[LO] [RE] [L]Ingeniería de Sistemas e Informática - UNAM [R]1

plain

UNIVERSIDAD NACIONAL DE MOQUEGUA

FACULTAD DE INGENIERÍA Y ARQUITECTURA

INGENIERÍA DE SISTEMAS E INFORMÁTICA

TESIS:

Desarrollo de una plataforma de validación de datos para la creación de corpus lingüísticos en desarrolladores, investigadores y validadores a nivel mundial.

Presentado Por:

Elmer Andres Collanqui Casapia

Asesor:

Moquegua, Diciembre de 2024

UNIVERSIDAD NACIONAL DE MOQUEGUA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ACTA DE SUSTENTACIÓN

Trabajo de investigación de titulación presentada por el bachiller Elmer Andres Collanqui Casapiaen el cumplimiento de los requisitos para obtener el título profesional de Ingeniero de Sistemas e Informática.
Cuyos resultados han sido los siguiente:
Calificativo del trabajo:.....
Calificativo de la sustentación: .....
Calificativo final: ..... Por lo expuesto, el bachiller Elmer Andres Collanqui Casapia
Ha sido declarado expedito para que se le confiera el titulo profesional de: Ingeniero de Sistemas e Informática.
En fe de ello queda asentada la presente acta.

Moquegua, 23 de diciembre de 2024

Prof. Dr. **Apellidos y Nombres**
PRESIDENTE

Prof. Dr. **Apellidos y Nombres**
VOCAL

Prof. Dr. **Apellidos y Nombres**
SECRETARIO

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mis padres, quienes han sido mi mayor fuente de inspiración y fortaleza. Su apoyo incondicional, sacrificios y palabras de aliento han sido fundamentales para que hoy alcance esta meta. A ustedes, les debo no solo mi formación académica, sino también los valores que me han guiado a lo largo de este camino.

Asimismo, extiendo mi sincero agradecimiento a la Universidad Nacional de Moquegua, por brindarme las herramientas académicas y profesionales necesarias para desarrollarme en el ámbito científico y personal. Gracias a sus docentes, por compartir su conocimiento y ser guías en este proceso, y a la institución por ofrecerme un entorno que promovió mi crecimiento como estudiante.

Abreviaturas

- **AI** – Inteligencia Artificial (Artificial Intelligence)
- **NLP** – Procesamiento de Lenguaje Natural (Natural Language Processing)
- **IA** – Inteligencia Artificial
- **API** – Interfaz de Programación de Aplicaciones (Application Programming Interface)
- **Ho** – Hipótesis Nula (Null Hypothesis)
- **Ha** – Hipótesis Alternativa (Alternative Hypothesis)
- **UI** – Interfaz de Usuario (User Interface)
- **DB** – Base de Datos (Database)
- **XML** – Lenguaje de Marcado Extensible (eXtensible Markup Language)
- **CSV** – Valores Separados por Comas (Comma-Separated Values)

Índice general

Índice de tablas

Índice de figuras

Resumen

La creación de corpus lingüísticos es crucial en diversas áreas, especialmente en la inteligencia artificial y la traducción automática. Este estudio presenta el desarrollo de una plataforma de validación de datos lingüísticos con el objetivo de mejorar la precisión y eficiencia en la creación de estos corpus. A través de un análisis detallado de los requisitos técnicos y funcionales, así como de la evaluación de su impacto en el proceso de validación, se muestra que la plataforma optimiza significativamente la validación de traducciones en comparación con métodos tradicionales. Los resultados obtenidos destacan mejoras en la precisión de las traducciones y en los tiempos de validación, lo que demuestra que esta plataforma es una solución efectiva y eficiente para la creación de corpus lingüísticos de alta calidad y su aplicación en la investigación de tecnologías emergentes como la inteligencia artificial.

Abstract

The creation of linguistic corpora is essential in areas such as artificial intelligence and machine translation. This study presents the development of a data validation platform aimed at improving the accuracy and efficiency of corpus creation. By analyzing technical and functional requirements and evaluating its impact on the validation process, the platform is shown to significantly optimize translation validation compared to traditional methods. The results highlight improvements in translation accuracy and validation times, demonstrating that this platform is an effective and efficient solution for creating high-quality linguistic corpora and its application in emerging technologies such as artificial intelligence.

Introducción

El presente proyecto tiene como objetivo el desarrollo de una plataforma para la validación de datos, con un enfoque particular en la creación de corpus lingüísticos para el idioma Aymara. La plataforma está diseñada para que los validadores puedan realizar un proceso colaborativo de validación y corrección de traducciones de palabras entre Aymara y español. Esta herramienta será útil no solo para mejorar la calidad de los corpus lingüísticos, sino también para promover la preservación y el uso correcto del idioma Aymara, que se encuentra en peligro de desaparecer debido a la falta de recursos educativos y tecnológicos adecuados.

La necesidad de una plataforma de validación surge de la creciente demanda de herramientas tecnológicas que faciliten la creación de recursos lingüísticos confiables y accesibles. Los corpus lingüísticos son fundamentales para el desarrollo de tecnologías de procesamiento de lenguaje natural (PLN), como la traducción automática, los sistemas de reconocimiento de voz y otros campos relacionados con la inteligencia artificial. Sin embargo, la calidad de estos recursos depende en gran medida de la precisión de las traducciones, lo que hace necesario contar con una plataforma eficaz para que los validadores puedan participar activamente en el proceso de validación.

El proyecto comenzó con una investigación sobre las necesidades y características del idioma Aymara, así como de las herramientas existentes para la validación de datos lingüísticos. A partir de esta investigación, se estableció el diseño y las funcionalidades esenciales de la plataforma, que incluyen un sistema de gestión de usuarios, un sistema de validación de traducciones y la posibilidad de realizar comentarios y sugerencias. A lo largo del desarrollo del proyecto, se utilizaron metodologías ágiles para asegurar una entrega progresiva de funcionalidades, lo que permitió realizar pruebas periódicas y ajustes en función de los resultados obtenidos.

En cuanto a los resultados obtenidos, la plataforma ha permitido crear una base sólida para la validación de traducciones y la recopilación de datos de calidad, que contribuirán significativamente a la creación de un corpus Aymara más completo y preciso. El proceso de validación ha mostrado ser eficiente, con la participación activa de los usuarios, lo que asegura la mejora continua de las traducciones y la calidad de los datos recolectados.

En conclusión, este proyecto ha alcanzado los objetivos establecidos al inicio, proporcionando una herramienta valiosa para la comunidad lingüística y ofreciendo un modelo escalable que podría implementarse en otros idiomas en el futuro. La plataforma continúa en constante evolución y está abierta a la incorporación de nuevas funcionalidades que respondan a las necesidades cambiantes de los usuarios y del campo de la lingüística computacional.

Capítulo 1

PLANTEAMIENTO DEL PROBLEMA

1.1. Planteamiento del Problema

1.1.1. Formulacion del Problema

- ¿De qué manera el desarrollo de una plataforma de validación de datos lingüísticos facilita la creación de corpus lingüísticos para desarrolladores e investigadores a nivel mundial?

1.1.2. Planteamiento del problema

El riesgo de desaparición de lenguas nativas es un problema global que afecta tanto la diversidad cultural como el desarrollo tecnológico. Se estima que más de 3,000 lenguas podrían desaparecer antes de que finalice este siglo, una tendencia que preocupa a organismos internacionales como la ONU y la UNESCO, que han declarado la Década Internacional de las Lenguas Indígenas (2022-2032) con el objetivo de frenar esta extinción (UNESCO, 2022c). Según estudios recientes, la digitalización y preservación de lenguas minoritarias dependen en gran medida de la calidad de los datos lingüísticos recolectados y validados, lo que requiere el uso de plataformas tecnológicas innovadoras (Bird, 2020). Además de las consecuencias culturales, la falta de datos lingüísticos verificados represen-

ta un desafío para el desarrollo de tecnologías avanzadas, como la inteligencia artificial, debido a la insuficiencia de corpus lingüísticos completos y precisos (UNESCO, 2022a; Tomaselli y Stella, 2021). El desarrollo de una plataforma de validación de datos que permita generar corpus lingüísticos verificados contribuiría significativamente al fortalecimiento de la preservación digital de las lenguas nativas y a la investigación y desarrollo tecnológico en este ámbito (Mager, Gutierrez-Vasques, Sierra, y Meza, 2018).

América Latina es una de las regiones con mayor diversidad lingüística en el mundo, pero muchas de sus lenguas indígenas están en peligro de desaparecer. Según la UNESCO, la región alberga más de 700 lenguas indígenas, pero gran parte de ellas no cuenta con los recursos ni las herramientas necesarias para ser preservadas y digitalizadas de manera efectiva (UNESCO, 2022a). Investigaciones recientes destacan que las iniciativas tecnológicas para la preservación de lenguas indígenas deben enfocarse en la validación colaborativa de datos lingüísticos, lo cual es crucial para garantizar la integridad y utilidad de los corpus generados (Gaur, Kursuncu, Sheth, Thirunarayan, y Daniulaityte, 2021). Sin estas plataformas de validación, es difícil garantizar la calidad y precisión de los corpus lingüísticos, lo que limita su aplicación en tecnologías modernas como la inteligencia artificial y el procesamiento del lenguaje natural (UNESCO, 2022c). El desarrollo de una plataforma para validar estos datos permitiría que los corpus resultantes fueran utilizados en la investigación académica y en la creación de herramientas tecnológicas avanzadas (May, 2019).

