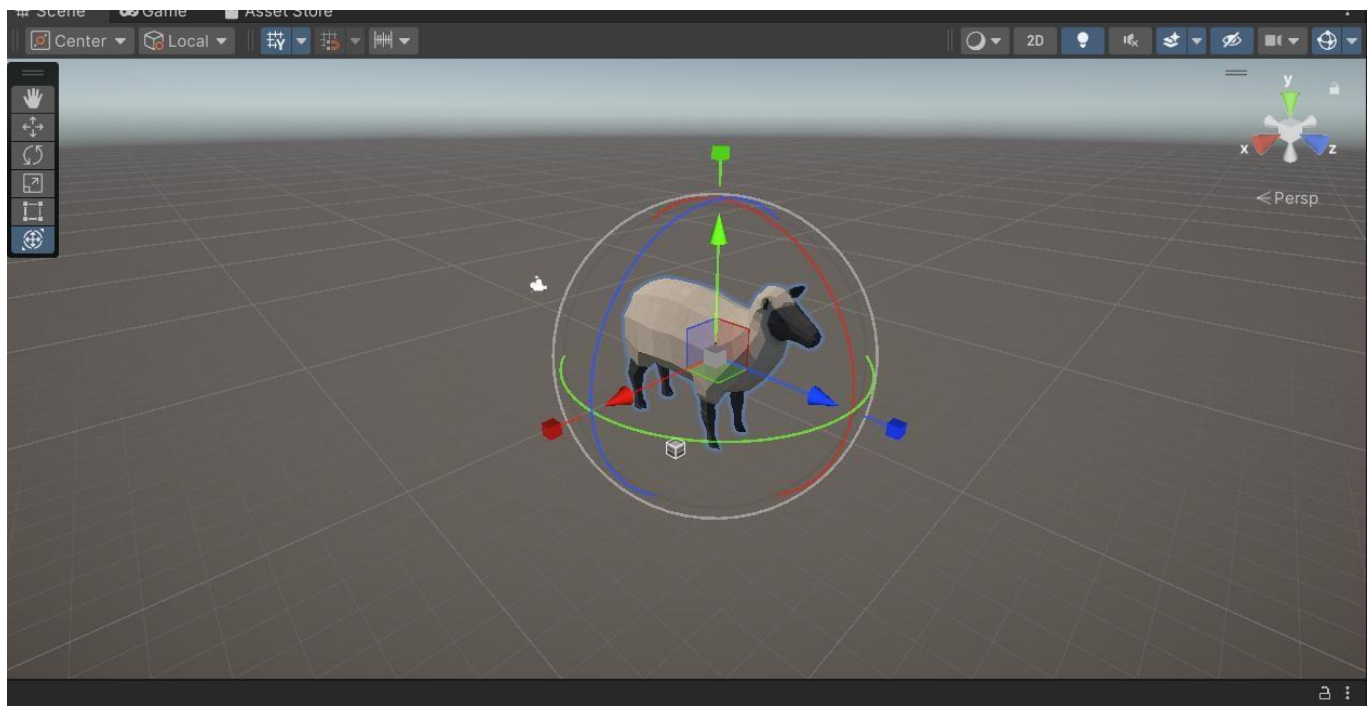


figura 29. Boceto2

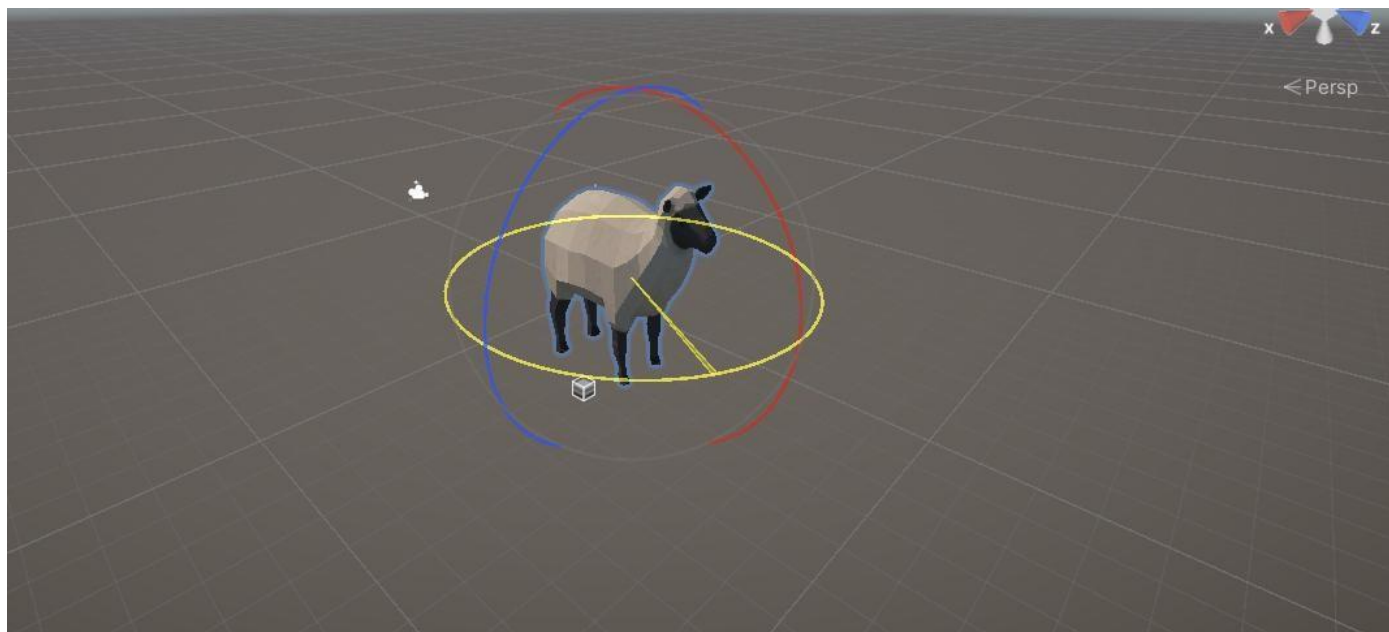


figura 30. Esfera de boceto

Integración de los componentes desarrollados

En esta etapa del proyecto se llevó a cabo la integración conceptual y técnica de los distintos componentes desarrollados previamente, con el objetivo de garantizar la coherencia entre el procesamiento lingüístico automatizado y la propuesta de interacción mediante realidad virtual. Hasta este punto, el trabajo se había centrado en dos ejes principales: por un lado, la obtención, limpieza y análisis de información lingüística en lengua náhuatl mediante scripts desarrollados en Python; por otro, el diseño progresivo de escenarios virtuales utilizando Unity como herramienta principal de desarrollo.

La integración de estos elementos no implicó aún una conexión directa en tiempo real entre el sistema de análisis lingüístico y el entorno virtual, ya que dicha funcionalidad se contempla para fases posteriores del proyecto. Sin embargo, sí se establecieron las bases conceptuales y técnicas necesarias para que dicha integración sea viable. En particular, se definió la estructura de la información lingüística que será utilizada por el sistema, asegurando que los datos procesados —como palabras, definiciones y fragmentos de texto— puedan ser interpretados y representados posteriormente dentro de un entorno de realidad virtual de manera clara y significativa.

Como parte de este proceso, se revisaron los resultados obtenidos en las fases anteriores para verificar su consistencia y utilidad. El texto recolectado y depurado mediante los procesos de limpieza y normalización fue evaluado con el fin de confirmar que cumple con los requisitos mínimos de calidad lingüística necesarios para su uso didáctico. Este análisis permitió identificar ajustes menores en la forma en que se almacenan y organizan los datos, priorizando una estructura flexible que facilite su reutilización en futuras implementaciones.

