Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación
Maestría en Ciencias de la Computación

Determinación automática del grado de dominio de una lengua extranjera usando modelos de lenguaje.

Tesis presentada para obtener el grado de:
Maestro en Ciencias de la Computación

Presenta:
Carlos Andrés Martínez González

Director de Tesis:
Dr. David Eduardo Pinto Avendaño

Codirectora de Tesis:
Dra. Darnes Vilariño Ayala

Puebla, Puebla, Septiembre 8 de 2025

Resumen

El proyecto de maestría tiene como objetivo principal determinar el grado de dominio de una lengua extranjera utilizando modelos de lenguaje. Para ello, se construirán diversos modelos que permitan evaluar aspectos clave del dominio del idioma, como la gramática, el vocabulario, la fluidez y la coherencia, entre otros. Estos modelos serán entrenados con datos de hablantes no nativos para identificar patrones y características que reflejen distintos niveles de competencia en la lengua extranjera.

El enfoque del proyecto se centra en el uso de técnicas de procesamiento de lenguaje natural (PLN) para analizar producciones lingüísticas de hablantes no nativos. Se seleccionarán corpus representativos que incluyan diferentes niveles de competencia, desde principiantes hasta avanzados, lo que permitirá desarrollar un modelo robusto y adaptativo capaz de distinguir entre los distintos grados de dominio del idioma. Además, se considerará el uso de algoritmos de aprendizaje supervisado y no supervisado para mejorar la precisión de las evaluaciones.

Un aspecto clave del proyecto será la evaluación de los modelos construidos. Se implementarán métricas específicas que permitan medir la exactitud y consistencia de las predicciones sobre el nivel de competencia lingüística. Estas métricas se validarán mediante comparaciones con evaluaciones humanas realizadas por expertos en enseñanza de lenguas. El objetivo es lograr un sistema que no solo sea eficiente en la clasificación automática, sino también interpretable, ofreciendo retroalimentación útil para los aprendices y los docentes.

Finalmente, se espera que los resultados de este proyecto contribuyan al campo de la evaluación automatizada de lenguas, ofreciendo una herramienta innovadora para medir el progreso en el aprendizaje de un idioma extranjero. A largo plazo, estos modelos podrían integrarse en plataformas educativas, brindando a los estudiantes no nativos

una manera accesible y precisa de monitorear su avance y mejorar en áreas específicas de su aprendizaje lingüístico.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Con el avance diario de la tecnología aplicada al campo computacional, se ha revolucionado la calidad de vida del ser humano, especialmente en áreas como la educación y la comunicación. En este contexto, los modelos de lenguaje han demostrado ser herramientas fundamentales para evaluar y mejorar la competencia lingüística en diversos idiomas. Este trabajo de investigación se centra en generar un modelo de aprendizaje automático que permita determinar de manera precisa el grado de dominio lingüístico en el idioma inglés como lengua extranjera. Se propone utilizar modelos de lenguaje basados en Procesamiento de Lenguaje Natural (PLN), para evaluar aspectos clave del dominio del idioma inglés, tomando en cuenta que el hablante no es nativo. El sistema evalúa competencias en comprensión, expresión oral y escrita, gramática, vocabulario y otros factores relevantes que determinan el nivel de competencia en inglés. Para ello, se utiliza un modelo GPT-J encargado de aplicar las pruebas, en conjunto con un submodelo interno que determina el nivel de dominio del idioma conforme a los parámetros establecidos por el Marco Común Europeo de Referencia para las Lenguas (MCER) (de Europa, 2002).

El sistema se implementa en una plataforma web, permitiendo la interacción fluida entre el usuario y el sistema de evaluación del grado de dominio del idioma inglés. Esta interfaz utiliza tanto texto como audio, permitiendo medir las competencias lingüísticas de los usuarios, lo cual es clave para el aprendizaje efectivo del idioma, obteniendo datos reales para refinar el algoritmo basado en los resultados obtenidos. Estas pruebas permiten ajustar el modelo de acuerdo con las necesidades de los usuarios y asegurar una tasa de precisión adecuada para su implementación a mayor escala.

A fin de garantizar una evaluación precisa y confiable, se emplean métricas como F1-score (V7 Labs, 2022), Perplexity (Jurafsky y Martin, 2009), ROUGE (Lin, 2004), BLEU

(Papineni y cols., 2002), WER (Wikipedia, 2025) y PER (He y Radfar, 2021); además de un corpus representativo de inglés estadounidense proveniente del Open American National Corpus (Ide y cols., 2008), Wikipedia (Wikimedia Foundation, 2024), Europarl (Koehn, 2005) y LibriSpeech (Panayotov y cols., 2015), con el propósito de seleccionar los datos más adecuados para el entrenamiento y validación del modelo propuesto. Este estudio no solo busca desarrollar un sistema de evaluación automatizada, sino también contribuir al campo del Procesamiento del Lenguaje Natural (PLN) con una metodología que permita evaluar de manera efectiva la adquisición de una segunda lengua, mediante modelos avanzados de Inteligencia Artificial (IA).

1.1. Planteamiento del problema.

El dominio del idioma inglés es un requisito esencial en diversos ámbitos académicos y profesionales, y su evaluación se ha convertido en un factor determinante para el acceso a oportunidades educativas y laborales. Sin embargo, los métodos tradicionales de evaluación, como exámenes estandarizados (TOEFL, IELTS, Cambridge, entre otros), presentan limitaciones significativas. Estas pruebas, aunque ampliamente aceptadas, se basan en estructuras fijas que pueden no reflejar con precisión la competencia comunicativa real de los hablantes, ya que evalúan el conocimiento del idioma de forma segmentada y en condiciones controladas, sin considerar plenamente la espontaneidad y variabilidad del lenguaje en situaciones cotidianas.

Además, la preparación para estos exámenes implica un alto costo económico, tanto en términos de la inscripción como en materiales de estudio y cursos preparatorios, lo que limita el acceso equitativo a la certificación del dominio del inglés. Esto afecta particularmente a estudiantes autodidactas o personas sin acceso a instituciones especializadas que faciliten la evaluación de su progreso en el aprendizaje del idioma.

Con el avance de la inteligencia artificial y los modelos de lenguaje natural, han surgido herramientas automatizadas que prometen evaluar el dominio del inglés de manera más accesible y eficiente. No obstante, muchas de estas herramientas dependen de software propietario, lo que restringe su disponibilidad, flexibilidad y transparencia en los criterios de evaluación. Asimismo, los modelos de IA generalistas, como los basados en redes neuronales profundas, no han sido diseñados específicamente para evaluar el inglés bajo criterios estandarizados como los del Marco Común Europeo de Referencia para las Lenguas (MCER), lo que genera incertidumbre en la validez de sus resultados.

Otro desafío radica en la falta de métricas cuantificables que permitan evaluar de manera objetiva aspectos esenciales como pronunciación, gramática, vocabulario y comprensión

lectora; mientras que los exámenes tradicionales ofrecen puntuaciones basadas en rúbricas predefinidas, los sistemas de IA suelen carecer de mecanismos estructurados que permitan medir con precisión la calidad de las respuestas generadas por los usuarios. Sin métricas adecuadas, la evaluación automatizada podría ser inconsistente o sesgada, lo que comprometería su fiabilidad en entornos educativos y profesionales.

Ante esta problemática, se plantea la necesidad de desarrollar un sistema de evaluación del inglés basado en modelos de lenguaje de código abierto que integre métricas cuantificables como F1 Score, Perplexity, ROUGE, BLEU, WER y PER. Este enfoque permitiría una evaluación más objetiva y accesible, asegurando que las respuestas de los usuarios sean analizadas bajo criterios estandarizados y con un nivel de precisión comparable al de los exámenes tradicionales.

El aprendizaje y la enseñanza de una lengua extranjera, por ejemplo, el inglés presenta diversos retos, especialmente en la evaluación precisa del grado de dominio lingüístico que una persona ha alcanzado, problemática también discutida en estudios recientes sobre modelos automatizados de evaluación (Escobar Hernández, 2021). Para ello, se puede automatizar dicha evaluación a través de modelos de lenguaje, centrándose en aspectos clave como gramática, vocabulario, pronunciación, fluidez y coherencia. Para abordar esto, nos planteamos el siguiente objetivo general y específicos y posteriormente, las siguientes preguntas de investigación.

1.2. Objetivos.

Objetivo general.

Desarrollar un modelo de lenguaje que pueda analizar aspectos del idioma inglés, como la gramática, el vocabulario, la fluidez, la coherencia y la pronunciación, utilizando técnicas de PLN como la lematización, tokenización y métricas de evaluación lingüística.

Objetivos específicos.

- Implementar una plataforma web que permita la interacción fluida entre el usuario y el sistema de evaluación del grado de dominio del idioma inglés. Esta interfaz utiliza tanto texto como audio, permitiendo medir las competencias lingüísticas de los usuarios, lo cual es clave para el aprendizaje efectivo del idioma.
- Evaluar la precisión del modelo en el idioma inglés utilizando las métricas especializadas en el procesamiento de lenguaje natural. Los resultados obtenidos por el sistema se compararán con exámenes tradicionales de certificación de

inglés, como el TOEFL o el Cambridge English, para validar la precisión del modelo en diferentes niveles de dominio del inglés.

- Realizar pruebas piloto del modelo con hablantes no nativos de inglés, obteniendo datos reales para refinar el algoritmo basado en los resultados obtenidos. Estas pruebas permitirán ajustar el modelo de acuerdo con las necesidades de los usuarios y asegurar una tasa de precisión adecuada para su implementación a mayor escala.

1.3. Preguntas de investigación.

a) ¿Cuáles son los parámetros más efectivos para evaluar el dominio del inglés en hablantes no nativos a través de modelos de lenguaje?.

b) ¿Cómo puede un modelo de lenguaje determinar con precisión el grado de dominio del inglés en términos de gramática, vocabulario, fluidez y pronunciación?.

c) ¿Qué técnicas de reconocimiento de voz son más adecuadas para la conversión de voz a texto en la evaluación de la pronunciación y fluidez en inglés?.

d) ¿De qué manera puede una plataforma de interfaz humano-computadora mejorar la evaluación del inglés, mediante la integración de modelos de lenguaje?.

Las respuestas a estas incógnitas guiarán el desarrollo de los objetivos de investigación.

1.4. Hipótesis de la investigación.

Existe la posibilidad de desarrollar un sistema automatizado basado en modelos de lenguaje que permita evaluar el grado de dominio de nivel de idioma.

1.5. Estructura de la tesis.

La siguiente estructura describe la composición de la tesis, en cada uno de los capítulos se desarrolla un aspecto clave para el estudio y su implementación. A continuación, se

describen los puntos relevantes de cada capítulo, desde los objetivos hasta el resultado esperado por cada apartado.

1.5.1 Capítulo 1: Introducción.

En este capítulo se aborda la información preliminar con relación al tema de investigación, desde el contexto en que se plantea el problema, hasta el análisis de los objetivos, preguntas de investigación e hipótesis correspondientes.

1.5.2 Capítulo 2: Marco Teórico.

En este capítulo de hace un análisis de los conceptos teóricos necesarios para el desarrollo del proyecto, abarcando medidas necesarias, así como las métricas empleadas desde el ámbito computacional.

1.5.3 Capítulo 3: Estado del Arte.

En este capítulo de hace un análisis profundo de los trabajos relacionados con el tema de investigación, evaluando y seleccionando los que sean cercanos a la idea de propuesta de solución, así como las tecnologías conocidas hasta el momento y que se ha hecho al respecto para afrontar los retos del PLN en la actualidad.

1.5.4 Capítulo 4: Propuesta Metodológica.

En este capítulo se hace una propuesta de solución que enlista las alternativas para crear un sistema automatizado de evaluación de nivel de idioma que emplee modelos de lenguaje para determinar el grado de dominio lingüístico sin intervención humana.

1.5.5 Capítulo 5: Resultados.

En este capítulo se visualizan los diversos resultados obtenidos de la experimentación utilizando la metodología propuesta en el capítulo anterior, así como las herramientas utilizadas, las técnicas, medidas empleadas de PLN y los volúmenes de datos adecuados para lograr los resultados esperados.

1.5.6 Capítulo 6: Conclusiones.

En este capítulo se describen los hallazgos arrojados por la investigación, así como plantear las diversas posibilidades, puntos a favor y oportunidades de mejora del modelo resultante y su aplicación en el sistema de evaluación automatizada.

1.5.7 Capítulo 7: Trabajo futuro.

En este capítulo se hace una retroalimentación de lo obtenido hasta el momento con el desarrollo del tema de investigación y se plantean escenarios posibles de desarrollo continuo posterior, asimismo, se describe las tecnologías necesarias que se podrían añadir a largo plazo para su perfeccionamiento.

Capítulo 2

Marco Teórico

El dominio de un idioma se refiere a la capacidad de utilizar una lengua de manera correcta y adecuada en diversas situaciones comunicativas. Este dominio implica no solo el conocimiento profundo de las reglas gramaticales y sintácticas, sino también el manejo apropiado del vocabulario, la pronunciación, la ortografía y la puntuación. El verdadero dominio de una lengua requiere que el hablante sea capaz de adaptar su uso a diferentes contextos, comprendiendo las sutilezas y normas que rigen el lenguaje en cada situación.

2.1. Medidas de evaluación: ámbito lingüístico.

Existen diversos aspectos del lenguaje que determinan el grado de dominio de un idioma, los cuales abarcan desde la compresión lectora, el uso adecuado de la gramática, el vocabulario e incluso habilidades de comunicación como la pronunciación y la capacidad de escuchar son fundamentales a la hora de interactuar con otras personas en el idioma a comprender. Para evaluar el nivel de dominio de un idioma, el Marco Común Europeo de Referencia de las Lenguas (MCER por sus siglas en inglés)(de Europa, 2002) es un estándar ampliamente utilizado, el cual define las competencias lingüísticas en tres ámbitos: recepción, producción e interacción. Este marco establece seis niveles de referencia (A1 a C2), que agrupan a los hablantes en categorías básicas, intermedias y avanzadas.

Para medir el dominio de una segunda lengua, existen diferentes métodos de evaluación que, generalmente, concluyen con la obtención de una certificación. Entre los más conocidos se encuentra el Test of English as a Foreign Language (TOEFL), una prueba que evalúa el nivel de inglés de hablantes no nativos, enfocándose principalmente en su

capacidad académica. Asimismo, el examen de Cambridge, emitido por la Universidad de Cambridge, otorga un diploma que acredita el nivel de inglés en diferentes grados de competencia. Ambos métodos son ampliamente reconocidos a nivel internacional.