En Perú, lenguas nativas como el quechua, el aymara y otras lenguas amazónicas enfrentan desafíos importantes en términos de preservación y digitalización. A pesar de los esfuerzos realizados por instituciones educativas y gubernamentales, la creación de corpus lingüísticos validados sigue siendo limitada, lo que impide su incorporación en tecnologías emergentes, como aplicaciones de inteligencia artificial y sistemas de procesamiento del lenguaje natural (UNESCO, 2022b). En un estudio reciente sobre la preservación de lenguas indígenas en Perú, se destaca la necesidad urgente de desarrollar plataformas colaborativas de validación de datos para mejorar la calidad de los corpus lingüísticos (Zuazo, 2020). Una plataforma de este tipo permitiría no solo preservar estas lenguas, sino también facilitar su inclusión en investigaciones y tecnologías avanzadas (UNESCO-IESALC, 2022).

1.2. Objetivos

1.2.1. Objetivo General

Explorar cómo el desarrollo de una plataforma de validación de datos lingüísticos facilita la creación del corpus lingüístico para desarrolladores e investigadores a nivel mundial.

1.2.2. Objetivos Específicos

1. Identificar los requisitos técnicos y funcionales necesarios para el diseño y desarrollo de la plataforma de validación de datos lingüísticos.
2. Desarrollar un modelo preliminar para la implementación de una plataforma de validación de datos lingüísticos.
3. Identificar los recursos y metodologías necesarios para la creación de un corpus lingüístico que contemple diversas fuentes de datos y categorías lingüísticas.
4. Evaluar las posibles aplicaciones de una plataforma de validación de datos lingüísticos en la investigación de inteligencia artificial y otras tecnologías emergentes.

1.3. Justificación

La investigación sobre el desarrollo de una plataforma de validación de datos y la creación de un corpus lingüístico es de gran conveniencia en el contexto actual, donde la demanda por herramientas tecnológicas que faciliten la gestión y análisis de datos lingüísticos está en aumento. Este estudio aborda la necesidad de contar con mecanismos eficientes que permitan la recopilación, organización y validación de datos lingüísticos, favoreciendo así la creación de bases de datos robustas que puedan ser utilizadas en diversas investigaciones y aplicaciones.

Desde el punto de vista de la relevancia social, la creación de un corpus lingüístico accesible y bien estructurado permitirá a investigadores, educadores y profesionales en el área de la lingüística y la educación realizar análisis más profundos y significativos. Esto contribuye al enriquecimiento del conocimiento lingüístico y a la preservación de lenguas en peligro, lo que es esencial para mantener la diversidad cultural y lingüística de nuestra sociedad.

Las implicaciones prácticas de este estudio son múltiples. La plataforma de validación facilitará el trabajo de los validadores de datos, quienes podrán colaborar de manera más efectiva y eficiente, asegurando la calidad y fiabilidad de los datos recolectados. Además, un corpus bien diseñado permitirá la utilización de estos datos en proyectos educativos, desarrollo de software lingüístico, y en la investigación académica, favoreciendo la innovación en el campo de la lingüística y la educación.

Desde una perspectiva de valor ético, es fundamental asegurar que los datos lingüísticos sean tratados de manera responsable y ética. Este estudio promueve la transparencia y la rigurosidad en el proceso de recolección y validación de datos, garantizando que los participantes y las comunidades lingüísticas sean respetadas y que sus contribuciones sean valoradas.

Por último, la utilidad metodológica de este trabajo radica en la posibilidad de desarrollar un marco metodológico que guíe la creación y validación de datos lingüísticos. Este marco servirá como referencia para futuras investigaciones, proporcionando directrices claras sobre cómo llevar a cabo estos procesos de manera efectiva y ética, y estableciendo un estándar de calidad que beneficie a toda la comunidad lingüística.

1.4. Delimitación del Estudio

Temática: Este estudio se centrará en la **validación de traducciones** de palabras del **Aymara** al **español**. Se investigarán las metodologías para validar estas traducciones en una plataforma, tomando en cuenta la participación de los **validadores** y los **comentarios** sobre las traducciones.

Espacial: La investigación se llevará a cabo en el **ámbito digital**, utilizando una **plataforma en línea** destinada a la validación de traducciones. El estudio también tomará en cuenta los **registros** almacenados en la base de datos, cuyo acceso se realiza en diversas ubicaciones geográficas, pero el desarrollo y pruebas del sistema serán realizadas dentro de las **instalaciones** de la universidad o el entorno de trabajo del proyecto.

Temporal: El estudio se realizará a lo largo del **año 2024**, con pruebas y validación del sistema dentro de un periodo de **6 meses** (de enero a junio). Los datos que se analizarán provienen de las **acciones de los validadores** desde el lanzamiento de la plataforma hasta el periodo de investigación (2024).

1.5. Hipótesis

El uso de la plataforma de validación de datos propuesta mejora significativamente la precisión y eficiencia en el proceso de creación de corpus lingüísticos, en comparación con los métodos tradicionales empleados por los validadores de datos.

1.6. Operacionalización de Variables

Tabla 1.1: Operacionalización de Variable Independiente

Objetivo: Analizar cómo el desarrollo de la plataforma de validación facilita la creación de corpus lingüísticos.		
Hipótesis: La implementación de una plataforma de validación de datos lingüísticos facilita la creación de corpus lingüísticos de calidad.		
Definición Conceptual: Creación de un sistema tecnológico para validad datos lingüísticos de manera eficiente y colaborativa.		
Definición Operacional: Implementación de una plataforma con funcionalidades especificas para la gestión y validación de datos lingüísticos.		
Dimensiones	**Indicadores**	**Ítems**
Plataforma de validación de datos lingüísticos.	Existencia de la plataforma (si/no).	¿La plataforma esta implementada correctamente?
	Número de validaciones realizadas.	¿Cuantas validaciones se realizaran por mes?
	Funcionalidades implementadas	¿Cuales son las principales funcionalidades de la plataforma?

Tabla 1.2: Operacionalización de Variable Dependiente

Objetivos: Evaluar como la plataforma de validación mejora la creación de corpus lingüísticos.		
Hipótesis: La plataforma de validación mejora la cantidad y calidad de los corpus lingüísticos generados.		
Definición Conceptual: Mejoras en la calidad y cantidad de corpus lingüísticos generados gracias a un proceso eficiente de validación		
Definición Operacional: Incremento en la cantidad de corpus lingüísticos generados y validados mediante el uso de la plataforma		
Dimensiones	**Indicadores**	**Ítems**
Creación y validación de corpus lingüísticos.	Numero de corpus generados.	¿Cuantos corpus lingüísticos se han generado?
	Porcentaje de validaciones completadas.	¿Que porcentaje de las validaciones han sido completadas?
	Tiempo promedio para validar un corpus.	¿Cuanto tiempo tarda en promedio la validacion de un corpus?.

Capítulo 2

Marco teórico

2.1. Marco Teórico

2.1.1. Estado del Arte

(Zevallos y otros, 2022) presentan 'Huqariq: Un corpus multilingüe de habla de lenguas nativas del Perú para el reconocimiento de voz'. Este estudio presenta un corpus de habla transcrito de lenguas nativas peruanas, diseñado para la investigación y desarrollo de tecnologías de reconocimiento de voz. Utilizando una metodología de crowdsourcing, se recopilaron 220 horas de audio transcrito en varias lenguas indígenas. Los resultados indican que el corpus es el más grande para lenguas nativas en Perú y se espera que contribuya significativamente al desarrollo de herramientas de reconocimiento automático de voz. La conclusión resalta la importancia del corpus para preservar y revitalizar lenguas en peligro .

(Zevallos y otros, 2023) abordan en su trabajo Avances en el reconocimiento automático de voz para lenguas indígenas: Quechua, Guaraní, Bribri, Kotiria y Wa'ikhana el desarrollo de modelos de reconocimiento automático de voz para diversas lenguas indígenas, incluyendo el Quechua. Se aplicó una metodología experimental con pruebas comparativas entre diferentes modelos. Los resultados muestran mejoras significativas en la precisión del reconocimiento para estas lenguas. La conclusión sugiere que estos avances

son cruciales para la creación de plataformas que faciliten la validación y uso práctico del reconocimiento de voz en lenguas indígenas .

(Pérez y otros, 2022) presentan Construcción de un recurso para lenguas en peligro en el aula: Dependencias universales para el Kakataibo. Este trabajo se centra en la creación de un árbol de dependencias para el Kakataibo, una lengua en peligro en la Amazonía peruana. Se utilizó un enfoque metodológico basado en la recopilación y análisis lingüístico. Los resultados indican que este recurso es fundamental para la enseñanza y validación del Kakataibo en entornos educativos. La conclusión enfatiza que este tipo de recursos son esenciales para preservar lenguas amenazadas mediante plataformas digitales .