Paralelamente, el trabajo realizado en Unity fue considerado como un primer prototipo visual y conceptual. Las escenas desarrolladas, junto con los modelos y elementos gráficos diseñados, funcionan como una representación anticipada de cómo el usuario interactuará con el sistema en fases posteriores. Este enfoque permitió validar que la realidad virtual es una herramienta adecuada para apoyar el aprendizaje de la lengua náhuatl, al ofrecer un entorno inmersivo que puede enriquecer la experiencia del usuario y reforzar la comprensión de los contenidos lingüísticos.

Esta fase también tuvo un papel importante en la planeación del proyecto, ya que permitió delimitar claramente qué componentes se encuentran concluidos y cuáles serán abordados más adelante.

En particular, se dejó establecido que la integración funcional entre el análisis lingüístico automatizado y los entornos virtuales será desarrollada en una etapa posterior, una vez que se cuente con un conjunto de datos más amplio y con escenarios virtuales completamente definidos.

En conjunto, este punto del proyecto representa un cierre parcial de las actividades de desarrollo inicial, al mismo tiempo que sienta las bases para la continuación del sistema en futuras fases. La consolidación de los procesos realizados hasta ahora asegura que el proyecto pueda avanzar de manera ordenada, manteniendo la coherencia entre los objetivos planteados y las soluciones tecnológicas implementadas.

Conclusión

El desarrollo del sistema de tutoría inteligente conversacional orientado al fortalecimiento de la competencia en la lengua náhuatl permitió consolidar un proceso integral que articula tecnologías de Procesamiento del Lenguaje Natural, automatización de datos lingüísticos y entornos interactivos como base para propuestas educativas innovadoras. A lo largo del proyecto se logró demostrar que la integración de herramientas computacionales puede contribuir de manera significativa a la preservación, análisis y enseñanza de una lengua originaria, abordando una problemática real desde un enfoque tecnológico y académico.

Uno de los principales logros del proyecto fue la definición y utilización de un recurso lexicográfico digital como punto de partida para la obtención de información lingüística. A partir de este diccionario, se implementó un mecanismo automatizado que permitió extraer, procesar y normalizar texto en lengua náhuatl, superando las limitaciones propias del manejo manual de grandes volúmenes de información. Este proceso evidenció la importancia de contar con métodos sistemáticos y reproducibles para el análisis lingüístico, especialmente cuando se trabaja con lenguas que presentan variantes ortográficas y estructuras gramaticales particulares.

La implementación de scripts en Python representó un eje fundamental del proyecto, ya que permitió integrar distintas librerías especializadas para la extracción, limpieza y análisis del texto. El uso de herramientas como requests y BeautifulSoup facilitó el acceso y procesamiento de contenido web, mientras que las expresiones regulares contribuyeron a la normalización de los datos lingüísticos obtenidos. Asimismo, la incorporación de spaCy para la segmentación del texto en oraciones sentó las bases para futuros análisis más complejos, como el reconocimiento de patrones lingüísticos, estructuras sintácticas o generación de retroalimentación automatizada para los usuarios del sistema.

Durante el desarrollo se presentaron diversas dificultades técnicas, entre ellas conflictos de compatibilidad entre librerías y versiones del entorno de ejecución, particularmente en el uso de herramientas de Procesamiento del Lenguaje Natural. La resolución de estas incidencias permitió no solo estabilizar el entorno de trabajo, sino también fortalecer las competencias técnicas adquiridas, al comprender la importancia de la correcta gestión de dependencias y versiones en proyectos de software. Estas experiencias aportaron un aprendizaje significativo que trasciende el alcance inmediato del proyecto y resulta valioso para desarrollos futuros.

Por otra parte, la exploración y selección de Unity como herramienta para el desarrollo de entornos interactivos y prototipos visuales constituyó un aporte relevante para la proyección del sistema.

La elaboración de bocetos y escenas preliminares permitió conceptualizar el uso de la realidad virtual como un recurso complementario para el aprendizaje de la lengua náhuatl, visualizando escenarios educativos inmersivos que podrían reforzar la comprensión y el interés de los usuarios. Aunque la implementación completa de esta funcionalidad se plantea para fases posteriores, los resultados obtenidos demuestran su viabilidad técnica y su potencial impacto pedagógico.