El PLN ha permitido avances significativos en la comprensión, interpretación y manipulación del lenguaje humano por parte de las máquinas. Dentro del PLN, los modelos de lenguaje representan el dominio del idioma mediante el uso de aprendizaje automático, cuyo objetivo es estimar la probabilidad de que una palabra, frase o símbolo ocurra en un texto. Uno de los métodos relevantes en el ámbito del PLN es la métrica BLEU, utilizada principalmente en traducción automática para evaluar la calidad de los textos generados por máquinas en comparación con traducciones humanas de alta calidad. Esta métrica mide la similitud entre ambos textos, teniendo en cuenta la precisión y penalizando las brevedades en el texto generado. Por otro lado, en el desarrollo de software para proyectos lingüísticos, la metodología de prototipo resulta esencial. Este enfoque permite la creación de prototipos iniciales que son evaluados y ajustados según la retroalimentación, favoreciendo la mejora continua y ofreciendo resultados tangibles en plazos cortos.

Además de las metodologías de evaluación mencionadas, el Marco Común Europeo de Referencia de las Lenguas (MCER) ha sido adoptado como una herramienta universal para estandarizar la medición de competencias lingüísticas en diversos idiomas. Este sistema no solo define los niveles de competencia, sino que también establece una guía detallada de las destrezas que los hablantes deben dominar en cada nivel, ya sea en comprensión auditiva, expresión oral, lectura o escritura. Esto ha permitido que instituciones educativas y empresas de todo el mundo utilicen el MCER como referencia para determinar el nivel de competencia lingüística de individuos, facilitando la movilidad internacional y el acceso a oportunidades laborales y académicas.

El uso de modelos de lenguaje en el PLN ha tenido un impacto considerable en la evaluación automática de idiomas. Estos modelos, entrenados con grandes cantidades de datos lingüísticos, son capaces de generar, predecir y corregir el lenguaje humano, lo que los convierte en una herramienta clave no solo para la traducción automática, sino también para la enseñanza de lenguas y la evaluación del dominio del idioma. Al integrar métodos de PLN con técnicas avanzadas de aprendizaje automático, como las redes neuronales y los transformers, los sistemas pueden ofrecer una retroalimentación precisa y en tiempo real sobre el nivel de competencia de un hablante, proporcionando una evaluación detallada y escalable en diferentes contextos educativos.

2.2. Métricas de evaluación: medidas con relación al modelo.

Para evaluar el rendimiento de los modelos de lenguaje, es fundamental utilizar medidas de evaluación especializadas. Estas permiten cuantificar la efectividad de un modelo así como guiar su optimización. A continuación, se describen las métricas disponibles que se utilizarán en el desarrollo de este proyecto.

Métrica F1.

Esta métrica (V7 Labs, 2022) se puede considerar como la más adecuada para el proyecto, debido a su capacidad de medir precisión del modelo de lenguaje, dado esto por la fórmula:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.1}$$

La fórmula describe el promedio armónico entre ***Precision*** (proporción de predicciones correctas respecto a todas las predicciones positivas) y ***Recall*** (proporción de predicciones correctas respecto a los datos positivos reales). Es útil para evaluar modelos cuando existe un desbalance entre las clases.

Métrica Perplexity.

De la misma forma, la métrica Perplexity (Jurafsky y Martin, 2009) también puede ser una opción viable a considerar tal como se ve en la fórmula:

$$\text{Perplexity} = 2^{H} \tag{2.2}$$

Donde se demuestra que, se mide qué tan bien un modelo de lenguaje predice una secuencia. La entropía cruzada ***(H)*** calcula la incertidumbre promedio en las predicciones del modelo. Valores bajos de perplexity indican un mejor ajuste del modelo al texto. Lo que permite medir la calidad de generación del modelo y la habilidad de construcción de frases en inglés.

Métrica ROUGE.

Esta métrica permite evaluar textos no necesariamente grandes (Lin, 2004), comparado con la métrica **BLEU**, lo que permite una mayor flexibilidad en la calidad de evaluación si se visualiza desde el ámbito de la conversación y el análisis de frases cortas, como se ve en la fórmula:

$$\text{ROUGE-N} = \frac{\text{n-grams in reference and generated text that match}}{\text{total n-grams in reference text (for recall) or generated text (for precision)}} \tag{2.3}$$

Donde se compara n-gramas generados por el modelo con los del texto de referencia, evaluando la similitud. Es ampliamente usado en tareas de resumen y generación de texto, midiendo precisión, recall y F1 a través de diferentes niveles de análisis. En otros aspectos, como las métricas necesarias para la evaluación de la pronunciación, existen diversos métodos, de los cuales los siguientes pueden ser de mayor utilidad para los objetivos de desarrollo y evaluación de nivel de idioma.

Métrica BLEU.

La métrica BLEU (Bilingual Evaluation Understudy por sus siglas en inglés) (Papineni y cols., 2002) es un método cuantitativo ampliamente utilizado para evaluar la calidad de traducciones automáticas y sistemas de generación de lenguaje natural. Esta métrica combina dos componentes principales: la precisión modificada de n-gramas (que evita premiar repeticiones excesivas) y un factor de penalización por brevedad (para castigar traducciones demasiado cortas). El resultado, es un puntaje normalizado entre 0 y 1, donde valores cercanos a 1 indican mayor concordancia con las referencias humanas, la cual se define en la siguiente fórmula:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.4}$$

Donde:

- *BP* es la penalización por brevedad para tener en cuenta la longitud del texto generado en comparación con los textos de referencia.
- *exp* es la función exponencial y se usa para convertir el promedio logarítmico (que puede ser negativo) en un valor positivo entre 0 y 1.
- w_n son los pesos asignados a cada n-grama.
- *n* es el índice que representa el orden del n-grama actual que se está evaluando.
- *N* es el orden máximo de n-gramas considerado (normalmente 4).
- P_n es la precisión del n-grama entre el texto generado y el texto o textos de referencia.

Métrica de tasa de error por palabra (WER).
El Word Error Rate (WER), o Tasa de Error de Palabras (Wikipedia, 2025), es una métrica ampliamente utilizada en la evaluación de sistemas de reconocimiento de voz. Esta medida cuantifica la precisión de un sistema al comparar la transcripción generada automáticamente con una transcripción de referencia considerada como correcta. El WER se calcula mediante la siguiente fórmula:

$$\text{WER} = \frac{S + D + I}{N} \tag{2.5}$$

Donde:

- *S* representa el número de sustituciones (palabras incorrectas).
- *D* denota el número de eliminaciones (palabras omitidas).
- *I* indica el número de inserciones (palabras añadidas incorrectamente).
- *N* es el número total de palabras en la transcripción de referencia.

Un WER más bajo indica un mayor nivel de precisión en el reconocimiento de palabras, siendo esta métrica fundamental para evaluar y comparar el rendimiento de sistemas de reconocimiento de voz.

Métrica de tasa de error de fonemas (PER).
El Phoneme Error Rate (PER), o Tasa de Error de Fonemas (He y Radfar, 2021), es una métrica que evalúa la precisión en el reconocimiento de unidades fonéticas. A diferencia del WER, el PER se centra en la comparación entre la transcripción fonética generada por un sistema y una transcripción fonética de referencia. Esta métrica es particularmente útil en tareas relacionadas con el análisis de la pronunciación y la evaluación de sistemas de reconocimiento de habla a nivel fonético. El PER se calcula de manera similar al WER como se ve en la siguiente expresión:

$$\text{PER} = \frac{S + D + I}{N} \tag{2.6}$$

Donde:

- *S* representa el número de sustituciones de fonemas.
- *D* denota el número de eliminaciones de fonemas.

- *I* indica el número de inserciones de fonemas.
- *N* es el número total de fonemas en la transcripción de referencia.

El PER es una herramienta valiosa para analizar la capacidad de un sistema de reconocer y transcribir correctamente los sonidos del habla, lo que resulta esencial en aplicaciones como la corrección de pronunciación o la síntesis de voz.

Divergencia de Kullback-Leibler.
Para comprender la naturalidad de una oración escrita en comparación con el vocabulario de un lenguaje de referencia, una de las técnicas empleadas es la divergencia de Kullback – Leibler (Kullback y Leibler, 1951). Esta expresión, tal como se ve en la fórmula:

$$D_{KL}(P \parallel Q) = \sum_{i} P(i) \log \left(\frac{P(i)}{Q(i)} \right) \tag{2.7}$$

Mide la diferencia entre dos distribuciones de probabilidad ***P*** (distribución real o esperada) y ***Q*** (distribución aproximada o predicción). Indica cuánto difiere un modelo de la distribución verdadera. Es fundamental en tareas que comparan distribuciones, como el análisis de texto. Dentro del flujo del programa, esta divergencia evalúa la similitud del vocabulario del estudiante en relación con el de los hablantes nativos.

Similitud de coseno.
Esta medida se basa en la comparación de la dirección de los términos en un espacio matemático (Salton y McGill, 1983), sin importar su tamaño o frecuencia absoluta, la cual se define por la siguiente expresión:

$$\textit{Similitud}_{\cos}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|} \tag{2.8}$$

En la ecuación se calcula la similitud entre dos vectores basándose en el ángulo entre ello, lo cual es ampliamente usado para comparar documentos o frases en términos de similitud semántica. Aquí:

- $\vec{A}$ y $\vec{B}$ son vectores de características.
- $\cdot$ representa el producto escalar entre los vectores.
- $\|\vec{A}\| \cdot \|\vec{B}\|$ son las magnitudes (normas) de los vectores $\vec{A}$ y $\vec{B}$.

Distribución de Zipf.

Para conocer el flujo de palabras y su aparición en el texto comparado, la distribución de Zipf fue esencial (Zipf, 1949), debido a que es un principio estadístico que describe cómo ciertas palabras en un idioma aparecen con mayor frecuencia que otras, siguiendo un patrón predecible; lo que facilita comprender la distribución léxica en un corpus, esta distribución se denota por la siguiente ecuación:

$$P(r) = \frac{C}{r^s} \tag{2.9}$$

Donde la frecuencia de una palabra ***P*** en un texto es inversamente proporcional a su rango ***r*** en términos de frecuencia. El exponente ***s*** determina la pendiente de la distribución, y ***C*** es una constante de normalización, lo que ayuda en comprender la distribución de frecuencia de palabras en un corpus.

Capítulo 3

Estado del arte

Dentro del estudio de la IA, diversas investigaciones han dado hincapié a su aplicación no únicamente en el desarrollo del dominio de una lengua extranjera, sino también en entornos virtuales donde los modelos de lenguaje puede ser una herramienta adicional al progreso humano, creando así diversas experiencias inmersivas donde la realidad da un paso más allá de la convivencia humana - computacional.

Con relación a la influencia de los modelos de lenguaje para determinar el grado de dominio de un idioma, existe una investigación que destaca cómo los modelos de lenguaje están revolucionando el aprendizaje del inglés al facilitar la creación de chatbots conversacionales que interactúan con los estudiantes mediante respuestas coherentes, mejorando así la práctica del idioma. Además, estos modelos permiten adaptar el contenido educativo a las necesidades individuales de cada estudiante, ofreciendo una experiencia de aprendizaje personalizada (Chicaíza Chicaíza y cols., 2023).

En un artículo de la Universidad de Florida (Dai y cols., 2021) sobre aplicaciones de la IA, se exploran diversas formas en que la IA se integra en modelos virtuales para evaluar y mejorar la enseñanza hacia los estudiantes. Se concluye que estos modelos son herramientas de apoyo docente efectivas, basadas en más de 20 horas de pruebas de rendimiento y funcionalidad.

Este mismo tema fue abordado por Villarroel (García Villarroel, 2021), quien señala cómo el aula tradicional ha evolucionado con la inclusión de tecnologías como la IA y técnicas de análisis de información como Big Data. Estas herramientas permiten personalizar las rutas de aprendizaje de los estudiantes y optimizar la gestión educativa. Además, la importancia de plataformas colaborativas como GitHub agilizan tareas rutinarias y ofrecen propuestas predictivas, así como la integración de la gamificación con IA, la cual mejora las habilidades cognitivas y fomenta un aprendizaje significativo y actualizado

con relación a las necesidades de la actualidad.

En el proyecto VALL-E X (Z. Zhang y cols., 2023),que propone un modelo de síntesis y traducción en varios lenguajes. Este modelo permite que una persona hable en un idioma extranjero manteniendo las características únicas de su voz. Utilizando métricas de evaluación como zero-shot learning, el sistema logra una traducción y síntesis eficiente en diversos idiomas sin necesidad de entrenamiento adicional, lo que mejora la precisión en la comparación de textos en diferentes lenguas al tiempo que conserva la identidad vocal del hablante.

Por otra parte, en un artículo realizado en la universidad del sur de Florida (Topsakal y Topsakal, 2022), exploran la combinación de la Realidad Aumentada (AR) con Voicebots y modelos de lenguaje como ChatGPT para la enseñanza de lenguas extranjeras a niños pequeños a través de un entorno agradable e interactivo. Además, el uso de la AR facilita una experiencia visual interactiva que atrae la atención de los niños, mientras que los Voicebots y modelos de lenguaje como ChatGPT permiten conversaciones personalizadas que refuerzan la inmersión lingüística. Esta combinación ayuda a que el aprendizaje de una lengua extranjera sea más efectivo y entretenido para los niños, permitiendo una mayor retención de conocimiento a través de la interacción constante.

En una investigación realizada por Gupta y otros investigadores (Gupta y cols., 2022), se analiza la evolución del reconocimiento de la actividad humana (HAR por sus siglas en inglés) en diversos aspectos tecnológicos, desde el avance de los dispositivos hasta el campo de la IA. El estudio enfatiza que el crecimiento de los dispositivos HAR está sincronizado con los avances en los marcos de trabajo de IA, lo que sugiere que la mejora de los modelos de lenguaje podría potenciar aún más las capacidades de HAR en diversas aplicaciones. Además, las recomendaciones del artículo proponen un marco de referencia para expandir el uso de modelos de lenguaje en HAR, explorando nuevas formas de integrar estas tecnologías para mejorar la calidad de vida humana.

Otra forma de evaluar el dominio de un idioma se aborda en la investigación de Muñoz y Fuertes (Muñoz-Basols y Fuertes Gutiérrez, 2024), que explora cómo la IA puede ser aplicada en el ámbito educativo; en particular, destacan que la interacción entre humanos y máquinas puede ser una herramienta valiosa en el aprendizaje de una segunda lengua. Esto incluye aspectos como el aprendizaje informal, la autonomía del estudiante y la (auto)evaluación.