(Mendoza y Salas, 2023) exploran Aprovechando el poder de la inteligencia artificial para revitalizar lenguas indígenas en peligro: Tecnologías y experiencias. Este estudio investiga cómo la inteligencia artificial puede aplicarse a la documentación y revitalización de lenguas indígenas en peligro, con ejemplos específicos de proyectos en Brasil y Perú. Se utilizó una metodología mixta que combina estudios de caso y análisis tecnológico. Los resultados muestran que las tecnologías emergentes pueden facilitar la creación y validación de corpora lingüísticos efectivos. La conclusión destaca que integrar IA en estas iniciativas es clave para su éxito a largo plazo .

Según (Aguilar Santiago y García Zúñiga, 2023) en el artículo 'Tecnologías del lenguaje aplicadas al procesamiento de lenguas indígenas en México: una visión general' presenta una visión general del estado actual de las tecnologías lingüísticas en México para el procesamiento de lenguas indígenas. Su objetivo es mostrar el panorama general de estas tecnologías y su potencial para ayudar en la conservación digital y reducir la brecha tecnológica que enfrentan las comunidades de hablantes indígenas. La metodología se basa en un censo preliminar que destaca recursos como el Corpus Axolotl y aplicaciones móviles como Vamos a aprender, que incluyen lenguas como el náhuatl, mixteco y purépecha. La conclusión subraya la importancia de estas tecnologías para preservar y difundir las lenguas indígenas mexicanas, y la necesidad de seguir desarrollando recursos específicos para reducir la brecha tecnológica y promover su uso en comunidades indígenas.

Según (Martínez Musiño y Valdez Ramos, 2015) en el artículo 'El uso de las nuevas tecnologías y las lenguas y culturas indígenas' presenta el Multimedia de Lengua y Cultura Nahua de la Huasteca como un recurso didáctico para la enseñanza de lenguas indígenas.

Su objetivo es mostrar cómo las nuevas tecnologías pueden apoyar la enseñanza y preservación de las lenguas indígenas. La metodología se centra en el diseño y las funciones didácticas del multimedia, que incluye recursos interactivos para facilitar el aprendizaje. Los resultados indican que este tipo de recursos pueden ser efectivos para la enseñanza de lenguas indígenas en entornos educativos. La conclusión destaca el papel de las nuevas tecnologías en la enseñanza de lenguas indígenas como segundas lenguas.

Según (Ahmadbek, 2024) en el artículo 'Determinación de problemas semánticos funcionales de términos logísticos mediante lingüística de corpus' busca identificar y analizar los problemas semánticos funcionales de términos logísticos utilizando técnicas de lingüística de corpus. Su objetivo es mejorar la comprensión y el uso adecuado de estos términos en contextos logísticos. La metodología se basa en el análisis de un corpus de textos relacionados con la logística, utilizando herramientas de procesamiento de lenguaje natural para identificar patrones y problemas semánticos en el uso de términos logísticos. Los resultados indican que existen problemas significativos en la interpretación y uso de términos logísticos debido a ambigüedades semánticas y falta de estandarización, destacando áreas específicas donde estos problemas son más comunes. La conclusión destaca la importancia de utilizar la lingüística de corpus para mejorar la precisión y consistencia en el uso de términos logísticos, lo cual puede contribuir a una comunicación más efectiva en el sector logístico.

Según (Sierra y cols., 2017) en el artículo ' Corpus lingüístico: estudio y aplicación en revitalización de lenguas indígenas' explora cómo los corpus lingüísticos pueden ser utilizados en la revitalización de lenguas indígenas, ofreciendo perspectivas valiosas para proyectos que buscan preservar y promover idiomas en peligro de extinción. La metodología se basa en el análisis de corpus lingüísticos existentes para lenguas indígenas, destacando su potencial para documentar y analizar estructuras lingüísticas, léxico y patrones de uso. Los resultados indican que los corpus lingüísticos son herramientas efectivas para la revitalización, ya que permiten la creación de recursos educativos, materiales didácticos y plataformas interactivas que fomentan el aprendizaje y uso de estas lenguas. La conclusión destaca la importancia de los corpus lingüísticos en la revitalización de lenguas indígenas, subrayando su capacidad para apoyar la documentación, enseñanza y promoción de idiomas en peligro de extinción.

Segun (Gutiérrez-Fandiño, Pérez-Fernández, Armengol-Estapé, Griol, y Callejas, 2022)

en el artículo ' esCorpius: A Massive Spanish Crawling Corpus ' presenta esCorpius, un corpus de rastreo en español obtenido de aproximadamente 1 petabyte de datos de Common Crawl. Su objetivo es proporcionar un recurso extenso y de alta calidad para el desarrollo de modelos de lenguaje en español. La metodología se basa en la extracción de datos de Common Crawl, utilizando técnicas avanzadas de limpieza y de duplicación para asegurar la calidad del corpus. Los resultados indican que esCorpius es el corpus más extenso en español, superando a otros en términos de tamaño y calidad. Se destaca su capacidad para mantener los límites de documento y párrafo, lo que permite a los modelos de lenguaje natural procesar el texto de manera similar a como lo hacen los humanos. La conclusión destaca la importancia de esCorpius para el desarrollo de modelos de lenguaje en español, ofreciendo un recurso valioso para la investigación y el desarrollo de aplicaciones lingüísticas avanzadas.

2.2. Bases Teóricas

2.2.1. Bases Teóricas

El desarrollo de una plataforma para la **validación de datos** en la creación de **corpus lingüísticos** está fundamentado en teorías y enfoques que abordan el procesamiento del lenguaje natural, la validación de datos y la creación de bases de datos lingüísticas. A continuación, se presentan las principales bases teóricas que sustentan esta investigación.

1. **Teoría del Procesamiento del Lenguaje Natural (PLN)**: El procesamiento del lenguaje natural (PLN) es un área de la inteligencia artificial que se ocupa de la interacción entre las computadoras y el lenguaje humano. Según (Jurafsky y Martin, 2020), el PLN permite a las máquinas comprender, interpretar y generar lenguaje de manera similar a los humanos. En el contexto de la creación de corpus lingüísticos, el PLN se aplica para analizar y estructurar grandes volúmenes de datos lingüísticos de manera eficiente, facilitando la creación de recursos como diccionarios, traducciones y modelos de lenguaje.

2. **Teoría de la Validación de Datos**: La validación de datos es un proceso crucial para garantizar la calidad y fiabilidad de los datos recopilados en cualquier campo

de estudio. Según (Redman, 2001), la validación de datos implica la comprobación y verificación de la exactitud, consistencia y relevancia de los datos antes de su uso. En el caso de los corpus lingüísticos, la validación se centra en asegurar que las traducciones y transcripciones sean precisas y estén alineadas con las normas lingüísticas.

3. **Teoría de la Creación de Corpus Lingüísticos**: Un corpus lingüístico es una colección estructurada de textos que representa un conjunto de datos lingüísticos de interés. Según (Sinclair, 2004), los corpus lingüísticos permiten el análisis empírico del lenguaje, siendo una herramienta esencial para los estudios lingüísticos, el desarrollo de modelos de lenguaje y la mejora de tecnologías de PLN. La creación de un corpus lingüístico de alta calidad requiere una validación rigurosa de las traducciones y datos lingüísticos para garantizar su fiabilidad.

2.2.2. Marco Conceptual

- **Corpus Lingüístico:** Un **corpus lingüístico** es una colección de textos que han sido recopilados de manera sistemática para su análisis lingüístico. Los corpus lingüísticos son herramientas fundamentales en la investigación lingüística, especialmente en áreas como el procesamiento del lenguaje natural (PLN), la traducción automática, y la lexicografía. Estos corpus pueden estar compuestos por textos escritos o hablados, y pueden ser utilizados para diversos fines, como la creación de diccionarios, la evaluación de modelos de traducción y la identificación de patrones lingüísticos.

  Un corpus lingüístico puede ser categorizado de diferentes maneras, dependiendo de los textos que lo componen. Por ejemplo, un **corpus paralelo** incluye textos que están disponibles en más de un idioma, lo que lo convierte en una fuente valiosa para el entrenamiento y la evaluación de sistemas de traducción automática (Bird, Klein, y Loper, 2009). El **corpus monolingüe**, por otro lado, consiste en textos en un solo idioma y se utiliza principalmente para estudiar características lingüísticas internas de un idioma particular.

- **Plataformas de Validación de Datos:** Las **plataformas de validación de datos** permiten a los usuarios interactuar con los datos para garantizar su precisión y coherencia. Estas plataformas son fundamentales en el desarrollo de aplicaciones que

requieren datos de alta calidad para entrenar modelos, como en la creación de sistemas de traducción automática, el análisis de sentimientos o el desarrollo de diccionarios digitales. Las plataformas de validación son herramientas que facilitan la participación de diferentes validadores, lo cual es crucial para garantizar que el corpus lingüístico esté libre de errores y sea representativo del uso real del idioma.