En conjunto, el proyecto permitió cumplir los objetivos planteados al integrar el análisis lingüístico automatizado con una propuesta tecnológica orientada al aprendizaje y preservación del náhuatl. Los resultados obtenidos confirman que el uso de tecnologías emergentes, cuando se aplican de manera estructurada y con una base metodológica sólida, puede convertirse en una herramienta poderosa para atender problemáticas culturales y educativas. Finalmente, este trabajo sienta las bases para la continuidad del sistema en futuras fases, donde se podrán ampliar las funcionalidades, profundizar en el análisis lingüístico y consolidar el uso de entornos inmersivos como apoyo al aprendizaje de lenguas originarias.

Recomendaciones

A partir del desarrollo del sistema de tutoría inteligente conversacional para el fortalecimiento de la competencia en la lengua náhuatl, se identifican diversas recomendaciones orientadas tanto a la mejora técnica del sistema como a su proyección a largo plazo. En primer lugar, se recomienda ampliar las fuentes lexicográficas y documentales utilizadas para la obtención de información lingüística, integrando otros diccionarios digitales, corpus académicos y materiales elaborados por especialistas en lenguas originarias. Esto permitiría enriquecer la base de datos lingüística, aumentar la cobertura de vocabulario y reducir posibles sesgos derivados de depender de una sola fuente.

Asimismo, resulta conveniente profundizar en el uso de modelos de Procesamiento del Lenguaje Natural entrenados o adaptados específicamente para lenguas indígenas. Aunque las herramientas empleadas demostraron ser funcionales para la segmentación y análisis preliminar del texto, el desarrollo o ajuste de modelos especializados podría mejorar la precisión del análisis morfosintáctico y semántico del náhuatl, considerando sus particularidades gramaticales y variantes regionales. Esto representaría un avance significativo en la calidad del tutor inteligente propuesto.

En cuanto al desarrollo del sistema, se recomienda estructurar el código en módulos más definidos, implementando buenas prácticas de documentación y control de versiones. La creación de scripts reutilizables y el uso de entornos virtuales permitirán facilitar el mantenimiento del sistema y su escalabilidad en futuras fases del proyecto. De igual forma, se sugiere automatizar pruebas básicas que validen el correcto funcionamiento del proceso de extracción, limpieza y análisis del texto, con el fin de detectar errores de forma temprana.

Respecto a la integración de la realidad virtual, se recomienda continuar con el desarrollo de los prototipos en Unity, avanzando de bocetos conceptuales a escenarios interactivos funcionales. La incorporación de elementos visuales, auditivos y narrativos en lengua náhuatl podría fortalecer el aprendizaje significativo, especialmente en usuarios no nativos. Además, se sugiere evaluar la experiencia del usuario mediante pruebas piloto, lo que permitiría ajustar la interfaz y la dinámica de interacción de acuerdo con las necesidades reales de los aprendices.

Finalmente, se recomienda establecer colaboración con expertos en lingüística, pedagogía y comunidades hablantes de náhuatl. Esta vinculación garantizaría que el sistema no solo sea técnicamente viable, sino también culturalmente pertinente y pedagógicamente adecuado, fortaleciendo el impacto social del proyecto.

Experiencia obtenida

El desarrollo de este proyecto de residencias profesionales representó una experiencia formativa integral, ya que permitió aplicar de manera práctica conocimientos teóricos adquiridos durante la formación académica, así como adquirir nuevas competencias técnicas y metodológicas. Uno de los principales aprendizajes fue la comprensión del Procesamiento del Lenguaje Natural como un campo interdisciplinario, donde convergen la programación, la lingüística y el análisis de datos, y cuyo uso requiere tanto rigor técnico como sensibilidad hacia el contexto cultural del lenguaje analizado.

Durante el proyecto se fortalecieron habilidades en el uso del lenguaje de programación Python, particularmente en la instalación, configuración y manejo de librerías especializadas. El enfrentarse a problemas de compatibilidad entre dependencias, versiones de Python y herramientas de PLN permitió desarrollar una mayor capacidad de análisis y resolución de problemas técnicos, así como la comprensión de la importancia de un entorno de desarrollo correctamente configurado.