De acuerdo con el artículo realizado en conjunto por Guano y otros investigadores (Guano Merino y cols., 2023) sobre el modelo de aprendizaje del idioma inglés utilizando algoritmos de machine learning, se analiza cómo estos algoritmos abarcan diversas áreas, como las ciencias psicológicas, sociales, lingüísticas y pedagógicas.

Aunque estos algoritmos pueden ser herramientas valiosas para personalizar el aprendizaje, evaluar habilidades y optimizar recursos educativos, también presentan riesgos, especialmente en la evaluación del dominio de una lengua extranjera. Entre estos riesgos se incluyen problemas de precisión en la evaluación, sesgo en los modelos y dependencia excesiva de la tecnología, lo que puede afectar tanto la calidad del aprendizaje como la equidad en la evaluación.

Existe, asimismo, un proyecto de aprendizaje continuo a lo largo de la vida (LLL por sus siglas en inglés) llamado LAMOL (Language Modeling for Lifelong Language Learning) (Sun y cols., 2019), que se presenta como una metodología innovadora en el ámbito del procesamiento del lenguaje natural (PLN). El proyecto permite que un modelo de lenguaje aprenda nuevas tareas sin olvidar las anteriores, logrando esto mediante la generación de pseudo-muestras de tareas previas mientras entrena para nuevas tareas, sin necesidad de memoria adicional o aumento en la capacidad del modelo. Este enfoque evita el problema del catastrophic forgetting (pérdida de lo aprendido al incorporar nueva información) y no muestra signos de intransigencia, es decir, dificultades para aprender nuevas tareas. Además, LAMOL ha demostrado ser capaz de realizar secuencialmente tareas de lenguaje muy diferentes utilizando un solo modelo, obteniendo resultados altamente eficientes dentro del enfoque de multitarea.

Dentro de los parámetros de evaluación sobre modelos de lenguaje, se encuentra un artículo que examina en profundidad los aspectos clave de la evaluación, centrándose en las preguntas: ¿qué evaluar?, ¿cómo evaluar? y ¿dónde evaluar? (Chang y cols., 2024) El artículo utiliza diversos estudios de rendimiento, protocolos y tareas para analizar no solo los avances realizados en los modelos de lenguaje, sino también para identificar áreas clave de mejora, estos se centran en métricas clave como perplexity y BLEU, señalando áreas específicas para mejorar la precisión y efectividad de estos modelos. Además, destaca la importancia de la evaluación en términos de aplicabilidad, ética y equidad, subrayando los desafíos que enfrentan estos modelos en cuanto a sesgo, generalización y manejo de tareas complejas.

En la investigación de Pérez (Perez y cols., 2022), se aborda el uso de modelos de lenguaje para evaluar y mejorar la robustez y seguridad de otros modelos, un proceso conocido como red teaming". Utilizan métricas como BLEU, zero-shot learning, y técnicas de "stochastic few-shot.º çonditional zero-shot"para evaluar la eficiencia y robustez de los modelos de lenguaje. A partir de estas evaluaciones, se clasifica el desempeño y la confiabilidad de los modelos. El objetivo es identificar vulnerabilidades y debilidades mediante la generación de entradas adversariales y el análisis de sus respuestas, para asegurar que los modelos sean más confiables y seguros.

En el artículo de Kasneci (Kasneci y cols., 2023), los autores exploran cómo los modelos de lenguaje grandes pueden ser utilizados de manera positiva en el ámbito educativo. Destacan que, además de determinar el grado de dominio de una lengua extranjera, estos modelos pueden apoyar diversos aspectos de la educación en distintas materias. Sin embargo, también abordan desafíos asociados con su inclusión, como la necesidad de proporcionar retroalimentación instantánea y facilitar el acceso a recursos educativos, así como problemas relacionados con la precisión de las respuestas. La ética del uso de IA en educación plantea preocupaciones sobre la dependencia excesiva y la privacidad.
En un artículo sobre comparación de rendimiento de modelos de lenguaje, los autores (Xu y cols., 2022), analizaron y compararon modelos de lenguaje de gran tamaño diseñados para trabajar con código, evaluando su capacidad de completar y sintetizar código desde descripciones en lenguaje natural. Destaca las limitaciones de modelos como Codex, que no son de código abierto, dejando incógnitas sobre sus decisiones de diseño. En este contexto, se evalúan modelos abiertos como GPT-J, GPT-Neo, GPT-NeoX-20B, y CodeParrot, mostrando que estos pueden acercarse al rendimiento de Codex en algunos lenguajes de programación.
En la investigación de (Boitel y cols., 2024) se compara el modelo GPT y el modelo BERT para el reconocimiento de emociones, analizando si un modelo general como GPT puede competir con modelos específicos como DeBERTa v3, se presentan las versiones de GPT (Davinci, Curie, Babbage y Ada) y su metodología de ajuste fino, mientras que para BERT se detalla un enfoque similar. La comparación revela que, aunque GPT tiene un rendimiento notable en diversas tareas, BERT muestra una mayor precisión en este caso específico. Sin embargo, se destaca el potencial de futuras mejoras en modelos como GPT-4, que podrían superar a los modelos actuales y ofrecer una mayor versatilidad. Esta información resulta especialmente relevante para proyectos que buscan implementar interacciones humano–modelo de lenguaje. Por ejemplo, en aplicaciones de análisis de sentimientos o evaluaciones idiomáticas, el uso de modelos como GPT para generar respuestas naturales podría complementarse con la precisión de BERT en tareas específicas, logrando una evaluación más completa y precisa del dominio del idioma.
En lo investigado con los autores (Ma y cols., 2023) sobre la corrección de errores en sistemas de reconocimiento del habla (ASR por sus siglas en ingés) se puede mejorar significativamente los resultados del reconocimiento de voz, especialmente cuando se utilizan modelos generativos como ChatGPT en configuraciones zero-shot o few-shot. Este enfoque, basado en las salidas N-best, ha mostrado ser eficaz incluso con arquitecturas avanzadas como transducers y modelos con atención. Modelos como T5

(Text-to-Text Transfer Transformer), que convierten cualquier tarea de procesamiento de lenguaje en un problema de generación de texto, se destacan por su versatilidad, haciendo que sean efectivos no solo para la corrección de errores ASR, sino también en diversas aplicaciones de PLN. Este enfoque podría integrarse al proyecto, mejorando la salida textual tras el reconocimiento de voz. Al aplicar técnicas similares, se lograría una evaluación más robusta y flexible de las habilidades lingüísticas en inglés, abarcando no solo la pronunciación, sino también la comprensión auditiva como parte integral del sistema.

En el artículo de Millière (Millière, 2024) se exploran las implicaciones de los modelos de lenguaje modernos, como los de la familia GPT, para la lingüística teórica. A pesar de que estos modelos están diseñados con objetivos de ingeniería, su capacidad para adquirir conocimiento lingüístico complejo a partir de grandes cantidades de datos plantea la necesidad de reconsiderar su relevancia en el ámbito de la teoría lingüística. Ésta información presenta evidencia empírica que sugiere que los modelos de lenguaje pueden aprender estructuras sintácticas jerárquicas y responder a fenómenos lingüísticos, aportando una colaboración más cercana entre lingüistas y científicos computacionales podría ofrecer valiosas perspectivas, especialmente en debates sobre el nativismo lingüístico.

En la investigación de (Chiang y Lee, 2023) se propone determinar si los modelos de lenguaje extensos (LLMs) pueden reemplazar evaluaciones humanas, realizaron experimentos en tareas de clasificación de texto, calificación automática de ensayos y análisis de contenido. Como resultado, Los LLMs demostraron ser consistentes y eficientes, sin embargo, aún necesitan supervisión humana para tareas complejas; lo que puede ser empleado en evaluaciones iniciales o como complemento en entornos educativos.

En lo propuesto por (Wang y cols., 2024) , se desarrolló un marco de evaluación para modelos grandes de lenguaje aplicados al audio (LLM-A). a través de proponer un conjunto de tareas de evaluación en diferentes modalidades: clasificación, generación, transcripción y síntesis. Incluye datasets preexistentes y adaptados para estas tareas. Este proyecto identificó las fortalezas y limitaciones de varios LLM-A existentes, destacando la necesidad de mejorar la coherencia en la generación y la calidad de la transcripción. Lo que puede favorecer el entender cómo integrar modelos de lenguaje para dominios específicos como audio, ofreciendo ideas reveladoras sobre posibles extensiones para el modelo de lenguaje aplicado al inglés.

Recientemente, se creó el proyecto AudioPaLM (Rubenstein y cols., 2023), en el cual se desarrolló un modelo de lenguaje multimodal que entiende y genera texto y audio con

alta calidad. Este modelo de lenguaje combina métodos de pre entrenamiento textual y de audio en un único modelo, utilizando datasets de alta calidad en ambas modalidades, también incluye aprendizaje de transferencia para tareas específicas como síntesis y traducción multimodal. Sus resultados mostraron rendimiento competitivo en tareas de síntesis y traducción, que igualan o superan a modelos especializados en audio. En pocas palabras, este modelo demuestra que puede haber un marco conceptual para expandir las capacidades de proyectos como éste hacia dominios multimodales en el futuro, permitiendo el manejo simultáneo de texto y audio en la evaluación de idiomas.

Otro proyecto relevante en el ámbito de la evaluación de audio por parte de modelos de lenguaje es SpeechVerse (Das y cols., 2024), el cual acorde a los autores su objetivo es diseñar un modelo escalable para tareas de lenguaje y audio con énfasis en generalización. Para ello, el modelo fue entrenado en un corpus masivo que incluye múltiples lenguajes y modalidades, también se usaron técnicas como Fine-Tuning progresivo y enmascaramiento predictivo. Este modelo de lenguaje especializado en audio destacó en tareas de generación y comprensión en diversos lenguajes, superando modelos previos en métricas como BLEU y F1, lo que indica que su enfoque puede ser relevante para evaluar cómo los modelos de lenguaje podrían manejar idiomas y sus variantes, especialmente entre dialectos o regiones.

En la investigación de (Yang y cols., 2024),se desarrolló AIR-Bench, una plataforma de evaluación diseñada específicamente para medir la comprensión generativa en modelos de audio, este sistema introduce tareas como la comprensión narrativa del audio y la generación de respuestas, utilizando datasets anotados manualmente. Los resultados evidenciaron que los modelos actuales son efectivos en tareas de comprensión básica, pero enfrentan desafíos en escenarios más complejos o especializados. Además, en futuras aplicaciones, AIR-Bench podría ofrecer métricas y tareas adaptables para evaluar componentes específicos de modelos de lenguaje aplicados al inglés, especialmente en contextos de evaluación avanzada.

Capítulo 4

Propuesta Metodológica

Para llevar a cabo la determinación del grado de dominio de una lengua extranjera usando modelos de lenguaje, se necesitará una metodología propuesta para llevar con éxito su desarrollo e implementación, tal metodología se concentra en los puntos siguientes:

- Realizar un estudio de las metodologías desarrolladas hasta el momento, en busca de un área de oportunidad.
- Comparar los modelos de lenguaje para determinar cuál se utilizará en este proyecto de investigación.
- Investigar la tecnología necesaria para llevar a cabo el desarrollo del proyecto.
- Obtener un corpus para el desarrollo del modelo de lenguaje a usar en el proceso de determinación del grado de dominio.
- Entrenar el modelo de lenguaje usando GPUs.
- Realizar las pruebas para determinar la eficiencia del modelo del lenguaje utilizando métricas adecuadas, por ejemplo, el puntaje F1 para la precisión del modelo o BLEU para la precisión de traducción de texto.
- Realizar los ajustes necesarios para mejora del modelo.
- Obtención del modelo ajustado.
- Evaluación del modelo generado.

4.1. Comparación de modelos

Como proceso inicial, se realizó un análisis de los diversos modelos de lenguaje que pueden ser aplicados para determinar el grado de dominio de una lengua extranjera, se encontraron diversas opciones tecnológicas que pueden ser aplicadas en el ámbito lingüístico, evaluando aspectos clave como la corrección ortográfica, semántica, gramática, vocabulario y naturalidad. Estos modelos, fueron seleccionados acorde a resultados de artículos previamente citados en el estado del arte, a continuación de describe brevemente sus cualidades tal como se ve en el Cuadro 4.1.

BERT	**GPT**	**RoBERTa**	**DeepSpeech**	**T5**
Tareas de comprensión del lenguaje y análisis de texto, evaluación de gramática y el vocabulario en oraciones.	Genera texto coherente para la interacción conversacional. Puede evaluar la fluidez y coherencia del habla.	Variante de BERT optimizada para un rendimiento superior en tareas de comprensión del lenguaje.	Modelo de reconocimiento de voz para evaluar la pronunciación y la fluidez del habla.	Convierte las tareas de procesamiento de lenguaje en problemas de texto a texto.

Cuadro 4.1: Modelos de lenguaje disponibles para el proyecto.

Acorde a lo mencionado, modelos avanzados como BERT y RoBERTa destacan en tareas relacionadas con textos amplios, como clasificación y comprensión, con un rendimiento robusto pero alto costo computacional. Por otro lado, DeepSpeech, especializado en audio, aborda aspectos específicos de reconocimiento de voz. Sin embargo, para el desarrollo del proyecto, sería necesario integrar funcionalidades prácticas de interacción texto-audio o texto-texto.
En este contexto, T5 ofrece capacidades avanzadas en tareas de generación de texto similar a BERT y RoBERTa, pero con un enfoque texto-texto más explícito. Sin embargo, tras analizar las capacidades de la familia GPT, se identifica un equilibrio clave entre capacidad conversacional (modelo humano-máquina) y un manejo razonable de recursos tecnológicos. Esto satisface la necesidad de evaluar el grado de dominio del idioma, haciendo que los modelos GPT sean los más adecuados para este proyecto.
Dentro de los modelos GPT disponibles, se realizó una comparación de características

entre las versiones GPT-Neo, GPT-J y GPT-3:

- GPT-Neo: Rendimiento competitivo en clasificación de texto, generación y traducción. Diseñado para optimizar costos computacionales.

- GPT-J: Análisis de las respuestas de los usuarios, clasificando la gramática y coherencia. Útil en aplicaciones como la enseñanza de idiomas y PLN.