En el contexto de la creación de un corpus lingüístico, las plataformas permiten que los traductores o validadores no solo verifiquen las traducciones, sino también que sugieran mejoras, corrijan errores o añadan contexto adicional. Estas plataformas deben ser accesibles, intuitivas y confiables para asegurar una alta tasa de participación y asegurar la calidad del trabajo realizado.

- **Índice Kappa y Fiabilidad Interevaluador:** El índice **Kappa de Cohen** es una métrica estadística que se utiliza para evaluar la concordancia entre dos o más evaluadores, teniendo en cuenta la posibilidad de que el acuerdo ocurra por azar. Esta métrica es especialmente relevante en la validación de datos, donde se requiere asegurar que diferentes validadores lleguen a conclusiones similares sobre la corrección o validez de los datos (Cohen, 1960). El índice Kappa se define como:

$$K = \frac{P_o - P_e}{1 - P_e}$$

Donde:

- P_o es la proporción de acuerdo observado entre los evaluadores.
- P_e es la proporción de acuerdo esperado por azar.

El valor de K varía entre -1 y 1, donde 1 indica un acuerdo perfecto, 0 indica que el acuerdo observado es el mismo que el esperado por azar, y valores negativos indican un acuerdo peor que el azar.

- **Interpretación del Índice Kappa:** El valor de Kappa se interpreta de la siguiente manera:

- $K = 1$: Acuerdo perfecto.
- $0.81 \leq K < 1$: Acuerdo casi perfecto.
- $0.61 \leq K < 0.80$: Acuerdo sustancial.

- $0.41 \leq K < 0.60$: Acuerdo moderado.
- $0.21 \leq K < 0.40$: Acuerdo bajo.
- $K \leq 0.20$: Acuerdo muy bajo o casi nulo.

En el contexto de validación de un corpus lingüístico, un índice Kappa alto es indicativo de que los validadores están de acuerdo en su evaluación de las traducciones o datos, lo cual es crucial para garantizar la calidad del corpus final (Viera y Garrett, 2005).

- **Fiabilidad y Validación en la Creación de Corpus:** La **fiabilidad** se refiere a la consistencia de los resultados obtenidos por un instrumento o evaluador. En el caso de un corpus lingüístico, la fiabilidad es esencial para asegurar que los datos validados sean consistentes a lo largo del tiempo y entre diferentes validadores. Una alta fiabilidad interevaluador, que se mide a través de índices como el Kappa de Cohen, indica que los evaluadores son consistentes en sus decisiones de validación, lo que contribuye a la calidad del corpus.

  Por otro lado, la **validación** es el proceso de asegurar que los datos son correctos y adecuados para su propósito. En el caso de un corpus lingüístico, la validación no solo involucra la verificación de las traducciones, sino también la identificación y corrección de errores, la adición de contexto cultural y la evaluación de la precisión de las traducciones en diferentes contextos. Este proceso es esencial para crear un corpus que sea representativo y útil para futuras investigaciones y aplicaciones.

- **Plataformas de Creación de Corpus y Herramientas de Validación:** Existen diversas herramientas tecnológicas que permiten la creación y validación de corpus lingüísticos. Entre estas, se incluyen sistemas de etiquetado automático, plataformas colaborativas y herramientas basadas en la inteligencia artificial. Estas plataformas permiten que los validadores no solo evalúen las traducciones, sino que también colaboren en tiempo real, sugieran mejoras y participen en el proceso de creación de corpus de manera eficiente.

  Algunas plataformas populares en la creación de corpus incluyen **ELAN** (una herramienta para la anotación lingüística), **TBL** (para etiquetado de texto), y **CrowdFlower** (una plataforma de crowdsourcing para la recolección y validación de datos). Estas herramientas permiten una colaboración eficaz entre múltiples usuarios,

asegurando que el corpus lingüístico final sea de alta calidad y adecuado para su propósito (Bird y cols., 2009).

2.3. Marco Legal

La Constitución del Perú y normativas específicas, como la Ley N° 29735, establecen un marco legal para la promoción y protección de los derechos lingüísticos de las comunidades indígenas. Estas disposiciones legales respaldan iniciativas que buscan preservar y revitalizar lenguas en peligro mediante la integración de tecnologías modernas. Este marco proporciona una base sólida para desarrollar plataformas tecnológicas que contribuyan a la preservación cultural y lingüística del país.

Capítulo 3

Marco Metodológico

3.1. Enfoque de Investigación

El presente estudio adopta un enfoque cuantitativo, el cual permite analizar de manera objetiva y medible las características y el desempeño de la plataforma de validación de datos para la creación de corpus lingüísticos. Este enfoque se fundamenta en la recolección y análisis de datos numéricos, con el objetivo de evaluar la efectividad de la herramienta desarrollada y su impacto en los procesos de validación y creación de corpus lingüísticos basándonos principalmente en las métricas Kappa explicadas en (de Ullibarri Galparsoro y Pita Fernández, 1999).

3.2. Tipo de Investigación

El presente estudio se clasifica como una investigación exploratoria y descriptiva, ya que busca indagar en un área poco estudiada, como es el uso de plataformas tecnológicas para la validación de datos lingüísticos, y, al mismo tiempo, describir de manera detallada las características y el desempeño de la herramienta desarrollada. Desde el enfoque exploratorio, se analiza el contexto, las problemáticas y las oportunidades relacionadas con los procesos de validación de datos en la creación de corpus lingüísticos. Por su parte, el

enfoque descriptivo permite documentar variables clave como el tiempo promedio de validación, la cantidad de palabras procesadas y la percepción de usabilidad de la plataforma. Esta combinación de enfoques facilita tanto la comprensión inicial del fenómeno como la generación de información clara y precisa que contribuya al avance en este campo de estudio.

3.3. Diseño de Investigación

El diseño de la investigación es de tipo no experimental y transversal. Este diseño se adecúa a los objetivos del estudio, ya que no se manipulan las variables de manera directa, sino que se observan y analizan tal como se presentan en su contexto natural. Además, se realiza un análisis en un único punto en el tiempo, permitiendo obtener una visión precisa y puntual del fenómeno estudiado.

En este caso, el diseño no experimental se justifica porque el desarrollo de la plataforma de validación de datos no requiere alterar las condiciones del entorno de los participantes (validadores, desarrolladores e investigadores). Más bien, se busca analizar cómo interactúan con la herramienta y cómo esta influye en los procesos de validación de datos y creación de corpus lingüísticos.

El diseño transversal se emplea para recolectar datos en un momento específico del desarrollo y uso de la plataforma, permitiendo identificar patrones, características y posibles limitaciones en su implementación. Este enfoque asegura una evaluación objetiva de las funcionalidades de la herramienta y del impacto que tiene en los procesos asociados (Hernández Sampieri, Fernández Collado, y Baptista Lucio, 2014).

3.4. Método de Investigación

hipotetco deductivo

3.5. Población y Muestra

La población objeto de estudio está constituida por los **investigadores, desarrolladores y validadores de datos** que participarán en la validación de las traducciones en la plataforma de creación de corpus lingüístico, tanto a nivel nacional como internacional. Debido a que la población total de estos participantes es incierta y se desconoce con exactitud, se considera una **población infinita** o suficientemente grande para aplicar las fórmulas estándar de muestreo en estadística.

Para determinar el tamaño de la muestra se utilizó la fórmula para **población infinita**, ajustada con un **margen de error del 10 %** y un **nivel de confianza del 95 %**. Este margen de error se seleccionó debido a que, al ser una población incierta y a nivel global, se acepta un margen de error ligeramente mayor para equilibrar la precisión de los resultados con la necesidad de contar con una muestra representativa.

La fórmula utilizada para calcular el tamaño de la muestra es la siguiente:

$$n = \frac{Z^2 \cdot p \cdot (1 - p)}{E^2}$$

Donde:

- $Z = 1.96$ es el valor de la distribución normal estándar, correspondiente a un nivel de confianza del 95 %. Este valor asegura que el intervalo de confianza cubra el 95 % de los posibles resultados.

- $p = 0.5$ es la proporción estimada de la población. Este valor se utiliza cuando no se tiene una estimación precisa de la proporción, ya que maximiza el tamaño de la muestra.

- $E = 0.10$ es el margen de error, que se estableció en 10 % debido a la incertidumbre sobre el tamaño exacto de la población. Este margen de error representa la precisión de los resultados, lo que implica que los resultados pueden variar hasta un 10 % respecto a la población total.

Sustituyendo los valores en la fórmula:

$$n = \frac{(1.96)^2 \cdot 0.5 \cdot (1 - 0.5)}{(0.10)^2}$$

$$n = \frac{(3.8416) \cdot 0.25}{0.01}$$

$$n = 96.04$$

Por lo tanto, el tamaño de la muestra requerido es de **96 participantes**. Este tamaño es suficiente para obtener conclusiones con un margen de error razonable, considerando la incertidumbre de la población y la necesidad de obtener datos representativos para evaluar el funcionamiento de la plataforma en una etapa temprana.