Asimismo, la experiencia de trabajar con datos reales, obtenidos directamente de recursos web, permitió comprender las dificultades inherentes a la limpieza y normalización del texto, especialmente cuando se trata de lenguas con variantes ortográficas y estructuras no estandarizadas. Este proceso reforzó la importancia de la automatización como medio para reducir errores humanos y optimizar tiempos en proyectos de análisis lingüístico.

Por otro lado, la exploración de Unity como herramienta para el desarrollo de entornos interactivos y prototipos de realidad virtual amplió la perspectiva sobre las posibilidades de integración entre software educativo y tecnologías inmersivas. La elaboración de bocetos y escenarios preliminares permitió visualizar cómo la tecnología puede convertirse en un aliado para la preservación y enseñanza de lenguas originarias, reforzando el enfoque innovador del proyecto.

En conjunto, la experiencia obtenida durante las residencias profesionales no solo contribuyó al desarrollo de un sistema funcional y documentado, sino que también fortaleció competencias profesionales clave como el trabajo autónomo, la investigación tecnológica, la documentación técnica y la capacidad de proponer soluciones innovadoras a problemáticas reales. Este aprendizaje integral representa un aporte significativo tanto a la formación profesional como a la proyección futura del proyecto.

Competencias desarrolladas

Durante el desarrollo del proyecto se fortalecieron diversas competencias profesionales, técnicas y metodológicas, las cuales contribuyeron de manera significativa a la formación integral y al desempeño en un contexto real de aplicación tecnológica. En primer lugar, se desarrollaron competencias en **análisis y solución de problemas**, al identificar necesidades específicas relacionadas con el aprendizaje de la lengua náhuatl y traducirlas en requerimientos funcionales para el diseño del sistema de tutoría inteligente. Esta capacidad permitió abordar retos técnicos y conceptuales de manera estructurada y fundamentada.

En el ámbito técnico, se consolidaron competencias en **desarrollo de software**, particularmente en el uso del lenguaje de programación Python y en la implementación de herramientas de Procesamiento del Lenguaje Natural. El manejo de librerías especializadas, la configuración de entornos virtuales y la resolución de conflictos de compatibilidad fortalecieron la habilidad para trabajar con tecnologías emergentes y adaptarse a distintos escenarios de desarrollo.

Asimismo, se desarrollaron competencias en **gestión de información y análisis de datos**, al realizar procesos de recolección, limpieza y estructuración de textos provenientes de fuentes digitales. Esta competencia resultó fundamental para garantizar la calidad de los datos utilizados en el análisis lingüístico, así como para interpretar los resultados obtenidos de manera crítica y objetiva.

En cuanto al diseño de soluciones innovadoras, se fortaleció la competencia de **integración tecnológica**, al vincular el análisis de lenguaje natural con el desarrollo de prototipos interactivos en Unity. Esta integración permitió comprender la importancia de la interoperabilidad entre diferentes plataformas y herramientas, así como el potencial de la tecnología para generar experiencias de aprendizaje significativas.

De igual manera, se desarrollaron competencias en **documentación técnica y comunicación escrita**, al elaborar reportes, diagramas, descripciones metodológicas y resultados de manera clara y estructurada. Esta habilidad facilitó la correcta transmisión de ideas, procesos y hallazgos, cumpliendo con los estándares académicos requeridos en un proyecto de residencias profesionales.

Finalmente, se fortalecieron competencias en **trabajo autónomo, responsabilidad profesional y ética**, al gestionar de manera independiente las actividades del proyecto, cumplir con los tiempos establecidos y considerar el impacto social y cultural del sistema desarrollado. Estas competencias reflejan una preparación adecuada para el ejercicio profesional y para la participación en proyectos tecnológicos con responsabilidad social.