- GPT-3: Sobresale en escritura creativa, soporte técnico automatizado, generación de código, y más.

Desde las métricas de evaluación en cuanto a precisión de modelos de lenguaje y servicios de traducción que serán utilizados para el desarrollo del proyecto, en el lado computacional las métricas posibles de uso se describen en el Cuadro 4.2.

Métrica	**Tipo de tarea**	**Cualidad**
F1 Score	Evaluación de modelo de lenguaje	Clasificación de errores y exactitud en componentes de evaluación.
BLEU	Tareas de textos grandes	Coincidencia de palabras y frases exactas, pero con rigidez en evaluaciones de competencia lingüística.
ROUGE	Tareas de textos diversos	Resumen y comprensión de lectura, útil cuando el texto generado tiene varias formas correctas de expresarse.
METEOR	Tareas de textos grandes (similar a BLEU)	Evaluación basada en sinónimos, variantes morfológicas y palabras semánticamente similares.
Perplexity	Evaluación de modelo y texto	Evalúa la calidad de generación del modelo y su habilidad de construcción de frases en inglés.
Exact Match	Clasificación booleana	Mide si el texto de salida coincide completamente con el de referencia (“correcto/incorrecto”).

Cuadro 4.2: Métricas de evaluación de modelos de lenguaje.

4.2. Análisis de datos.

Para realizar la comparación de las respuestas obtenidas por el usuario, se emplearon volúmenes de datos específicos que permitieran verificar aspectos como la ortografía, la pronunciación, entre otros factores, con el fin de catalogarlos correctamente. Como resultado, se determinaron dos tipos de corpus de información.
El primero se enfoca en los fonemas obtenidos al convertir audio a texto. Para ello, se utilizaron herramientas avanzadas de captura de audio, las cuales permitieron comparar y verificar si el usuario pronunció correctamente en inglés.
El segundo corpus, por su parte, se centró en la comprensión de textos, así como en la ortografía, la gramática y el vocabulario. Para este análisis, se tomaron ejemplos de artículos de periódicos y enlaces de Wikipedia, con el objetivo de evaluar las habilidades de escritura en inglés. Como resultado, se pudo proporcionar retroalimentación al usuario sobre su nivel de dominio del idioma.
A continuación, se describe a detalle la información obtenida a partir de los corpus disponibles y utilizados en este proceso, así como las tecnologías empleadas y las métricas aplicadas.

4.2.1. Corpus para la evaluación de audio.

El desarrollo de modelos de lenguaje aplicados a la evaluación del dominio del inglés requiere fuentes de datos adecuadas tanto en texto como en audio. Los corpus de texto permiten evaluar aspectos como gramática, vocabulario y comprensión lectora, mientras que los corpus de audio son esenciales para el análisis de la pronunciación. Además, el uso de bibliotecas especializadas en conversión de audio a texto facilita la integración de estos recursos en modelos de evaluación automatizados. Dentro de la investigación sobre la evaluación de la pronunciación del idioma inglés estadounidense, se destacan los siguientes corpus disponibles.

Common Voice (Mozilla).
Es un corpus de audio de código abierto desarrollado por Mozilla (Mozilla, 2015), diseñado para mejorar el reconocimiento de voz en distintos idiomas. Dentro de sus principales características incluyen:

- Variedad de acentos y edades: Permite evaluar pronunciación en distintos contextos.

- Archivos de audio emparejados con transcripciones: Facilita el entrenamiento y evaluación de modelos de reconocimiento de voz.
- Frecuencia de actualización: Asegura un flujo constante de datos nuevos para mejorar la robustez de los modelos.

Como tal, este corpus resulta útil para evaluar cómo distintos hablantes pronuncian palabras y frases en inglés estadounidense, lo que lo convierte en una herramienta esencial para evaluar variedad de repuestas.

LibriSpeech.
Éste corpus se encuentra basado en audiolibros del Proyecto Gutenberg (Panayotov y cols., 2015), el cual se caracteriza por los siguientes aspectos.

- Alta calidad de audio: Donde se encuentran grabaciones bien pronunciadas y sin ruido de fondo.
- Transcripciones alineadas: Lo que permite comparar la pronunciación real con la esperada.
- Segmentación en conjuntos çleanz .ºther": El cual permite realizar pruebas en distintos niveles de claridad y dificultad.

Este recurso es particularmente útil para evaluar pronunciación en escenarios de lectura estructurada, complementando otras bases de datos que incluyen habla más espontánea.

Wall Street Journal (WSJ Speech Corpus).
El WSJ Speech Corpus (Paul y Baker, 1992) es un conjunto de grabaciones de artículos periodísticos leídos en voz alta, proporcionando:

- Lenguaje formal y estructurado, ideal para evaluar pronunciación en contextos profesionales.
- Pronunciación estándar del inglés estadounidense, lo que permite comparar hablantes con un estándar de referencia.

Dado que este corpus no es de libre acceso, su uso está sujeto a licencias, pero su contenido es relevante para estudios de pronunciación en ámbitos formales.

Por otra parte, para integrar los corpus de audio en el modelo de evaluación, es necesario emplear bibliotecas que conviertan la entrada de voz en texto. Algunas de las herramientas más relevantes en este proceso podrían ser los siguientes:

Whisper (OpenAI).
En este modelo de reconocimiento de voz de código abierto (OpenAI, 2022) se ofrece una alta precisión en la transcripción de audio a texto, asimismo la capacidad de reconocer múltiples acentos y dialectos del inglés; y, por último, procesamiento de ruido ambiental y habla espontánea. La aplicación de esta tecnología podría ayudar en la evaluación del idioma al obtener transcripciones detalladas para comparar la pronunciación del hablante con la referencia estándar.

SpeechRecognition (Python Library).
Esta biblioteca es una interfaz para diversas APIs de reconocimiento de voz (A. Zhang, 2017) y se destaca por soportar múltiples motores como Google Speech-to-Text y Sphinx. Del mismo modo es muy accesible para integrar con Python al procesar archivos de audio en tiempo real. En este caso se puede emplear para realizar pruebas iniciales de reconocimiento de voz y comparación con modelos más avanzados como Whisper.

Praat.
Es una herramienta para el análisis fonético del habla que permite extraer características acústicas como formantes y entonación (Boersma y Weenink, 2023), comparar la pronunciación del usuario con valores estándar y evaluar la fluidez y claridad de la pronunciación. Su implementación en proyectos de reconocimiento de voz se enfoca en evaluar la calidad de la pronunciación en función de parámetros fonéticos objetivos.

4.2.2. Corpus para la evaluación de texto.

En el ámbito de la evaluación del dominio del inglés, los corpus textuales constituyen una herramienta fundamental para analizar aspectos clave como la precisión gramatical, la riqueza léxica y la capacidad de comprensión lectora. Estos recursos permiten examinar el uso del lenguaje en contextos reales, lo que resulta esencial para diseñar metodologías de evaluación efectivas. A continuación, se describen algunas de las principales fuentes utilizadas en este estudio.

Corpus of Contemporary American English (COCA).
El Corpus Contemporáneo de Inglés Americano (COCA por sus siglas en inglés) (Davies, 2008) es un corpus lingüístico de referencia para el estudio del inglés estadounidense contemporáneo. Creado por Mark Davies en 2008, es uno de los recursos más completos y ampliamente utilizados en investigaciones lingüísticas. Sus características principales incluyen estos tres aspectos:

- Un volumen superior a los 1.000 millones de palabras, provenientes de registros tanto escritos como orales.
- Una diversidad de dominios textuales, tales como noticias, ficción, textos académicos y conversaciones cotidianas.
- La posibilidad de analizar la frecuencia y el uso contextual del vocabulario, aspecto fundamental para evaluar la riqueza léxica de los usuarios.

Este corpus resulta particularmente útil para el entrenamiento de modelos de lenguaje, ya que permite trabajar con estructuras gramaticales y vocabulario representativo del inglés contemporáneo.

Open American National Corpus (OANC).
El Corpus Abierto Nacional Americano (OANC por sus siglas en inglés) (Ide y cols., 2008) es un corpus de acceso abierto diseñado para representar el inglés estadounidense en su diversidad de registros y contextos. Desarrollado como parte de un proyecto colaborativo, este recurso se destaca por su enfoque en la representatividad lingüística y su marcación detallada, del cual uno de sus principales aportes se encuentra en una amplia variedad de textos que abarcan desde documentos científicos hasta conversaciones informales, lo que permite evaluar el lenguaje en distintos niveles de formalidad. Del mismo modo también existe una marcación gramatical y semántica, facilitando el análisis estructural y funcional del lenguaje.
Este corpus es especialmente valioso para evaluar la competencia gramatical en contextos diversos, así como para analizar la adaptabilidad del usuario a diferentes registros lingüísticos.

Wikipedia.
La Wikipedia (Wikimedia Foundation, 2024) es un recurso de texto masivo y dinámico que, debido a sus características únicas, ha sido ampliamente adoptado como un corpus

para diversas tareas de investigación en procesamiento del lenguaje natural (PLN) y lingüística computacional. Se caracteriza por los siguientes aspectos:

- Gran Escala y Cobertura Amplia: Es uno de los corpus de texto más grandes disponibles públicamente. La Wikipedia en inglés, por ejemplo, cuenta con más de 7 millones de artículos y más de 4.8 mil millones de palabras (a mayo de 2025), cubriendo una vasta gama de temas y dominios. Su tamaño y diversidad de temas la hacen adecuada para entrenar modelos de lenguaje generales.

- Multilingüe: Existe en cientos de idiomas, lo que la convierte en una fuente excepcional para la investigación en procesamiento de lenguaje cruzado y multilingüe, incluyendo traducción automática, alineación de textos paralelos y estudios comparativos de idiomas.

- Estructura Semiestructurada: Aunque no es un corpus "limpio.[en] el sentido tradicional (contiene marcado Wiki, enlaces internos, categorías, infoboxes, etc.), esta estructura ofrece información valiosa. Los enlaces internos y las categorías pueden ser utilizados para crear redes semánticas, ontologías y para tareas de clasificación de texto y vinculación de entidades.

- Actualización Constante: Al ser una enciclopedia colaborativa, la Wikipedia se actualiza y expande continuamente, lo que permite a los investigadores trabajar con datos relativamente actuales y observar la evolución del lenguaje y los temas a lo largo del tiempo.

- Variedad de Estilos y Géneros: Si bien predomina el estilo enciclopédico, la gran cantidad de colaboradores y temas resulta en una diversidad de estilos de escritura, desde artículos muy técnicos hasta descripciones más generales, lo que añade a su riqueza lingüística.

- Disponibilidad Pública: Los dumps (volcados) de la base de datos de Wikipedia son freely available para su descarga, lo que facilita su uso por parte de la comunidad investigadora global.

Si bien es un recurso invaluable, su naturaleza colaborativa también presenta desafíos, como la variabilidad en la calidad del contenido, posibles sesgos sistémicos y la necesidad de preprocesamiento para eliminar el marcado Wiki y obtener texto plano.

Europarl.
El corpus Europarl (Koehn, 2005) es una colección de textos paralelos provenientes de las actas del Parlamento Europeo. Se caracteriza por los siguientes aspectos:

- Multilingüe: Contiene textos en 21 idiomas europeos diferentes, lo que lo convierte en un recurso invaluable para la investigación en traducción automática y lingüística comparada.
- Dominio específico: El contenido se centra en discursos y documentos parlamentarios, lo que lo hace ideal para el estudio del lenguaje político y legal.
- Gran tamaño: Con millones de palabras en cada idioma, Europarl ofrece un volumen significativo de datos para entrenar y evaluar modelos lingüísticos.
- Transducciones alineadas a nivel de frase: Permite la comparación directa de frases entre diferentes idiomas, facilitando la investigación sobre la traducción y la correspondencia entre lenguas.

Este recurso es particularmente útil para tareas de traducción automática estadística y neuronal, así como para estudios de lingüística comparada y terminología política. Su naturaleza multilingüe y su dominio específico lo hacen complementario a otros corpus que se centran en diferentes tipos de lenguaje o dominios.
Todas estas fuentes complementan los corpus estructurados al ofrecer material actualizado y representativo de la lengua en uso, lo que enriquece la evaluación de la competencia lectora y la capacidad de interactuar con textos auténticos.

LanguageTool (Python library).
LanguageTool es un corrector gramatical de código abierto que se utiliza para detectar errores de ortografía, gramática, estilo y puntuación en textos . El software, inicialmente concebido por Daniel Naber para su tesis (LanguageTool, 2025), ha evolucionado a través de la contribución de una extensa comunidad de desarrolladores y lingüistas. Al no depender de un único autor, se le considera un proyecto colaborativo y comunitario, lo que garantiza su continua actualización y mejora.
Dentro del contexto de la investigación, LanguageTool sirve como una herramienta de preprocesamiento invaluable. Se utiliza para estandarizar la calidad de los corpus de texto, asegurando que los datos de entrada estén libres de errores comunes antes de ser sometidos a análisis más profundos. Su naturaleza multilingüe, con soporte para más

de 30 idiomas, permite aplicarlo en diversos proyectos de procesamiento de lenguaje natural, mejorando la fiabilidad y consistencia de los resultados obtenidos

En resumen, la combinación de corpus de referencia como OANC y Europarl con herramientas de preprocesamiento como LanguageTool garantiza un corpus de texto de alta calidad y representatividad. Mientras los corpus preexistentes proporcionan un material lingüístico vasto y estructurado para tus análisis, la herramienta de LanguageTool te permite refinar y estandarizar la calidad del texto a nivel de gramática y estilo. Esta metodología integral asegura que los datos de tu corpus sean robustos y confiables, lo cual es fundamental para obtener resultados precisos y consistentes en cualquier proyecto de PLN.

Capítulo 5

Resultados

En esta sección se presentan y analizan los resultados obtenidos a partir de la implementación del diseño metodológico descrito en capítulos anteriores. El propósito fundamental fue evaluar el grado de dominio del idioma inglés de los usuarios mediante un modelo de lenguaje automatizado, empleando distintas métricas de PLN. Para ello, se seleccionaron y organizaron distintos conjuntos de datos que sirvieron tanto como corpus de referencia como entradas de validación durante la fase de pruebas. Estos datos incluyeron fragmentos representativos del corpus Europarl y de Wikipedia, junto con audios de evaluación generados por los propios usuarios en entornos controlados.