El tamaño de la muestra seleccionado permite obtener resultados válidos sin requerir una gran cantidad de participantes, lo que es especialmente útil dada la fase preliminar de la plataforma. Sin embargo, los resultados obtenidos con esta muestra de **96 participantes** se consideran representativos para el análisis de la eficacia de la plataforma en la validación de datos y la creación del corpus lingüístico, con un margen de error aceptable.

3.6. Técnicas e Instrumentos de Recolección de Datos

Para la recolección de datos en esta investigación, se emplearán técnicas e instrumentos adecuados que permitan obtener la información necesaria para evaluar la eficacia y el funcionamiento de la plataforma de validación de datos para la creación de corpus lingüístico. A continuación se detallan las técnicas, los instrumentos utilizados, y las consideraciones sobre la validez y confiabilidad de los mismos.

3.6.1. Técnicas de Recolección de Datos

La técnica principal que se utilizará es la **encuesta**, ya que permite recopilar información de una muestra amplia de participantes en un corto período de tiempo. Dado que la investigación se orienta a una plataforma digital y se pretende obtener la opinión de los validadores de datos y usuarios, se considera que la encuesta es el método más adecuado para recolectar datos de forma estructurada.

Adicionalmente, se podrá emplear la **observación directa** en algunos casos específicos, para identificar cómo los validadores interactúan con la plataforma y detectar posibles dificultades o problemas de usabilidad.

3.6.2. Instrumentos de Recolección de Datos

Los instrumentos que se utilizarán en esta investigación incluyen:

- **Cuestionarios**: Se diseñará un cuestionario estructurado que será enviado a los usuarios y validadores de datos. El cuestionario incluirá preguntas cerradas y abiertas sobre la usabilidad de la plataforma, la claridad de las instrucciones, la facilidad de uso, y la precisión de las traducciones. Las preguntas estarán enfocadas en la experiencia del usuario y en la efectividad de la plataforma como herramienta de validación.
- **Registros de interacción en la plataforma**: Además de las encuestas, se utilizarán los registros de interacción dentro de la plataforma para obtener datos cuantitativos sobre el uso de las funcionalidades, el tiempo de interacción y los errores comunes reportados por los usuarios.

3.6.3. Validez de los Instrumentos

La validez de los instrumentos de recolección de datos será garantizada mediante el proceso de **validación de contenido**. Esto implicará la revisión de los cuestionarios por expertos en el área de lingüística y tecnologías de la información, quienes evaluarán si las

preguntas cubren adecuadamente los aspectos relevantes de la plataforma y si permiten obtener información significativa para el análisis. Además, se realizará una prueba piloto con una pequeña muestra de usuarios para identificar posibles fallos en la interpretación de las preguntas y asegurar que el cuestionario sea adecuado para el propósito de la investigación.

3.6.4. Confiabilidad de los Instrumentos

La confiabilidad de los instrumentos será evaluada mediante el cálculo del **coeficiente de fiabilidad de Cronbach** (α). Este coeficiente mide la consistencia interna del cuestionario y asegura que las preguntas estén relacionadas entre sí de manera coherente. Se espera obtener un valor de α mayor a 0.7, lo que indicaría que el cuestionario es confiable para medir las variables de interés de manera consistente. Si el coeficiente es bajo, se revisarán las preguntas y se harán ajustes para mejorar la coherencia interna.

Adicionalmente, la **observación directa** también se complementará ya que se basaran netamente en cálculos estadísticos que sacaremos delos datos que obtengamos.

3.7. Procedimiento

El procedimiento para la creación de la plataforma de validación de datos consta de varias etapas fundamentales, que van desde la planificación y el diseño hasta la implementación y evaluación. Este proceso incluye el desarrollo tanto de la parte técnica como de los aspectos metodológicos, garantizando que la plataforma cumpla con los requisitos de validación y que sea adecuada para su uso por parte de los validadores de datos.

- **Fase 1: Planificación y Definición de Requisitos**
  En esta fase inicial, se definen los requisitos generales de la plataforma. Esto incluye la identificación de los usuarios finales (validadores de datos), el tipo de datos que se van a validar (traducciones de palabras), y las funcionalidades esenciales de la plataforma, como la capacidad de gestionar usuarios, realizar validaciones, almacenar comentarios, y generar reportes de progreso. Además, se establece el diseño de

la base de datos y las interacciones de usuario necesarias para que la plataforma sea eficiente y fácil de usar.

- **Fase 2: Diseño de la Plataforma**
  En esta fase, se desarrollan los planos detallados de la interfaz de usuario (UI) y la arquitectura de la plataforma. El diseño de la interfaz debe ser intuitivo, permitiendo a los validadores interactuar con los datos sin dificultades. También se especifica el diseño de la base de datos, que incluye tablas para usuarios, traducciones, validaciones, y comentarios. Se elige la tecnología adecuada para el desarrollo, considerando la escalabilidad, la seguridad y la facilidad de mantenimiento de la plataforma.

- **Fase 3: Desarrollo e Implementación**
  Durante esta fase, se lleva a cabo el desarrollo real de la plataforma. Esto incluye la programación de la interfaz de usuario, la implementación de las funcionalidades principales (como el registro de usuarios, validación de datos, y la generación de reportes), y la conexión con la base de datos. El equipo de desarrollo utiliza tecnologías como HTML, CSS, JavaScript, y frameworks como React para el front-end, y Node.js con Express para el back-end. Además, se configuran los sistemas de autenticación y autorización de usuarios para garantizar la seguridad de la plataforma.