Asimismo, la propuesta de integrar recursos de realidad virtual en fases posteriores del proyecto refuerza el enfoque cultural del sistema, ya que abre la posibilidad de crear entornos de aprendizaje inmersivos que favorezcan la interacción con elementos simbólicos, lingüísticos y contextuales propios de la lengua náhuatl. Este enfoque innovador permite vincular la tecnología con la identidad cultural, promoviendo experiencias educativas más significativas y respetuosas con el origen y uso de la lengua.

En conjunto, estas competencias reflejan una preparación adecuada para el ejercicio profesional, al combinar habilidades técnicas con una **conciencia social y cultural**, así como para la participación en proyectos tecnológicos que buscan generar un impacto positivo en la sociedad. La capacidad de desarrollar soluciones tecnológicas con responsabilidad social demuestra una formación integral, orientada no solo a la eficiencia técnica, sino también al compromiso con la preservación cultural y el desarrollo sostenible del conocimiento.

Referencias

- IBM. (s. f. c). Redes neuronales recurrentes. Recuperado de <https://www.ibm.com/mx-es/think/topics/recurrent-neural-networks>
- Python Software Foundation. (2025). Python 3.13.2 documentation. Recuperado de <https://docs.python.org/3/>
- BeautifulSoup. (s. f.). BeautifulSoup documentation. Recuperado de <https://www.crummy.com/software/BeautifulSoup/>
- Requests. (s. f.). Requests: HTTP for humans. Recuperado de <https://docs.python-requests.org/>
- spaCy. (s. f.). spaCy: Industrial-strength NLP. Recuperado de <https://spacy.io/>
- Explosion AI. (2024). spaCy usage examples. Recuperado de <https://spacy.io/usage>
- Unity Technologies. (2023). Unity User Manual (2022 LTS). Recuperado de <https://docs.unity3d.com/Manual/>
- Unity Technologies. (2023). Unity VR development overview. Recuperado de <https://unity.com/solutions/vr>
- Unreal Engine. (s. f.). Realidad virtual con Unreal Engine. Recuperado de <https://www.unrealengine.com/>
- Godot Engine. (s. f.). Godot Engine documentation. Recuperado de <https://docs.godotengine.org/>
- National Center for Biotechnology Information (NCBI). (2019). Natural language processing overview. Recuperado de <https://www.ncbi.nlm.nih.gov/>
- O'Reilly Media. (2017). Python for Data Analysis. Recuperado de <https://www.oreilly.com/>
- Manning Publications. (2018). Deep Learning with Python. Recuperado de <https://www.manning.com/>
- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3.^a ed.). Pearson.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.
- Crystal, D. (2000). *Language death*. Cambridge University Press.

UNESCO. (2019). *Atlas de las lenguas del mundo en peligro*. Recuperado de <https://unesdoc.unesco.org/>

Instituto Nacional de Lenguas Indígenas (INALI). (2022). Lengua náhuatl: Variantes lingüísticas de México. Recuperado de <https://www.inali.gob.mx/>

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*.

Freina, L., & Ott, M. (2015). A literature review on immersive virtual reality in education. *eLearning and Software for Education*.

Radianti, J., Fromm, J., & Wohlgenannt, I. (2020). Immersive VR applications in higher education. *Education and Information Technologies*.

Pressman, R. S., & Maxim, B. R. (2020). *Ingeniería de software: Un enfoque práctico* (9.^a ed.). McGraw-Hill.

Sommerville, I. (2016). *Software engineering* (10.^a ed.). Pearson.

Google Developers. (s. f.). Introduction to Natural Language API. Recuperado de <https://cloud.google.com/natural-language>

Hugging Face. (s. f.). transformer models documentation. Recuperado de <https://huggingface.co/docs>

Stanford University. (2018). CoreNLP: A suite of NLP tools. Recuperado de <https://stanfordnlp.github.io/CoreNLP/>

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. Recuperado de https://www.cs.toronto.edu/~kriz/imagenet_classification_with_deep_convolutional.pdf

Benetech. (s. f.). Technology for language preservation. Recuperado de <https://benetech.org/>