Los resultados aquí reportados fueron extraídos a partir del análisis comparativo entre las respuestas generadas por los usuarios y los textos de referencia previamente establecidos. Dichos análisis incluyeron el uso de métricas como BLEU, ROUGE, F1 Score, Perplexity, así como la tasa de error en reconocimiento de voz mediante WER (Word Error Rate) y PER (Phoneme Error Rate). A través de estas herramientas, se logró cuantificar de forma objetiva aspectos como la coherencia gramatical, la precisión léxica, la fluidez, y la calidad fonética en las producciones del usuario.

5.1. Selección de modelo auxiliar.

Para integrar estas técnicas y determinar cuál modelo puede ser funcional para el desarrollo del proyecto, se realizó el siguiente diagrama de flujo para proceder a evaluar textos de prueba y poder comparar a detalle el rendimiento de cada modelo, como se ve en la Figura 5.1.

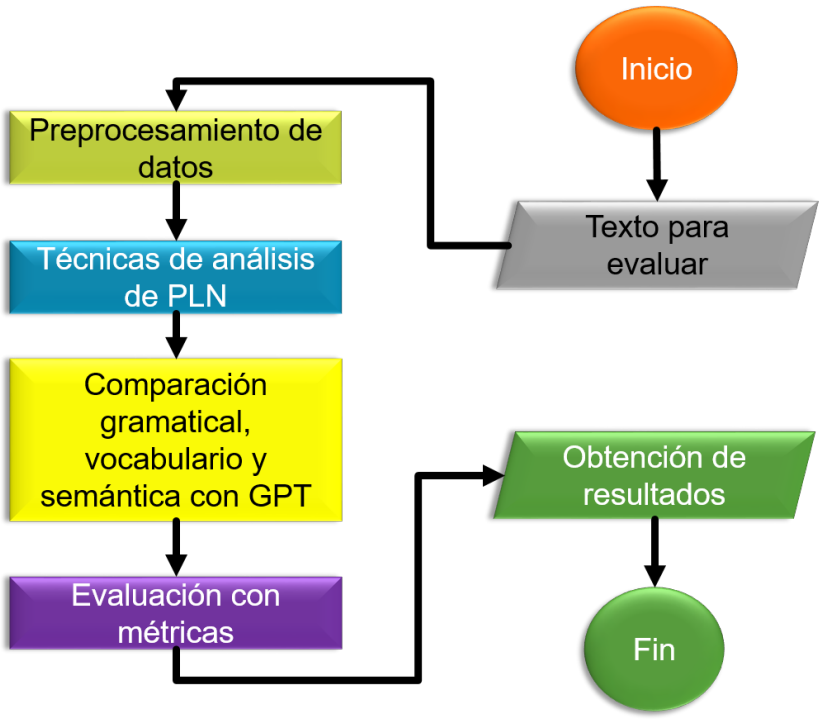

Figura 5.1: Funcionamiento algorítmico de prueba.

Por consiguiente, se realizó la evaluación de los modelos GPT-Neo, GPT-3 y GT-J como modelos seleccionados, utilizando las técnicas de análisis PLN antes mencionadas, por lo cual se demostró una eficiencia óptima en el modelo GPT-J y GPT-3 en comparación con el modelo GPT-Neo, en la Figura 5.2, se visualizan los resultados obtenidos con la distribución de la ley de Zipf.

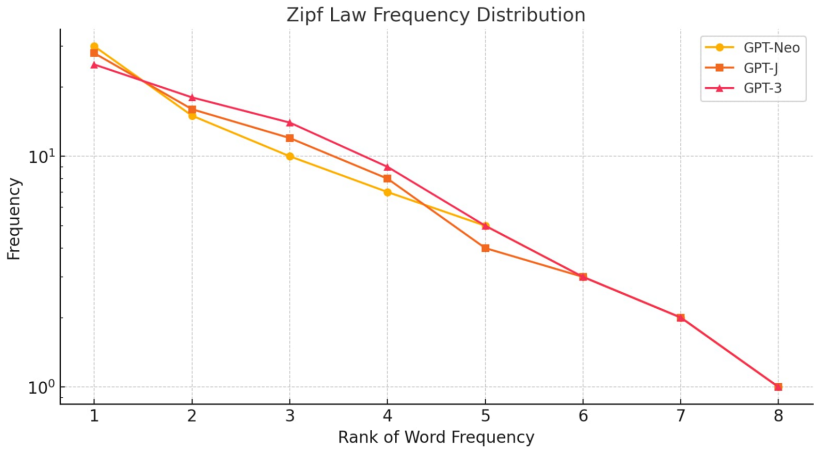

Figura 5.2: Gráfica de comparación de modelos GPT (Ley Zipf).

Del mismo modo, el la Figura 5.3, se puede comprender el rendimiento de los modelos utilizando la divergencia de Kullback-Leibler, donde acorde a las especificaciones de su uso, el rendimiento de GPT-J se encuentra en un rendimiento mayor frente a los resultados arrojados por el modelo GPT-Neo, y, por lo tanto, muy cercano al rendimiento óptimo del modelo GPT-3.

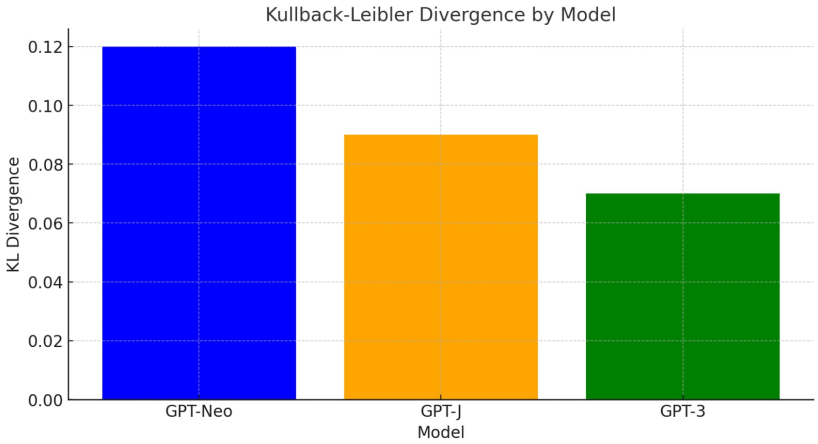

Figura 5.3: Gráfica de comparación de modelos GPT (Kullback-Leibler).

Con base en lo anterior mencionado, las características de los modelos seleccionados para pruebas y los resultados obtenidos en los diversos análisis de texto, se puede determinar que GPT-J tiene un rendimiento mayor en comparación a GPT-Neo ya que tiene mayor precisión en la comparación de resultados obtenidos por las técnicas de Zipf, Divergencia Kullback-Leibler, etc. Así mismo, es una herramienta de uso libre que permite un número indeterminado de pruebas sin restricción de licencia contrario al modelo GPT-3.

5.2. Diagrama de modelo final.

Comprendiendo la información obtenida sobre la comparación de modelos de lenguaje, GPT-J fue seleccionado para utilizarse en conjunto con el modelo de lenguaje que evaluará el grado de dominio del idioma haciendo énfasis en el idioma inglés, por otra parte se puede hacer uso de las técnicas del análisis PLN como la divergencia de Kullback-Leibler, la ley de coseno y Zipf para el análisis de respuestas y complementarlo con las herramientas de NLTK para la evaluación gramatical, vocabulario y comprensión lectora. Por otro lado, las herramientas como Whisper y SpeechRecognition pueden emplearse para la evaluación de la pronunciación.

Si a todo esto se le hace adición de métricas de evaluación de idioma como perplexity y rouge-n para el aspecto del texto y WER Y PER para la calidad de la pronunciación comprendida en apoyo de los corpus como LibriSpeech y el corpus realizado con la información de Wikipedia y textos periodísticos; se puede determinar un flujo de trabajo del funcionamiento de todo el sistema, el cual se puede ver representado en la Figura 5.4.

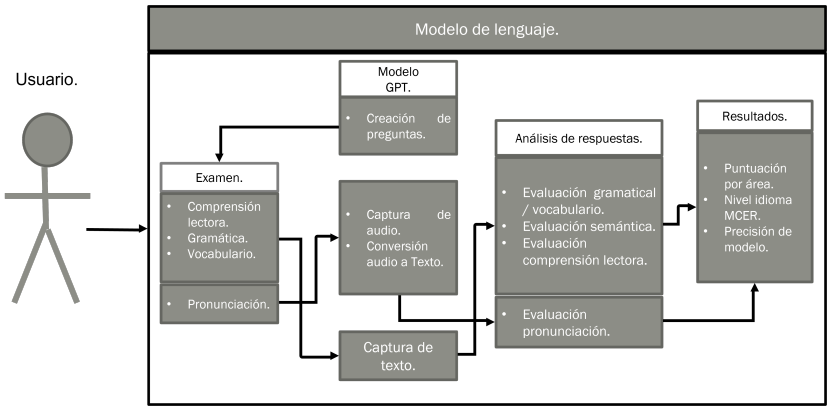

Figura 5.4: Diagrama de casos de uso modelo de lenguaje.

Aquí se puede apreciar que el modelo GPT genera las pruebas de evaluación de idioma, mientras que el usuario responde según el área de pregunta que presenta, el modelo de lenguaje creado se encargará de analizar y evaluar cada respuesta a través de diversos módulos especificados en la sección “Análisis de respuestas” utilizando todas las herramientas anteriormente mencionadas, como consecuencia los resultados se traducirán en términos de nivel de idioma acorde a los parámetros establecidos por el Marco Común Europeo de Referencia para las Lenguas, asimismo, la información arrojada por las métricas computacionales demostrarán que el modelo de lenguaje es capaz de evaluar el grado de dominio de idioma y al mismo tiempo se medirá la eficiencia del modelo de lenguaje construido.

5.3. Datos obtenidos del funcionamiento del modelo.

Al considerar que el modelo GPT-J genera las pruebas de evaluación del idioma, mientras que el usuario responde según el área de pregunta que presenta, el modelo de lenguaje creado se encarga de analizar y evaluar cada respuesta a través de diversos módulos especificados en la sección “Análisis de respuestas” utilizando todas las herramientas anteriormente mencionadas. Como consecuencia, los resultados se

pueden interpretar en términos del nivel de idioma acorde a los parámetros establecidos por el Marco Común Europeo de Referencia para las Lenguas, asimismo, la información arrojada por las métricas computacionales, demuestran que el modelo de lenguaje es capaz de evaluar el grado de dominio del idioma y al mismo tiempo se mide la eficiencia del modelo de lenguaje construido, en el Cuadro 5.1 se puede comprender a profundidad los primeros resultados de texto obtenidos de las pruebas de nivel del idioma utilizando el corpus de OANC.

#	Corrección Gramatical	ROUGE	Comprensión Lectora	BLEU	F1 Score	Perplexity
1	my mother was born on 1935	0.0042	0.0	0.0	0.0174	6.0
2	how are you ?	0.0001	0.0	0.0	0.0005	4.0
3	my favorite animas is cats , dogs and scorpions	0.0055	0.0	2.05e-282	0.0128	9.0
4	my name is andy and i am an english teacher and computer systems engineer	0.0084	0.0	9.79e-273	0.0079	14.0
5	oh , really ? think you do not	0.0068	0.0	3.19e-273	0.0190	7.0
6	the sun rises in the east	0.0032	0.0	1.45e-200	0.0152	5.0
7	what time is it now ?	0.0015	0.0	0.0	0.0038	3.5
8	i love programming with python	0.0091	0.0	4.75e-180	0.0215	11.0
9	artificial intelligence is the future	0.0078	0.0	5.89e-150	0.0107	8.0
10	today is a beautiful day	0.0027	0.0	7.12e-170	0.0143	6.5

Cuadro 5.1: Resultados obtenidos de las pruebas de texto (Corpus OANC).

La evaluación mediante ROUGE y otras métricas reveló que el modelo basado en GPT obtiene una puntuación comparable a modelos comerciales en términos de similitud textual, validando su uso en la evaluación del dominio del idioma inglés. En contraste, la divergencia de Kullback-Leibler indica que el modelo genera respuestas con una

distribución de probabilidad distinta a la de hablantes nativos, sugiriendo la necesidad de ajustes adicionales en el entrenamiento del corpus. Por otro lado, los análisis de similitud del coseno reflejaron una correlación moderada entre las respuestas generadas y los textos de referencia, lo que indica que el modelo capta estructuras lingüísticas clave, aunque con ciertas limitaciones en la comprensión semántica. Finalmente, la evaluación basada en la Ley de Zipf permitió identificar desviaciones en la frecuencia de uso de términos clave, lo que sugiere la necesidad de integrar corpus más representativos del inglés contemporáneo para mejorar la naturalidad del lenguaje generado.

En cuanto a las pruebas de pronunciación, los resultados describen diversos valores en las métricas en cuanto a fonemas detectados en la pronunciación (aplicación de la métrica PER), comparados con las transcripciones del corpus antes mencionado, mientras que los valores de la métrica WER reflejan la calidad de la transcripción obtenida de la respuesta del usuario. En el Cuadro 5.2 se presentan en detalle los resultados de pronunciación junto con observaciones cualitativas de cada experimento, utilizando el corpus Librispeech como referencia para la comparación de las respuestas de audio.

#	Texto Transcrito	WER	PER	Observaciones por experimento
1	The cat is on the table	0.12	0.08	Buena pronunciación
2	She want go to school	0.22	0.15	Errores en conjugación
3	I am happy today	0.10	0.05	Pronunciación clara
4	We going to the park	0.18	0.12	Problema con la gramática
5	This is a apple	0.25	0.18	Error en artículo
6	My brother play soccer	0.20	0.14	Error en tiempo verbal
7	They are studying English	0.08	0.04	Excelente pronunciación
8	The book on the table	0.30	0.22	Omisión de verbo
9	He is doctor	0.15	0.10	Falta de artículo
10	We have a meeting tomorrow	0.05	0.03	Pronunciación fluida

Cuadro 5.2: Resultados obtenidos de las pruebas de pronunciación.