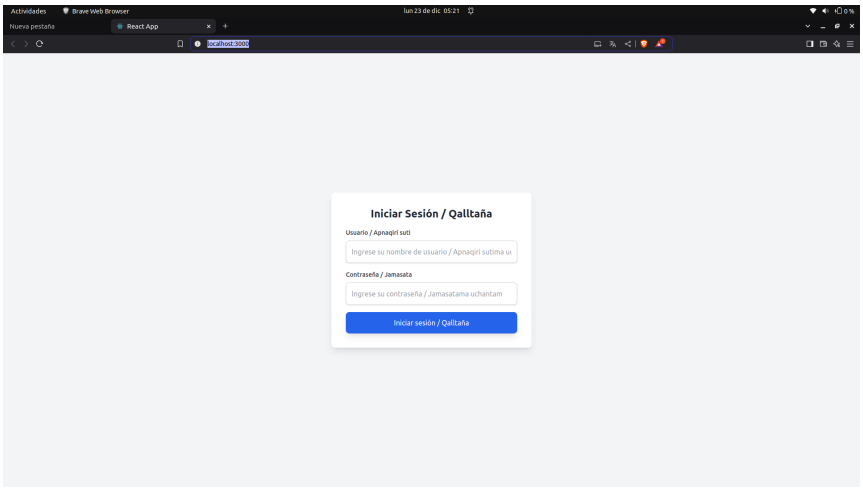

Figura 3.1: Login de la página

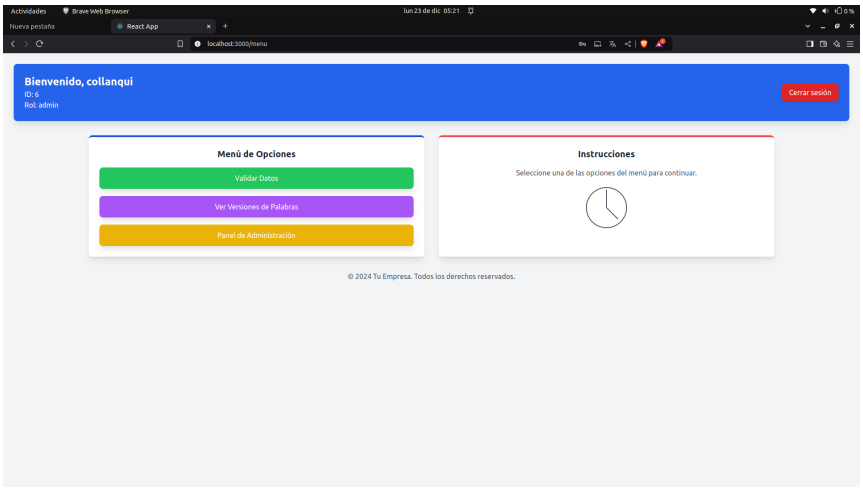

Figura 3.2: Interfaz principal

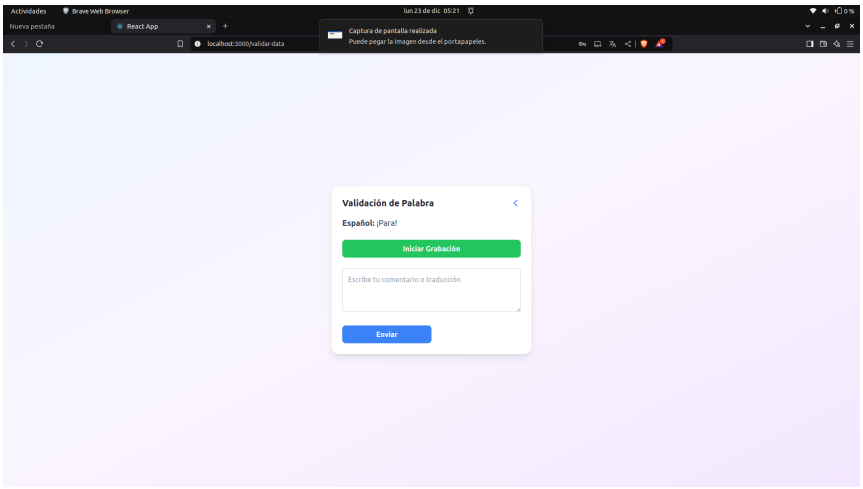

Figura 3.3: Interfaz para validar palabra

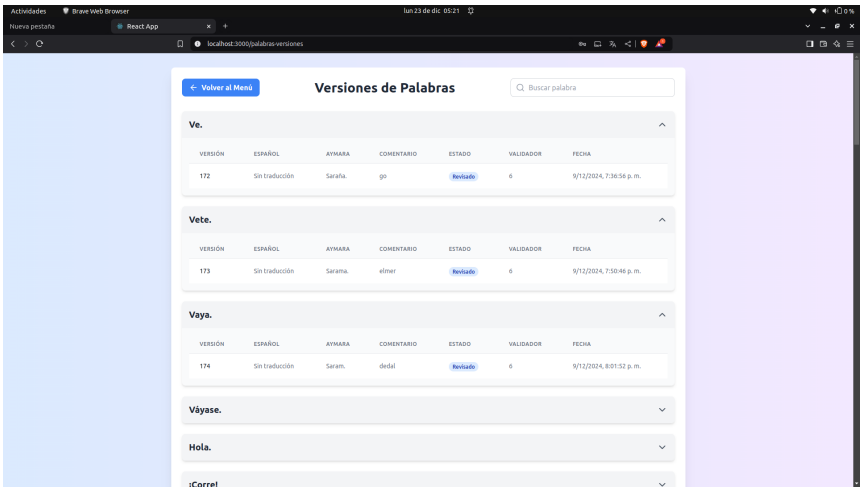

Figura 3.4: Interfaz para ver el corpus

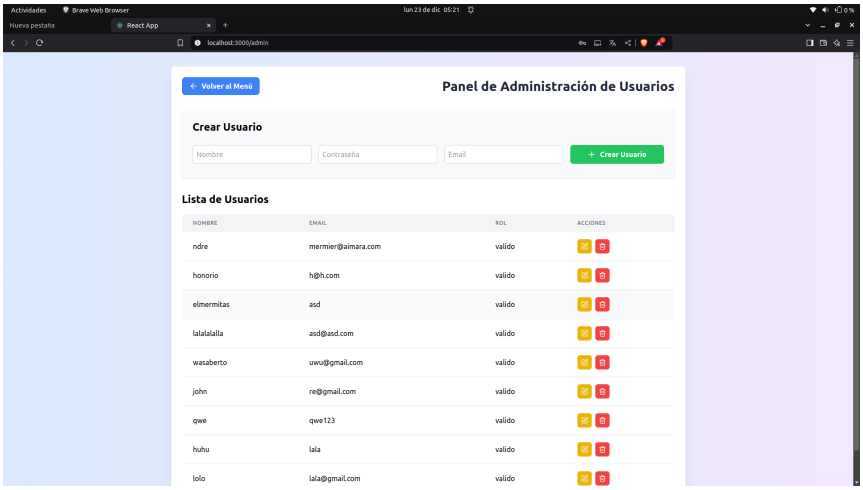

Figura 3.5: Interfaz para administrar cuentas

- **Fase 4: Integración de la Plataforma con la Base de Datos**
  Una vez que la plataforma ha sido desarrollada, se realiza la integración con la base de datos. Se asegura que las interacciones entre la plataforma y la base de datos sean eficientes, utilizando consultas SQL para la gestión de usuarios, traducciones y validaciones. La base de datos se estructura para almacenar no solo la información sobre los usuarios y las traducciones, sino también el historial de validaciones y comentarios de los validadores. Este sistema es crucial para mantener un registro completo de las acciones realizadas dentro de la plataforma.

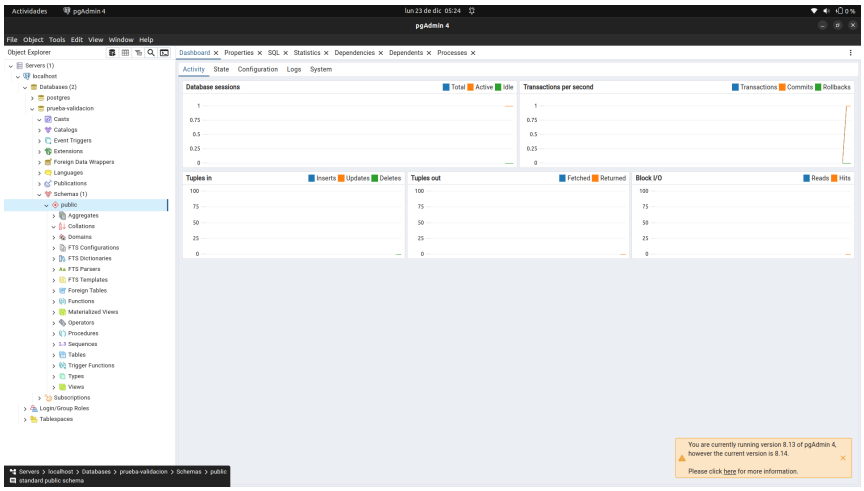

Figura 3.6: Relación con la BD

- **Fase 5: Pruebas y Validación**
  Una vez que la plataforma está integrada y funcionando, se realiza un conjunto de pruebas para verificar que todos los componentes funcionan correctamente. Esto incluye pruebas de funcionalidad para asegurarse de que las validaciones se registran adecuadamente, las interacciones de los usuarios son correctas, y la base de datos responde de manera eficiente. Se llevan a cabo pruebas de usabilidad para confirmar que la interfaz es fácil de usar, y también se realizan pruebas de seguridad para asegurar que la plataforma es resistente a vulnerabilidades comunes.

- **Fase 6: Implementación Piloto**
  Después de la fase de pruebas, se realiza una implementación piloto de la plataforma con un grupo selecto de validadores. Este piloto permite recoger comentarios de los usuarios sobre el funcionamiento de la plataforma y detectar posibles áreas de mejora. Durante esta fase, se realiza un monitoreo continuo del sistema para asegurar que la plataforma está funcionando sin problemas y que los usuarios pueden completar sus tareas de validación de manera eficiente.

- **Fase 7: Ajustes y Mejoras**
  Basado en los comentarios del piloto, se implementan ajustes y mejoras en la plataforma. Esto puede incluir modificaciones en la interfaz de usuario, optimización de las consultas a la base de datos, o la corrección de errores detectados durante el uso

real de la plataforma. Una vez realizados los ajustes necesarios, la plataforma está lista para su implementación completa.

- **Fase 8: Lanzamiento Final y Monitoreo Continuo**
  En esta fase final, la plataforma es lanzada oficialmente para su uso generalizado. A partir de este momento, el sistema se mantiene bajo monitoreo continuo para asegurar que los usuarios puedan acceder sin problemas y que las validaciones se realicen con la mayor calidad posible. El monitoreo también permite la detección temprana de posibles fallos técnicos, los cuales pueden ser corregidos rápidamente para mantener la integridad de la plataforma.

Capítulo 4

Análisis e Interpretación de los Resultados

4.1. Análisis e Interpretación de los Resultados

4.1.1. Participación de los Validadores

La participación activa de los validadores es crucial para garantizar la eficacia del sistema de validación. Durante el período analizado, se observó un alto nivel de compromiso por parte de los usuarios, lo que se refleja en la cantidad de palabras validadas y los comentarios enviados. La distribución de las validaciones entre los usuarios muestra una contribución equilibrada, con algunos validadores destacando por su mayor aporte. Este comportamiento demuestra que la plataforma incentiva la colaboración activa, lo que es esencial para mantener la calidad y confiabilidad del sistema.