En las pruebas de pronunciación, la evaluación basada en las métricas WER y PER mostró que la transcripción automática de los audios presenta variaciones en la precisión, dependiendo de la claridad del hablante y la complejidad de las frases evaluadas. Se observó que las frases con estructuras gramaticales más simples

obtuvieron menores tasas de error, mientras que aquellas con mayor longitud o con pronunciaciones poco convencionales generaron una mayor tasa de error. Además, la comparación con el corpus de referencia reveló que el modelo mantiene una correlación aceptable en la identificación de fonemas comunes en inglés, aunque con errores recurrentes en palabras de pronunciación ambigua. Esto sugiere que el desempeño del sistema puede optimizarse mediante mejoras en el preprocesamiento de audio y la incorporación de técnicas de alineación fonética más robustas. Por otra parte, en las pruebas realizadas con el corpus de Europarl se obtuvieron resultados significativos que ofrecen una visión detallada de la capacidad del modelo para discernir entre distintos niveles de dominio lingüístico. Este corpus, ampliamente utilizado en tareas de procesamiento de lenguaje natural por su riqueza léxica y estructura gramatical formal, permitió evaluar el desempeño del sistema bajo condiciones que simulan interacciones reales en contextos diplomáticos y legislativos. Los hallazgos obtenidos son esenciales no solo para validar la robustez del modelo, sino también para comprobar su aplicabilidad práctica en escenarios reales donde la evaluación del nivel de idioma es crítica, tales como exámenes de certificación, procesos de selección académica o implementación de sistemas inteligentes de tutoría. Además, los resultados permitieron identificar qué métricas mostraban mayor sensibilidad ante variaciones léxicas, sintácticas y semánticas, fortaleciendo así la utilidad diagnóstica del sistema propuesto. En el Cuadro 5.3 se pueden visualizar los resultados obtenidos en las pruebas de texto obtenidas con el corpus de Europarl.

#	Corrección Gramatical	ROUGE	Comprensión Lectora	BLEU	F1 Score	Perplexity
1	Hello, you AR you?	0.000	0.000	0.000	0.000	4.00
2	The person that I think... is Andy	0.111	0.000	3.39×10^{-234}	0.135	11.00
3	Woah!, it is amazing!	0.025	0.000	1.67×10^{-247}	0.043	4.00
4	I don't know what is forbidden	0.043	0.000	5.80×10^{-244}	0.032	6.00
5	Dear Ralph, I hope you are good...	0.019	0.000	2.08×10^{-235}	0.022	10.00
6	UATs your opinion abut mater...	0.000	0.000	0.000	0.000	7.00
7	I hate school, Iron now UAI...	0.028	0.000	2.00×10^{-238}	0.020	8.00
8	Oh!, sounds good, I'm surprised	0.000	0.000	0.000	0.000	5.00
9	Do you want to spin French, please?	0.019	0.000	2.19×10^{-237}	0.024	7.00
10	Okay but...eats is the question? don't understand yo	0.108	0.000	1.70×10^{-234}	0.073	8.00

Cuadro 5.3: Resultados obtenidos de las pruebas de texto (Corpus Europarl).

Se puede visualizar a detalle los resultados de los experimentos realizados utilizando el corpus Europarl como referencia para evaluar el desempeño del sistema de evaluación desarrollado utilizando las métricas antes descritas y empleadas en el otro corpus. En lo referente a la similitud textual (BLEU y F1), los valores obtenidos son bajos, lo

cual era previsible debido a que las respuestas del usuario son abiertas y no buscan reproducir literalmente las respuestas de referencia. La métrica ROUGE-1 también refleja esta divergencia, evidenciando coincidencias limitadas a nivel de unigramas. No obstante, este comportamiento no necesariamente indica una comprensión deficiente, sino más bien una variación esperada en el uso del lenguaje natural. Mientras que perplexity, empleada como indicador de fluidez gramatical, se sitúa entre valores de 4 y 11, lo cual sugiere que las respuestas generadas son sintácticamente válidas y comprensibles, aunque puedan diferir del estilo o contenido del corpus utilizado como base de comparación.

En cuanto a la Similitud de Coseno, se observa una puntuación constante de 1.000, resultado de haber comparado las respuestas corregidas consigo mismas o con textos muy similares. Este comportamiento será ajustado posteriormente al aplicar un modelo de generación distinto durante la fase de entrenamiento. Por último, la Divergencia de Kullback-Leibler reporta un valor de 0.000 en todos los casos, lo cual indica que las distribuciones de probabilidad comparadas son idénticas. Este resultado se debe a la configuración actual del sistema de comparación, la cual será refinada al integrar a GPT-J como generador de texto base para la evaluación automática.

La utilización de dos corpus distintos (Europarl y OANC) permitió realizar una comparación representativa que evidencia la capacidad del modelo propuesto para evaluar el grado de dominio del idioma inglés bajo distintos contextos lingüísticos. Mientras que Europarl aporta un lenguaje más estructurado, formal y sintácticamente homogéneo, OANC representa un uso más natural, cotidiano y diverso del idioma, lo que implica un mayor grado de variabilidad en las expresiones de los usuarios. Los resultados obtenidos reflejan que el modelo mantiene una base sólida para la evaluación automatizada, ya que logra adaptarse a ambos contextos sin comprometer las métricas fundamentales. Sin embargo, también se observaron diferencias clave en el rendimiento, especialmente en métricas sensibles a la variación léxica, como BLEU, F1 Score y ROUGE.

Estas variaciones sugieren que, aunque el sistema es funcional, se requieren ajustes adicionales para mejorar su precisión cuando se enfrenta a respuestas menos estructuradas o fuera del patrón formal del corpus de referencia; por último, el análisis destaca la necesidad de incorporar métricas complementarias, especialmente aquellas que permitan evaluar la fluidez, la riqueza léxica y la coherencia semántica de forma más flexible. Asimismo, optimizar el uso de técnicas de normalización contextual y posibles enfoques de entrenamiento adaptativo que consideren la naturaleza del corpus como variable de entrada, mejorará significativamente una evaluación más robusta,

representativa y alineada con los estándares de los marcos de certificación como el MCER.

5.4. Entrenando el sistema con Word2Vec.

Considerando los resultados obtenidos preliminarmente, se realizaron ajustes en el sistema con el fin de aumentar la fiabilidad en relación con las métricas previamente descritas. En los experimentos iniciales, las pruebas no se generaban directamente a partir del corpus, por lo que en este ajuste GPT-J extrae fragmentos representativos del mismo y, posteriormente, elabora las pruebas para el usuario tomando como base dichos extractos. Con ello, la evaluación adquiere mayor coherencia y pertinencia para determinar el grado de dominio del idioma. Asimismo, se integró un proceso de entrenamiento basado en embeddings semánticos, cuyo objetivo es permitir que el sistema evaluador identifique y acepte términos no contemplados en el corpus base. Este mecanismo incrementa la flexibilidad en la calificación de las pruebas sin incrementar de manera significativa el coste computacional.

El entrenamiento se estructura a partir de un diccionario en formato JSON que alimenta un modelo Word2Vec, empleado como recurso auxiliar al corpus de referencia. Dicho diccionario incorpora terminología de uso común en el idioma inglés, incluyendo modismos, regionalismos, acrónimos y expresiones de carácter informal que, incluso cuando no presentan una ortografía normativa, son de uso cotidiano y socialmente aceptados (por ejemplo: Gosh, Cuz, thnkx, LOL, entre otros). Con esta estrategia, se busca reducir el sesgo latente en la calidad de la evaluación, al flexibilizar el universo léxico disponible para medir el grado de dominio del idioma.

Durante la fase de evaluación, si un término en la respuesta del usuario no coincide de manera exacta con el fragmento de corpus resumido, el sistema calcula su similitud vectorial mediante el modelo entrenado. Si el nivel de similitud resulta suficientemente alto, el término se considera aceptable en la evaluación. En consecuencia, el sistema combina el corpus seleccionado con el diccionario complementario entrenado en Word2Vec. Este procedimiento disminuye penalizaciones innecesarias en las métricas y, en paralelo, fortalece la consistencia general del proceso evaluador. De esta manera, el mecanismo propuesto fortalece el análisis de ROUGE-3 en comprensión lectora, al tiempo que en la evaluación semántica sustenta la construcción de vectores representativos con Word2Vec; por último, ROUGE-1 se mantiene como métrica principal para la evaluación de vocabulario.

El proceso de entrenamiento ha sido diseñado con un enfoque modular, lo que garantiza

que no interfiera en los flujos principales de ejecución del sistema. Puede activarse como parte de una rutina de mantenimiento o como una mejora iterativa del modelo. Tras su implementación, se repitieron las pruebas de eficiencia con los corpus OANC y Europarl a fin de comparar el desempeño antes y después de la incorporación de las mejoras propuestas. En una primera instancia, tomando como referencia el corpus Europarl, los experimentos iniciales se centraron en comparar la precisión gramatical y léxica respecto a la versión previa del sistema. Los resultados presentados en el Cuadro 5.4 evidencian una mejora significativa en la evaluación gramatical. En este proceso, el uso de herramientas como LanguageTool resultó clave para la corrección ortográfica y gramatical de las respuestas, gracias a la amplitud de su base de reglas y la precisión de su diccionario interno. Esto permitió obtener evaluaciones más fiables a través de la métrica ROUGE-1, al comparar la respuesta corregida con la original. Adicionalmente, se aplicó la Ley de Zipf para medir la distribución de frecuencias léxicas tanto en la respuesta corregida como en la original, y para identificar posibles cambios estructurales introducidos en el proceso de corrección ortográfica y gramatical.

#	Respuesta usuario	Respuesta corregida	ROUGE-1	Zipf Usuario	Zipf Corregido	Zipf Ambos textos
1	the sEntenc "now is th time for the EU to show its gogwill"means to demostrte with actions the goodwil from EU	The sentence "now is the time for the EU to show its goodwill"means to demonstrate with actions the goodwill from EU	0.8636	0.6397	0.5298	0.3535
2	the meaning of .EU is being put under presure by the threat of comisioners being removedïs if the commissioners are removd can make a political imbalans	The meaning of EU is being put under pressure by the threat of commissioners being removed is if the commissioners are removed can make a political imbalance	0.8518	0.7239	0.8103	0.3162
3	the sentence means thatt the EU is forcing to member state to tak ideas or desicions can afect their politics.	The sentence means that the EU is forcing to member state to take ideas or decisions can affect their politics.	0.8000	0.7227	0.7227	0.0
4	the EU caN answr to the acceSion of turkey in diferenT waiys, EU has stoped the negotiations to conerns about the rule of laww	The EU can answer to the accession of Turkey in different ways, EU has stopped the negotiations to concerns about the rule of law	0.7083	0.5566	0.5566	0.0
5	the sentence is saying it ´ s dificolt to figur how EU can become efficient and cuick to respond to isues related to the treatye	The sentence is saying it's difficult to figure how EU can become efficient and quick to respond to issues related to the treaty	0.7659	0.5137	0.5137	0.0

Cuadro 5.4: Resultados evaluación gramatical (Europarl y Word2Vec).

Asimismo, se observó que en los casos donde las respuestas presentaron un mayor número de correcciones gramaticales o léxicas, la desviación de Zipf tendió a alejarse de cero. Este comportamiento indica que las intervenciones correctivas modificaron la distribución de frecuencias al introducir vocabulario distinto al contemplado en la producción original del usuario. En consecuencia, se genera una mayor irregularidad en la proporción entre palabras de alta y baja frecuencia, lo cual refleja una mayor rigidez en la estructura léxica del corpus Europarl. Dicho resultado contrasta con aquellos casos en los que la desviación fue nula, donde la corrección ortográfica no alteró la naturalidad estadística del enunciado y, por tanto, la distribución de Zipf se mantuvo estable.

En cuanto a la evaluación semántica, se llevó a cabo un análisis de la relevancia y significado de las respuestas del usuario en comparación con los extractos de texto provenientes del corpus de referencia (Europarl). Mediante el modelo Word2Vec, las palabras fueron representadas en vectores distribucionales, lo que permitió capturar sus relaciones de significado. Posteriormente, la similitud de coseno se aplicó para calcular el grado de cercanía semántica entre la respuesta corregida del usuario y el fragmento resumido del corpus. Los resultados, presentados en el Cuadro 5.5, muestran valores de similitud que oscilan entre 0.86 y 0.95, lo que refleja una alta correspondencia semántica entre ambos textos. Estos valores cercanos a 1 indican que las respuestas no solo son gramaticalmente válidas, sino que además mantienen coherencia con el significado expresado en el corpus de referencia.

#	Respuesta corregida	Extracto de texto	Similitud de coseno
1	The sentence "now is the time for the EU to show its goodwill"means to demonstrate with actions the goodwill from EU	Now is the time for the European Union to demonstrate its goodwill through concrete actions, not only in words but also in deeds	0.8696
2	The meaning of .EU is being put under pressure by the threat of commissioners being removedïs if the commissioners are removed can make a political imbalance	The European Union is currently facing significant challenges regarding the credibility of its institutions. Pressure has increased due to the potential removal of commissioners, which could lead to political instability and undermine public trust.	0.8895
3	The sentence means that the EU is forcing to member state to take ideas or decisions can affect their politics	Member States have expressed concerns that the European Union is imposing decisions without sufficient consultation. The growing perception is that the EU is pressuring governments to adopt measures that directly affect national political agendas.	0.8991
4	The EU can answer to the accession of Turkey in different ways, EU has stopped the negotiations to concerns about the rule of law	The accession of Turkey to the European Union has raised questions about the rule of law and democratic standards. Negotiations have slowed down as the EU seeks to ensure that fundamental values are respected before further progress is made.	0.9550
5	The sentence is saying it's difficult to figure how EU can become efficient and quick to respond to issues related to the treaty	It is also difficult to determine how the European Union can become more responsive when the responsibilities of newly created posts remain vaguely defined in the treaty. Clearer institutional roles are needed to ensure efficiency and accountability.	0.9053

Cuadro 5.5: Resultados evaluación semántica (Europarl y Word2Vec).

En conclusión, el proceso de evaluación semántica demuestra que el uso combinado de embeddings distribucionales y similitud coseno permite establecer con precisión el nivel de coherencia y relevancia contextual entre las respuestas del usuario y los fragmentos del corpus. Este resultado respalda la eficacia del módulo para medir el significado más allá del nivel léxico o superficial.

Para la evaluación de comprensión lectora, reflejada en el Cuadro 5.6, se utilizaron las métricas ROUGE-3 y la Divergencia de Kullback-Leibler (KLD). La métrica ROUGE-3 mide la coincidencia de n-gramas entre la pregunta generada a partir del corpus y la respuesta del usuario, mientras que KLD evalúa la diferencia en la distribución probabilística del lenguaje entre ambos textos, lo que refleja el grado de variación con respecto al estilo del corpus de referencia. Tras la integración del diccionario generado mediante el entrenamiento con Word2Vec, se fortaleció la comparación en casos donde la respuesta del usuario no reproduce de manera exacta la formulación de la pregunta. El uso de palabras validadas y relaciones semánticas aprendidas permite que el sistema tolere variaciones léxicas y ofrezca una evaluación más flexible y cercana al contexto.

#	Prueba evaluación	Respuesta corregida	ROUGE-3	Kullback-Leibler
1	What does the sentence "Now is the time for the EU to show its goodwill"mean?	The sentence "now is the time for the EU to show its goodwill"means to demonstrate with actions the goodwill from EU	0.7059	4.1712
2	What does it mean that the EU is being put under pressure by the threat of commissioners being removed?	The meaning of .EU is being put under pressure by the threat of commissioners being removedïs if the commissioners are removed can make a political imbalance	0.5238	6.7685
3	What does it mean that the EU is putting pressure on Member States?	The sentence means that the EU is forcing to member state to take ideas or decisions can affect their politics.	0.2069	14.2886
4	How can the EU respond to the accession of Turkey?	The EU can answer to the accession of Turkey in different ways, EU has stopped the negotiations to concerns about the rule of law	0.2	9.3577
5	What does ït is also difficult to work out how the EU is to become more responsive if the responsibilities of the newly created posts are only outlined in the treaty"mean?	The sentence is saying it's difficult to figure how EU can become efficient and quick to respond to issues related to the treaty	0.0	9.3302

Cuadro 5.6: Resultados comprensión lectora (Europarl y Word2Vec).

Por lo tanto, los resultados de la evaluación de comprensión lectora muestran una mejora en los valores de ROUGE-3, mientras que los valores de KLD indican variaciones en la naturalidad del lenguaje usado por los usuarios frente al corpus base. Estos hallazgos confirman que el módulo es capaz de medir no solo la literalidad de la respuesta, sino también su coherencia en contexto, lo que valida la incorporación de Word2Vec como mecanismo de refuerzo en la evaluación.

Por último, tomando en cosideración los resultados de las métricas F1, BLEU y Perplexity; utilizando el corpus de Europarl, también hubo una mejoria en cuanto al sentido de las métricas debe de dar, los resultados ya no son exorbitantes ni mucho menos incoherentes, además se suavizaron las restricciones de las metricas para lograr una flexibilidad en la evaluación. El Cuadro 5.7 representa los resultados de las métricas antes mencionadas utlizando el corpus de Europarl y el entrenamiento con Word2Vec.

#	BLEU	F1 Score	Perplexity
1	0.012738779005338265	0.16896120150187735	21.999999999999996
2	0.03310872107062791	0.25510204081632654	27.0
3	0.003409006438787469	0.08771929824561403	19.999999999999996
4	0.01592574183994333	0.16523867809057527	24.000000000000004
5	0.011758872928004553	0.16163793103448276	23.0

Cuadro 5.7: Resultados comprensión lectora (Europarl y Word2Vec).

En otra instancia, utilizando el corpus OANC, se realizaron nuevamente los experimentos con el fin de verificar si, de manera análoga, se observaba una mejora en las métricas y herramientas de PLN descritas con anterioridad. Siguiendo la misma estructura de análisis, en la evaluación de gramática y vocabulario se identificó una mejora al integrar el análisis de unigramas, la distribución de Zipf y la corrección ortográfica con respecto a las respuestas proporcionadas por los usuarios. A continuación, en el Cuadro 5.8 se detallan los resultados obtenidos.

#	Respuesta usuario	Respuesta corregida	ROUGE-1	Zipf Usuario	Zipf Corregido	Zipf Ambos textos
1	the qustion "how did hee try to not tell such talss?"dont have cntext to answr	The question "how did he try to not tell such tales?"don't have context to answer	0.6875	0.6242	0.6399	0.0
2	the sentenc "such tak"meenans that is an kind of conversatiion that has beeen previosly within cntext	The sentence "such talk"means that is a kind of conversation that has been previously within context	0.5882	0.6399	0.6399	0.0
3	the mening of "talesïs abouut the descripotion of storyes arent entirely true	The meaning of "talesïs about the description of stories aren't entirely true	0.6923	0.6327	0.6494	0.0
4	"such talkrefers into a conversation wher theres a specific converstion withnn a contex	"Such talkrefers into a conversation where there's a specific conversation within a context	0.7142	0.4741	0.4968	0.0
5	amplifiyng the threath factor means makigng a danger greater or more sgnifcant	Amplifying the threat factor means making a danger greater or more significant	0.6666	0.7413	0.7413	0.0

Cuadro 5.8: Resultados evaluación gramatical (OANC y Word2Vec).

En el caso de este corpus, los resultados obtenidos mediante la Ley de Zipf mostraron una desviación de 0.0 entre el texto original y el corregido. Esto refleja que, a pesar de las correcciones ortográficas y gramaticales aplicadas, la distribución estadística de las frecuencias de palabras permaneció estable, sin alterar la estructura léxica de los

enunciados. Dicho hallazgo resulta positivo en tanto sugiere que las correcciones no afectan la naturalidad estadística del texto; sin embargo, es necesario señalar que una desviación nula no implica necesariamente una mejora en el desempeño lingüístico, sino únicamente la neutralidad de las correcciones respecto a la distribución esperada de palabras.

Mientras tanto, tal como se observa en el Cuadro 5.9 correspondiente a la evaluación semántica, se aprecia una mejora notable en los valores obtenidos mediante la similitud de coseno. Estos resultados reflejan que el modelo logra una correspondencia más sólida entre el significado y el contexto de las respuestas generadas y los fragmentos de referencia extraídos del corpus OANC. Esta ganancia puede atribuirse, en parte, a la naturaleza del corpus, el cual contiene textos diversos en inglés estadounidense, más próximos al registro y uso lingüístico cotidiano que el sistema busca evaluar. Dicha diversidad aporta un entrenamiento más equilibrado y menos sesgado hacia estructuras formales o diplomáticas, predominantes en el corpus Europarl. En consecuencia, el módulo de evaluación semántica muestra que la variación en los corpus de referencia influye directamente en la precisión de la correspondencia contextual y en la robustez del análisis de significado.

#	Respuesta corregida	Extracto de texto	Similitud de coseno
1	The question "how did he try to not tell such tales?"don’t have context to answer	The character struggles to avoid recounting stories that may not be fully true, indicating hesitation in revealing the whole truth	0.8505
2	The sentence "such talk"means that is a kind of conversation that has been previously within context	The phrase refers to a type of conversation that has already occurred within the given context, suggesting familiarity among the speakers.	0.7977
3	The meaning of "talesïs about the description of stories aren’t entirely true	The term “tales” refers to stories or accounts that are often exaggerated or not entirely accurate.	0.7074
4	"Such talkrefers into a conversation where there’s a specific conversation within a context	The phrase indicates a specific kind of conversation tied to a particular context or situation in the narrative.	0.7538
5	Amplifying the threat factor means making a danger greater or more significant	The expression implies making a danger appear greater or more significant than it initially was.	0.8636

Cuadro 5.9: Resultados evaluación semántica (OANC y Word2Vec).

En cuanto a la evaluación de la comprensión lectora, reflejada en el Cuadro 5.10, se comparó la pregunta generada a partir del corpus OANC con la respuesta proporcionada por el usuario, aplicando las métricas ROUGE-3 KLD. La primera mide la coincidencia n-grama, mientras que KLD evalúa la diferencia en la distribución probabilística del lenguaje entre ambos textos, lo que refleja la cercanía con el estilo lingüístico del corpus. Los resultados muestran que, tras el entrenamiento con embeddings semánticos, el modelo logra respuestas más cercanas a las correctas no solo desde el punto de vista léxico, sino también con una estructura y coherencia más próximas a las de un hablante

nativo. Esto confirma que el sistema es capaz de captar variaciones en el uso natural del idioma, fortaleciendo la evaluación de la comprensión lectora en contextos más diversos como los contenidos en OANC.

#	Prueba evaluación	Respuesta corregida	ROUGE-3	Kullback-Leibler
1	How did he try to not tell such tales?	The question "how did he try to not tell such tales?"don’t have context to answer	0.6364	8.7016
2	What does "such talk"mean?	The sentence "such talk"means that is a kind of conversation that has been previously within context	0.1111	15.7446
3	What does he mean by "tales¿	The meaning of "talesïs about the description of stories aren’t entirely true	0.0	16.5212
4	What does "such talkrefer to?	"Such talkrefers into a conversation where there’s a specific conversation within a context	0.1176	15.5672
5	What does .ªmplifying the threat factor"mean?	Amplifying the threat factor means making a danger greater or more significant	0.4	13.6332

Cuadro 5.10: Resultados comprensión lectora (OANC y Word2Vec).

Finalmente, el Cuadro 5.11 presenta las métricas de rendimiento BLEU, F1 y Perplexity, donde la tendencia de mejora se mantiene. El BLEU, aunque bajo, es más consistente con las condiciones de generación del modelo y la complejidad del lenguaje evaluado, mientras que el F1 Score y la Perplexity muestran valores que indican una reducción de incertidumbre en las predicciones. En conjunto, estos resultados confirman que el entrenamiento con embeddings de Word2Vec y un corpus más alineado con el contexto de evaluación logra optimizar la calidad, coherencia y precisión de las respuestas generadas.

#	BLEU	F1 Score	Perplexity
1	0.06289486652545105	0.1746031746031746	15.999999999999998
2	0.028491638576128988	0.09788654060066741	17.0
3	0.004819482498019795	0.02571595558153127	13.0
4	0.011186950208457277	0.0541871921182266	14.000000000000004
5	0.009693551957930502	0.09419152276295134	12.0

Cuadro 5.11: Resultados comprensión lectora (Europarl y Word2Vec).

Comparando los resultados de ambos corpus, se observa que Europarl mostró una mayor sensibilidad ante errores ortográficos y gramaticales, lo que refleja la influencia de su léxico especializado y repetitivo propio del lenguaje parlamentario. Esta característica hace que cualquier alteración puntual en las palabras tenga un impacto más evidente en las métricas de análisis, como la desviación respecto a la Ley de Zipf. En contraste, OANC, caracterizado por su diversidad textual y léxica, presentó desviaciones menos marcadas, ya que la amplitud de vocabulario y estilos de escritura permite diluir el efecto de errores o variaciones aisladas, ofreciendo un panorama más equilibrado de la competencia lingüística del evaluado.

El uso de un corpus más representativo del inglés cotidiano, junto con el entrenamiento mediante embeddings semánticos y la tecnología Word2Vec, proporciona un marco más robusto para la evaluación integral del dominio del idioma. Este enfoque no solo mejora la detección de la adecuación léxica y la coherencia global en las respuestas generadas, sino que también permite incorporar términos de uso frecuente, regionalismos y modismos, los cuales no suelen aparecer en textos formales como Europarl o en corpus académicos tradicionales.

Asimismo, la integración de métricas automáticas como ROUGE-1 para vocabulario, ROUGE-3 para comprensión lectora, Divergencia de Kullback-Leibler, Similitud de Coseno para semántica, F1, BLEU y Perplexity; posibilita un análisis más detallado de distintos aspectos del lenguaje, desde la precisión léxica hasta la consistencia semántica y la naturalidad de las respuestas. Esto contribuye a reducir sesgos en la evaluación, ofreciendo un sistema más flexible, realista y cercano a contextos comunicativos auténticos.

En conclusión, los resultados obtenidos validan la pertinencia de integrar corpus representativos y técnicas de aprendizaje basadas en embeddings semánticos para fortalecer la capacidad del sistema en la evaluación del dominio lingüístico de manera integral. Este avance no solo mejora el rendimiento actual del modelo, sino que también

abre nuevas líneas de investigación, tales como la incorporación de modelos más recientes de representación lingüística, entrenamiento con datos multimodales, o la adaptación dinámica de los criterios de evaluación según el perfil y nivel del usuario, promoviendo un sistema de evaluación más personalizado y efectivo.

Capítulo 6

Conclusiones

En esta investigación se desarrolló un modelo de evaluación del grado de dominio del idioma inglés basado en modelos de lenguaje, integrando técnicas de procesamiento de lenguaje natural (PLN) y métricas especializadas. Para ello, se compararon distintos modelos de la familia GPT, incluyendo GPT-Neo, GPT-J y GPT-3, determinando que GPT-J resultó el más adecuado para la generación de pruebas, mientras que el modelo **MAENI (Modelo Automático de Evaluación del Nivel de Inglés)** fungió como evaluador principal, encargado de analizar y calificar las respuestas del usuario.

Para la evaluación de la pronunciación, se integraron herramientas como Whisper y SpeechRecognition junto con el corpus fonético de LibriSpeech, lo que permitió una comparación objetiva entre las transcripciones generadas por los usuarios y las referencias auditivas estándar. En el análisis textual, se aplicaron procesos de preprocesamiento con NLTK, corrección ortográfica, análisis de frecuencia léxica y evaluación de vocabulario. La comprensión lectora se valoró con el Open American National Corpus (OANC), lo que permitió examinar con detalle la semántica, la coherencia y la estructura gramatical de los textos escritos.

En cuanto al desempeño global, se aplicaron métricas clave como F1 Score, Perplexity, ROUGE y BLEU para la comprensión textual, así como WER y PER para la pronunciación. Los resultados evidencian que el modelo puede operar de manera autónoma, reduciendo significativamente la necesidad de intervención humana y mostrando potencial para su adaptación a otros idiomas mediante ajustes en el corpus y parámetros lingüísticos.

Adicionalmente, se realizaron experimentos con el corpus Europarl, reconocido por su riqueza léxica y formalidad sintáctica. Esto permitió validar la robustez del modelo en un contexto de mayor complejidad lingüística. Asimismo, se incorporó un módulo de entrenamiento con embeddings semánticos utilizando Word2Vec, que reforzó la capacidad de MAENI para procesar entradas generadas automáticamente, aplicar corrección ortográfica y evaluar respuestas del usuario de acuerdo con métricas estándar en PLN. Aunque las métricas como BLEU, F1 y ROUGE reflejaron variaciones debido a la libertad en la redacción de respuestas frente a los textos de referencia, el sistema demostró consistencia, coherencia y estabilidad, validando su diseño modular y adaptable.

Entre las limitaciones identificadas se encuentran la dependencia de la comparación semántica directa y la necesidad de corpus más amplios y anotados. No obstante, incluso en su estado actual, MAENI constituye una alternativa viable para la evaluación automatizada del dominio del inglés, con aplicaciones en educación, autoevaluación, certificación y soporte didáctico inteligente. Además, representa un paso hacia sistemas de evaluación más flexibles, personalizados y accesibles, abriendo la puerta a futuras mejoras mediante técnicas avanzadas de aprendizaje supervisado, integración multimodal y adaptación dinámica a distintos perfiles de usuario.