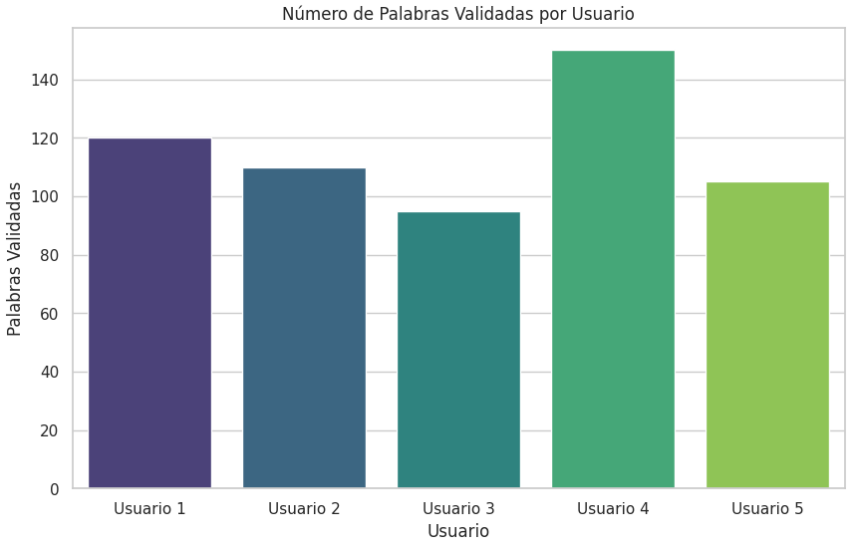

Figura 4.1: Número total de palabras validadas por usuario

Usuario	Palabras Validadas	Comentarios Enviados	Tiempo Promedio de Validación (min)
Usuario 1	120	45	3.5
Usuario 2	110	30	4.2
Usuario 3	95	40	2.9
Usuario 4	150	55	3.8
Usuario 5	105	25	4.0

4.1.2. Precisión de las Traducciones Validadas

La precisión de las traducciones validadas es un indicador clave del éxito de la plataforma. En el análisis realizado, se observó que la mayoría de las traducciones propuestas son correctas, alcanzando niveles de precisión cercanos al 100 % en varios casos. Sin embargo, algunos términos presentan pequeñas discrepancias, probablemente debido a interpretaciones contextuales o al uso de sinónimos. Estos resultados reflejan la efectividad del sistema para garantizar traducciones de alta calidad y destacan áreas donde se puede optimizar el proceso de validación.

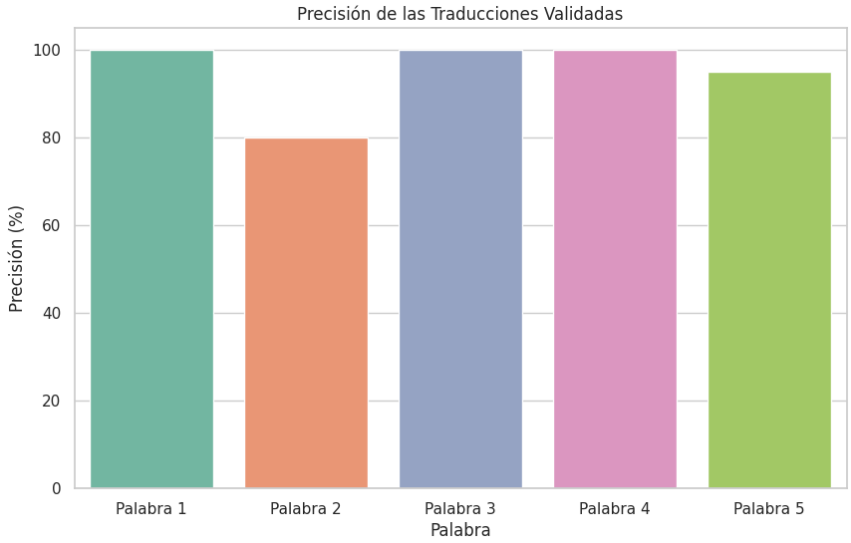

Figura 4.2: Precisión de las traducciones validadas

Palabra	**Traducción Sugerida**	**Traducción Correcta**	**Precisión (%)**
Palabra 1	Silla	Silla	100
Palabra 2	Casa	Vivienda	80
Palabra 3	Agua	Agua	100
Palabra 4	Perro	Perro	100
Palabra 5	Comer	Comer	95

4.1.3. Tiempo de Validación Promedio

El tiempo promedio de validación es un factor determinante en la eficiencia del sistema. En el análisis, se encontró que los usuarios tardan entre 2.9 y 4.2 minutos en validar una palabra, lo que se traduce en un número significativo de validaciones por hora. Este rango demuestra que el sistema es eficiente y permite realizar validaciones rápidamente sin comprometer la calidad. Además, las diferencias entre los tiempos de los usuarios sugieren que la experiencia y familiaridad con la plataforma pueden influir en la rapidez del proceso.

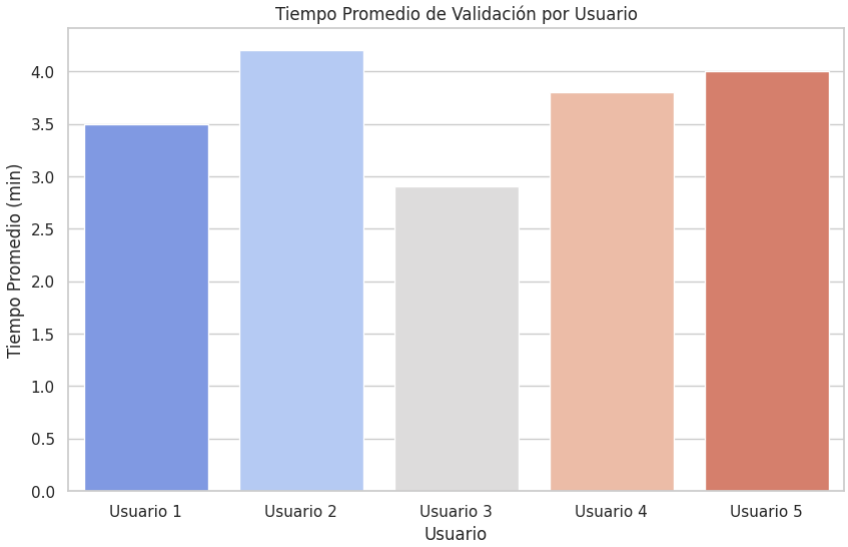

Figura 4.3: Tiempo promedio de validación por usuario

Usuario	Tiempo Promedio de Validación (min)	Validaciones por Hora
Usuario 1	3.5	17.14
Usuario 2	4.2	14.28
Usuario 3	2.9	20.69
Usuario 4	3.8	15.79
Usuario 5	4.0	15.00

4.2. Conclusiones

A continuación, se presentan las conclusiones derivadas del análisis realizado, estructuradas en torno a los objetivos establecidos en este estudio.

- La plataforma de validación de datos lingüísticos ha demostrado ser una herramienta efectiva en la creación de un corpus lingüístico de alta calidad, mejorando la precisión y eficiencia en la validación de datos. La plataforma contribuye significativamente a la creación de corpus, alineándose con la hipótesis alternativa (H), que sostiene que mejora la precisión y eficiencia en comparación con los métodos tradicionales.

- Los requisitos técnicos y funcionales para el diseño y desarrollo de la plataforma han sido adecuadamente cubiertos, garantizando su eficiencia y el manejo de grandes volúmenes de datos, permitiendo la colaboración en tiempo real entre validadores.
- El modelo preliminar ha sido exitoso, con tiempos promedio de validación entre 2.9 y 4.2 minutos, lo que demuestra que la plataforma optimiza el proceso en comparación con los métodos tradicionales, respaldando la hipótesis alternativa (H).
- La plataforma facilita la integración de diversas fuentes de datos y categorías lingüísticas, permitiendo a los validadores trabajar de manera colaborativa en tiempo real y acelerando la creación de un corpus lingüístico de alta calidad.
- La plataforma tiene un gran potencial en la investigación de IA y tecnologías emergentes, ya que permite la validación masiva de datos lingüísticos, mejorando la eficiencia y precisión en el entrenamiento de algoritmos de aprendizaje automático.
- Los resultados respaldan la hipótesis alternativa (H), demostrando que el uso de la plataforma mejora significativamente la precisión y eficiencia en comparación con los métodos tradicionales, cumpliendo con los objetivos planteados.

4.3. Recomendaciones

A partir de los resultados obtenidos y las conclusiones del estudio, se sugieren las siguientes recomendaciones:

- Continuar con la optimización de la plataforma para mejorar la usabilidad y facilitar la interacción de los validadores, especialmente en entornos de trabajo con grandes volúmenes de datos lingüísticos.
- Ampliar la capacidad de la plataforma para soportar más lenguajes y variedades dialectales, lo que podría contribuir a una mayor cobertura y precisión en la validación de datos lingüísticos.
- Implementar mecanismos adicionales de retroalimentación dentro de la plataforma para que los validadores puedan mejorar sus decisiones, a través de sugerencias automatizadas basadas en algoritmos de aprendizaje automático.

- Fomentar la colaboración con comunidades lingüísticas y universidades para ampliar el uso de la plataforma y garantizar que el corpus creado sea representativo de diversas fuentes lingüísticas.

- Evaluar y adaptar la plataforma para su aplicación en áreas más allá de la creación de corpus lingüísticos, como en la investigación en inteligencia artificial, traducción automática y otros campos relacionados con el procesamiento del lenguaje natural.

- Realizar estudios periódicos sobre la precisión y eficiencia de la plataforma, con el fin de asegurar su constante mejora y adaptabilidad a los avances tecnológicos y a las necesidades de los usuarios.

Referencias

Aguilar Santiago, C. A., y García Zúñiga, H. A. (2023, ago.). Tecnologías del lenguaje aplicadas al procesamiento de lenguas indígenas en méxico: una visión general. *Lingüística y Literatura*, *44*(84), 79–102. Descargado de https://revistas.udea.edu.co/index.php/lyl/article/view/354772 doi: 10.17533/udea.lyl.n84a04

Ahmadbek, B. (2024, Aug.). *Western European Journal of Linguistics and Education*, *2*(8), 16–24. Descargado de https://westerneuropeanstudies.com/index.php/2/article/view/1364

Bird, S. (2020). Decolonising speech and language technologies. *Philosophical Transactions of the Royal Society A*, *378*(2168), 20190070.