Capítulo 7

Trabajo Futuro

Hasta este punto, el sistema **MAENI** presenta una estabilidad en cuanto a la evaluación de nivel de grado de domino del idioma, empleando métricas apropiadas para determinar la calidad de lenguaje por parte del hablante. La integración de embebdings semánticos como parte del entrenamiento del mismo, mejora significativamente los experimentos realizados, permitiendo una evaluación más realista, flexible y con diversas opciones de mejorar en fases futuras. A continuación, se describen las oprtunidades de mejora a largo plazo asi como las consideraciones éticas y legales que el sistema deberá adquirir en etapas posteriores a esta primera fase de desarrollo.

7.1. Oportunidades de mejora.

Aunque el modelo **MAENI** presenta una primera versión funcional capaz de aplicar evaluaciones del dominio del idioma inglés mediante GPT-J y métricas cuantificables, se identificaron oportunidades claras para su evolución en versiones futuras, así como un mejor manejo en el sesgo cultural que puede surgir en este proyecto:

- **Ampliación del corpus de entrenamiento:** Realizar pruebas adicionales con conjuntos de datos más representativos del inglés conversacional y cotidiano, como *Common Voice* de Mozilla, con el fin de evitar la sobrecorrección de expresiones válidas en registros informales o no estándar.
- **Ampliación de variantes dialectales:** Adaptar el sistema para reconocer otras variantes del inglés (británico, australiano, entre otros), mediante la incorporación de corpus específicos y ajustes semánticos que contemplen diferencias léxicas y fonéticas.

- **Optimización de la verificación semántica:** Fortalecer el módulo que evalúa la correspondencia semántica de las respuestas del usuario, integrando coincidencia léxica flexible y el uso de índices de sinónimos, con el fin de aceptar múltiples formulaciones correctas.

- **Mejora del módulo de autoentrenamiento supervisado:** Avanzar en la estrategia incremental de aprendizaje, utilizando los propios datos recolectados para mejorar el rendimiento del modelo, reduciendo la redundancia en el análisis de respuestas y minimizando sesgos relacionados con vocabulario, modismos o expresiones no convencionales.

- **Escalabilidad e infraestructura tecnológica:** Preparar la infraestructura del sistema para una implementación a gran escala, incorporando servidores optimizados que gestionen con eficiencia el reconocimiento de voz, el almacenamiento de corpus y las evaluaciones simultáneas.

- **Entorno inmersivo con avatares 3D:** Desarrollar una interfaz gráfica basada en avatares tridimensionales mediante motores como **Unity**, que permita al usuario interactuar con el sistema a través de simulaciones conversacionales realistas, conectadas al evaluador por medio de WebRequests.

En conjunto, estas oportunidades de mejora consolidan el potencial de MAENI como un sistema flexible, escalable e innovador, capaz de evolucionar hacia escenarios de evaluación más realistas, multimodales y personalizados, manteniendo como eje central la objetividad y equidad en la valoración de las competencias lingüísticas.

7.2. Consideraciones éticas y futuras certificaciones.

Dado que el sistema propuesto realiza evaluaciones automatizadas del nivel de dominio del idioma inglés, es esencial atender las implicaciones éticas asociadas a su uso. Entre las principales consideraciones se encuentran:

- Transparencia: el sistema debe comunicar al usuario cómo se evalúa su desempeño y qué métricas se emplean.

- Justicia: se minimiza el sesgo en la evaluación al incluir un entrenamiento supervisado que amplía el espectro de vocabulario reconocido, evitando penalizar injustamente respuestas válidas no contempladas inicialmente.

- Privacidad: el tratamiento de datos personales, especialmente en módulos de voz, debe cumplir con normas de protección de datos (como el GDPR en Europa o la LFPDPPP en México).

Por lo tanto, es crucial contar con el respaldo legal para garantizar la privacia y confidencialidad de la información haciendo énfasis en la captura de audio, dejando en claro, que este sistema debe ser de uspo libre, mejora constante siguiendo la metodología de desarrollo de software basado en prototipos y sin fines de lucro, para que aquellos que no pudieran contar con los recursos ecnonómicos o que aprendan por cu cuentan pedan contar con la herramienta necesaria para evaluar sus cualidades linguisticas sin tenr riesgo de ver comprometida su información. Además, se plantea como línea futura el someter el sistema a procesos de evaluación por parte de organismos certificadores reconocidos internacionalmente, como Cambridge Assessment English o ETS (TOEFL), para validar la calidad y confiabilidad de las pruebas generadas. Estas validaciones permitirían extender el alcance del sistema hacia contextos educativos formales o plataformas de certificación digital.

Referencias

Boersma, P., y Weenink, D. (2023). *Praat: Doing phonetics by computer.* Obtenido de https://www.fon.hum.uva.nl/praat/. (Recuperado de https://www.fon.hum.uva.nl/praat/)

Boitel, E., Mohasseb, A., y Haig, E. (2024). A comparative analysis of gpt-3 and bert models for text-based emotion recognition: Performance, efficiency, and robustness. En *Advances in computational intelligence systems* (pp. 567–579). doi: 10.1007/978-3-031-47508-5_44

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., y cols. (2024, marzo). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. doi: 10.1145/3641289

Chiang, C.-H., y Lee, H.-Y. (2023, May). Can large language models be an alternative to human evaluations? *arXiv preprint*. (doi: https://doi.org/10.48550/arXiv.2305.01937)

Chicaíza Chicaíza, R. M., Camacho Castillo, L. A., Ghose, G., Castro Magayanes, I. E., y Gallo Fonseca, V. T. (2023). Aplicaciones de chat gpt como inteligencia artificial para el aprendizaje de idioma inglés: avances, desafíos y perspectivas futuras. *Latam: Revista Latinoamericana de Ciencias Sociales y Humanidades*, *4*(2), 2610–2628. (doi: https://doi.org/10.35588/latam.v4i2.781)

Dai, C.-P., Ke, F., Dai, Z., West, L., Bhowmik, S., y Yuan, X. (2021). *Nsf public access repository* (Inf. Téc.). National Science Foundation. (Recuperado de https://par.nsf.gov/servlets/purl/10343548)

Das, N., Dingliwal, S., Ronanki, S., Paturi, R., Huang, Z., Mathur, P., y cols. (2024, mayo). Speechverse: A large-scale generalizable audio language model. *arXiv preprint*. (Recuperado de https://arxiv.org/abs/2405.08295) doi: 10.48550/arXiv.2405.08295

Davies, M. (2008). *The corpus of contemporary american english (coca).* Brigham Young University. (Recuperado de https://www.english-corpora.org/coca/)

de Europa, C. (2002). *Marco común europeo de referencia para las lenguas: Aprendizaje, enseñanza, evaluación*. Madrid: Instituto Cervantes. (Recuperado de https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf)

Escobar Hernández, J. C. (2021). La inteligencia artificial y la enseñanza de lenguas: una aproximación al tema. *Decires: Revista del Centro de Enseñanza para Extranjeros*, *21*, 29–44. doi: 10.22201/cepe.14059134e.2021.21.25.3

García Villarroel, J. J. (2021). Implicancia de la inteligencia artificial en las aulas virtuales para la educación superior. *Orbis Tertius UPAL*, *5*(10), 31–52. doi: 10.59748/ot.v5i10.98

Guano Merino, D. F., Herrera Andrade, Z. V., y Vallejo Barreno, C. F. (2023). Modelo de aprendizaje del idioma inglés utilizando algoritmos de machine learning. *Explorador Digital*, *7*(1), 29–43. doi: 10.33262/exploradordigital.v7i1.2451

Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., y Rashidi, P. (2022). Human activity recognition in artificial intelligence framework: A narrative review. *Artificial Intelligence Review*, *55*, 4755–4808. doi: 10.1007/s10462-021-10116-x

He, B., y Radfar, M. (2021, 2 de 8). The performance evaluation of attention-based neural asr under mixed speech input. *arXiv preprint*. Descargado de https://arxiv.org/abs/2108.01245 doi: 10.48550/arXiv.2108.01245

Ide, N., Suderman, K., y Pustejovsky, J. (2008). *The open american national corpus (oanc).* Descargado de https://www.anc.org/ (Recuperado de https://www.anc.org/)

Jurafsky, D., y Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Pearson Prentice Hall.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., y Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. (Recuperado de https://doi.org/10.1016/j.lindif.2023.102274) doi: 10.1016/j.lindif.2023.102274

Koehn, P. (2005). *Europarl: A parallel corpus for statistical machine translation.* Presentado en MT Summit 2005. Descargado de http://www.statmt.org/europarl/ (Recuperado de http://www.statmt.org/europarl/)

Kullback, S., y Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. doi: 10.1214/aoms/1177729694

LanguageTool. (2025). *Languagetool – correcteur grammatical en ligne et d'orthographe.* https://www.languagetool.org/. (Consultado el 8 de agosto de 2025)

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. En *Proceedings of the workshop on text summarization branches out (was 2004)* (pp. 74–81). Association for Computational Linguistics.

Ma, R., Qian, M., Manakul, P., Gales, M., y Knill, K. (2023, September). Can generative large language models perform asr error correction? *arXiv preprint arXiv:2307.04172*. Descargado de https://arxiv.org/abs/2307.04172 (Recuperado de https://arxiv.org/abs/2307.04172) doi: 10.48550/arXiv.2307.04172

Millière, R. (2024). Language models as models of language. *arXiv preprint arXiv:2408.07144*. Descargado de https://arxiv.org/abs/2408.07144 (Recuperado de https://arxiv.org/abs/2408.07144) doi: 10.48550/arXiv.2408.07144

Mozilla. (2015). *Mozilla common voice.* Descargado de https://commonvoice.mozilla.org/ (Recuperado de https://commonvoice.mozilla.org/)

Muñoz-Basols, J., y Fuertes Gutiérrez, M. (2024). Oportunidades de la inteligencia. la enseñanza del español mediada por tecnología. En *Tecnologías en la enseñanza del español* (pp. 344–360). (Recuperado de https://doi.org/10.4324/9781003146391-18) doi: 10.4324/9781003146391-18

OpenAI. (2022). *Introducing whisper | openai.* Descargado de https://openai.com/research/whisper (Recuperado de https://openai.com/research/whisper)

Panayotov, V., Chen, G., Povey, D., y Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. En *Proceedings of the ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210). (Recuperado de https://arxiv.org/abs/1506.03088) doi: 10.48550/arXiv.1506.03088

Papineni, K., Roukos, S., Ward, T., y Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the association for computational linguistics (acl)* (pp. 311–318). Descargado de https://aclanthology.org/P02-1040.pdf (Recuperado de https://aclanthology.org/P02-1040.pdf) doi: 10.3115/1073083.1073135

Paul, D. B., y Baker, J. M. (1992, 23 de feb). The design for the wall street journal-based csr corpus. En *Speech and natural language: Proceedings of a workshop held at harriman* (pp. 357–362). Descargado de https://aclanthology.org/H92-1073/ (Recuperado de https://aclanthology.org/H92-1073/)

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., y Irving, G. (2022). Red teaming language models with language models. En *Proceedings of the 2022 conference on empirical methods in natural language processing (emnlp)* (pp. 3419–3448). Descargado de https://aclanthology.org/2022.emnlp-main.225 (Recuperado de https://aclanthology.org/2022.emnlp-main.225) doi: 10.18653/v1/2022.emnlp-main.225

Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., de Chaumont Quitry, F., y Frank, C. (2023, 22 de 6). Audiopalm: A large language model that can speak and listen. *arXiv preprint*. Descargado de https://arxiv.org/abs/2306.12925 (Recuperado de https://arxiv.org/abs/2306.12925) doi: 10.48550/arXiv.2306.12925

Salton, G., y McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Sun, F.-K., Ho, C.-H., y Lee, H.-Y. (2019). Lamol: Language modeling for lifelong language learning. En *International conference on learning representations (iclr).* Descargado de https://arxiv.org/abs/1909.03329 (Recuperado de https://arxiv.org/abs/1909.03329) doi: 10.48550/arXiv.1909.03329

Topsakal, O., y Topsakal, E. (2022, 12). Framework for a foreign language teaching software for children utilizing ar, voicebots and chatgpt (large language models). *The Journal of Cognitive Systems*, *7*(2). doi: 10.52876/jcs.1227392

V7 Labs. (2022). *F1 score in machine learning: Intro & calculation.* Descargado de https://www.v7labs.com/blog/f1-score-guide (Recuperado de https://www.v7labs.com/blog/f1-score-guide)

Wang, B., Zou, X., Lin, G., Sun, S., Liu, Z., Zhang, W., y Chen, N. F. (2024, 23 de 6). Audiobench: A universal benchmark for audio large language models. *arXiv preprint*. Descargado de https://arxiv.org/abs/2406.16020 (Recuperado de https://arxiv.org/abs/2406.16020) doi: 10.48550/arXiv.2406.16020

Wikimedia Foundation. (2024). *Wikipedia dump database.* Descargado de https://dumps.wikimedia.org/ (Recuperado de https://dumps.wikimedia.org/)

Wikipedia. (2025, 6 de 2). *Word error rate.* Descargado de https://es.wikipedia.org/w/index.php?title=Word_Error_Rate&oldid=125822424 (Recuperado de https://es.wikipedia.org/w/index.php?title=Word_Error_Rate&oldid=125822424)

Xu, F. F., Alon, U., Neubig, G., y Hellendoorn, V. J. (2022, 13 de 6). A systematic evaluation of large language models of code. En *Maps 2022: Proceedings of the 6th acm sigplan international symposium on machine programming* (pp. 1–10). Descargado de https://doi.org/10.1145/3520312.3534862 (Recuperado de https://doi.org/10.1145/3520312.3534862) doi: 10.1145/3520312.3534862

Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., y Zhou, J. (2024, 12 de 2). Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint*. Descargado de https://arxiv.org/abs/2402.07729 (Recuperado de https://arxiv.org/abs/2402.07729) doi: 10.48550/arXiv.2402.07729

Zhang, A. (2017). *Github: Uberi/speech_recognition.* Descargado de https://github.com/Uberi/speech_recognition/ (Repositorio en https://github.com/Uberi/speech_recognition/)

Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., y Wei, F. (2023, 7 de 3). Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint*. Descargado de https://arxiv.org/abs/2303.03926 doi: 10.48550/arXiv.2303.03926

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.