Bird, S., Klein, E., y Loper, E. (2009). *Natural language processing with python*. O'Reilly Media.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

de Ullibarri Galparsoro, L., y Pita Fernández, S. (1999). Medidas de concordancia: el índice de kappa. *Cad Aten Primaria*, *6*, 169–171.

Gaur, M., Kursuncu, U., Sheth, A., Thirunarayan, K., y Daniulaityte, R. (2021). Ai and nlp

to aid the endangered languages: a tech challenge for language preservation. *IEEE Access*, *9*, 31403-31415.

Gutiérrez-Fandiño, A., Pérez-Fernández, D., Armengol-Estapé, J., Griol, D., y Callejas, Z. (2022). *escorpius: A massive spanish crawling corpus.* Descargado de https://arxiv.org/abs/2206.15147

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (2014). *Métodos de investigación*. McGraw-Hill Interamericana.

Jurafsky, D., y Martin, J. H. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson.

Mager, M., Gutierrez-Vasques, X., Sierra, G., y Meza, I. (2018). Challenges of language technologies for the indigenous languages of the americas. En *Proceedings of the workshop on computational modeling of polysynthetic languages* (pp. 55–61).

Martínez Musiño, C., y Valdez Ramos, J. (2015). Las lenguas indígenas mexicanas en internet: Análisis webométrico. *Lenguas y Literaturas Indoamericanas*, *17*, 122-144. doi: 10.5678/jll.1234

May, S. (2019). Linguistic diversity in the digital age. *Journal of Multilingual and Multicultural Development*, *40*(10), 875-887.

Mendoza, R., y Salas, M. (2023). Aprovechando el poder de la inteligencia artificial para revitalizar lenguas indígenas en peligro: Tecnologías y experiencias. *ARXIV*. Descargado de https://aclanthology.org/2023.lrec-1.789.pdf

Pérez, A., y otros. (2022). Construcción de un recurso para lenguas en peligro en el aula: Dependencias universales para el kakataibo. *ARXIV*. Descargado de https://aclanthology.org/2022.lrec-1.456.pdf

Redman, T. C. (2001). *Data quality: The field guide*. Digital Press.

Sierra, L. M., Meza, E., Montenegro, E. D., Castaño, G. G., González, J. V., y Cobos, C. (2017). *Corpus lingüístico: estudio y aplicación en revitalización de lenguas indígenas* (1.ª ed.). Universidad del Cauca. Descargado 2024-12-22, de http://www.jstor.org/stable/j.ctv1pbwvx0

Sinclair, J. (2004). *Corpus language and linguistic theory: Papers from the 2002 clt conference*. Rodopi.

Tomaselli, M., y Stella, E. (2021). Digital futures for endangered languages. *Language Documentation and Conservation*, *15*, 215-234.

UNESCO. (2022a). *Decade of indigenous languages (2022-2032): Strategic plan*. Paris,

France: UNESCO.

UNESCO. (2022b). *Educational and cultural challenges of indigenous peoples in latin america*. Paris, France: UNESCO.

UNESCO. (2022c). *Global report on indigenous languages*. Paris, France: UNESCO.

UNESCO-IESALC. (2022). *The role of higher education in the preservation of indigenous languages*. Caracas, Venezuela: UNESCO-IESALC.

Viera, A. J., y Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Zevallos, J., y otros. (2022). Huqariq: Un corpus multilingüe de habla de lenguas nativas del perú para el reconocimiento de voz. *ARXIV*. Descargado de https://aclanthology.org/2022.lrec-1.537.pdf

Zevallos, J., y otros. (2023). Avances en el reconocimiento automático de voz para lenguas indígenas: Quechua, guaraní, bribri, kotiria y wa'ikhana. *ARXIV*. Descargado de https://aclanthology.org/2023.lrec-1.123.pdf

Zuazo, K. (2020). Linguistic challenges in the digital age: The case of quechua and aymara in peru. *Revista Peruana de Lingüística*, *2*(1), 45-62.

Apéndice A

Anexos

Este es el contenido del primer anexo.

A.1. Creación del Titulo

Desarrollo de una plataforma de validación de datos para la creación de corpus lingüísticos en desarrolladores, investigadores y validadores a nivel mundial.

- **Causa:** Desinterés global por las lenguas nativas.
- **Efecto:** Baja investigación y desarrollo tecnológico relacionado con estas lenguas debido a la falta de datos validados.
- **Aporte:** Desarrollo de una plataforma de validación de datos lingüísticos.
- **Qué:** la creación de corpus lingüísticos
- **Quién:** Investigadores y desarrolladores de tecnología.
- **Dónde:** A nivel mundial.

A.2. Encuesta

Introducción

Gracias por participar en esta encuesta. Su colaboración es fundamental para mejorar nuestra plataforma de validación de datos. Por favor, responda con sinceridad las siguientes preguntas. Las respuestas serán utilizadas con fines de investigación y desarrollo.

Datos Demográficos

1. ¿Cuál es su edad?
    - 18-25 años
    - 26-35 años
    - 36-45 años
    - 46 años o más
2. ¿Cuál es su nivel educativo?
    - Secundaria completa
    - Técnico superior
    - Universitario
    - Posgrado
3. ¿En qué área trabaja?
    - Tecnología
    - Investigación
    - Educación
    - Otra (especificar) _____

Preguntas sobre la Plataforma

1. ¿Cómo calificaría la facilidad de uso de la plataforma?
    - Muy fácil
    - Fácil
    - Regular
    - Difícil
    - Muy difícil
2. ¿La plataforma cumplió con sus expectativas en cuanto a precisión y confiabilidad de los datos?
    - Totalmente
    - Parcialmente
    - No
3. ¿Considera que la plataforma facilita el trabajo de validación de datos en su campo?
    - Sí
    - No
4. ¿Qué aspectos de la plataforma cree que podrían mejorarse? (Respuesta abierta)

    ---

5. ¿Qué características adicionales le gustaría ver implementadas en la plataforma? (Respuesta abierta)

    ---

6. ¿Recomendaría la plataforma a otros usuarios?
    - Sí
    - No
7. ¿Cuál es su nivel general de satisfacción con la plataforma?

- Muy satisfecho
- Satisfecho
- Neutral
- Insatisfecho
- Muy insatisfecho

A.3. Codigo para generacion de graficos

A continuación se presenta el código Python utilizado para generar los gráficos en el análisis de la plataforma de validación de datos.

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Datos de los usuarios para la participaci n
usuarios = ['Usuario 1', 'Usuario 2', 'Usuario 3', 'Usuario 4', '
    Usuario 5']
palabras_validas = [120, 110, 95, 150, 105]
comentarios = [45, 30, 40, 55, 25]
tiempo_validacion = [3.5, 4.2, 2.9, 3.8, 4.0]

# Crear un DataFrame
df = pd.DataFrame({
    'Usuario': usuarios,
    'Palabras Validadas': palabras_validas,
    'Comentarios Enviados': comentarios,
    'Tiempo Promedio de Validaci n (min)': tiempo_validacion
})

# Gr fico 1: Palabras validadas por usuario
plt.figure(figsize=(10, 6))
sns.barplot(x='Usuario', y='Palabras Validadas', data=df, palette='
    viridis')
plt.title('N mero de Palabras Validadas por Usuario')
plt.xlabel('Usuario')
plt.ylabel('Palabras Validadas')
plt.show()

# Gr fico 2: Tiempo Promedio de Validaci n por Usuario
plt.figure(figsize=(10, 6))
sns.barplot(x='Usuario', y='Tiempo Promedio de Validaci n (min)', data
    =df, palette='coolwarm')
plt.title('Tiempo Promedio de Validaci n por Usuario')
plt.xlabel('Usuario')
plt.ylabel('Tiempo Promedio (min)')
```

```
plt.show()

# Gr fico 3: Comparaci n de Palabras Validadas y Comentarios Enviados
plt.figure(figsize=(10, 6))
df[['Usuario', 'Palabras␣Validadas', 'Comentarios␣Enviados']].set_index
    ('Usuario').plot(kind='bar', figsize=(10, 6))
plt.title('Comparaci n␣de␣Palabras␣Validadas␣y␣Comentarios␣Enviados␣
    por␣Usuario')
plt.xlabel('Usuario')
plt.ylabel('Cantidad')
plt.xticks(rotation=0)
plt.show()

# Gr fico 4: Precisi n de las traducciones (Palabra vs Precisi n)
palabras = ['Palabra␣1', 'Palabra␣2', 'Palabra␣3', 'Palabra␣4', '
    Palabra␣5']
precision = [100, 80, 100, 100, 95]

plt.figure(figsize=(10, 6))
sns.barplot(x=palabras, y=precision, palette='Set2')
plt.title('Precisi n␣de␣las␣Traducciones␣Validadas')
plt.xlabel('Palabra')
plt.ylabel('Precisi n␣(%)')
plt.show()
```

Listing A.1: Código Python para Análisis de Datos